Aix*Marseille université
Socialement engagée

NNT : 2023AIXM0001

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 18 mars 2024 par

Pierre DE LANGEN

Cartographie et signification biologique des régions intergéniques transcrites du génome humain

Discipline
Biologie santé

Spécialité
Bioinformatique et Génomique

École doctorale
ED 62 - Sciences de la vie et de la santé

Laboratoire/Partenaires de recherche
TAGC - Inserm U1090

Composition du jury

Sarah DJEBALI IRSD	Rapportrice
Charles LECELLIER IGMM & LIRMM	Rapporteur
Andrée DELAHAYE-DURIEZ Université Sorbonne Paris Nord & AP-HP	Examinatrice
Salvatore SPICUGLIA TAGC	Président du jury
Benoît BALLESTER TAGC	Directeur de thèse
Lionel SPINELLI TAGC & CIML	Membre invité

umr U 1090 TAGC
theories and approaches of genomic complexity

Affidavit

Je soussigné, Pierre de Langen, 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 8 Janvier 2024



Cette œuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.

Liste de publications et participation aux conférences

Liste des publications et/ou brevets réalisées dans le cadre du projet de thèse :

1. MUFFIN : A suite of tools for the analysis of functional sequencing data Pierre de Langen, Benoit Ballester bioRxiv 2023.12.11.570597; doi : https ://doi.org/10.1101/2023.12.11.570597 This article is a preprint and has not been certified by peer review.
2. Pierre de Langen, Fayrouz Hammal, Elise Guéret, Jean-Christophe Mouren, Lionel Spinelli, Benoit Ballester, Characterizing intergenic transcription at RNA polymerase II binding sites in normal and cancer tissues, Cell Genomics, Volume 3, Issue 10, 2023
3. Fayrouz Hammal, Pierre de Langen, Aurélie Bergon, Fabrice Lopez, Benoit Ballester, ReMap 2022 : a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments, Nucleic Acids Research, Volume 50, Issue D1, 7 January 2022, Pages D316–D325

Participation aux conférences et écoles d'été au cours de la période de thèse :

1. ISMB / ECCB 2023 Lyon (Talk)
2. Colloque ED62 2022 (Poster)
3. JOBIM 2022 Rennes (Poster)
4. Ecole d'été Centuri 2021 (Tuteur)

Résumé

Selon le dogme central de la biologie, l'ADN est transcrit en ARN puis traduit en protéines. Cependant, seulement un à deux pourcents du génome semblent coder pour des protéines, alors qu'une majorité du génome semble pouvoir être transcrite en ARN. Il a été montré que certains de ces ARNs non-codants, étant en général moins abondants et plus instables, étaient nécessaires pour certaines identités cellulaires, notamment l'identité pluripotente. Dans les cancers, une expression aberrante de ces transcrits non-codants a été observée. Une majorité des variants génétiques associés à des maladies ou traits humains sont situés dans ces régions non-codantes.

Dans mon travail de thèse, nous avons réanalysé plus de 900 expériences ChIP-seq ciblant l'ARN Polymérase II, l'enzyme responsable de la génération de ces transcrits. J'ai pu identifier, dans le génome intergénique, plus de 180 000 régions fixées par l'ARN polymérase II, donc probablement transcrites, et identifier dans quels tissus celles ci étaient actives. Nous avons également analysé le signal transcriptionnel au niveau de ces régions dans près de 29 000 expériences RNA-seq provenant d'ENCODE, GTEx et TCGA. Dans les données de cancers issues de TCGA, cela a permis de mettre en évidence de nouvelles régions génomiques pouvant servir de marqueur dont l'expression est associée avec l'état tumoral du tissu ou à la survie du patient. Ce travail a également permis le développement de méthodes d'analyses de données génomique fonctionnant avec un signal bas et des régions non-codantes, qui ont été implémentées dans un package python, Muffin.

Mots clés : ARN Polymérase II, génomique, bioinformatique, Régulation génomique, enhancer, ChIP-seq

Abstract

According to the central dogma of biology, DNA is transcribed into RNA and then translated into proteins. However, only one to two percent of the genome appears to code for proteins, while a majority of the genome seems capable of being transcribed into RNA. It has been shown that some of these non-coding RNAs, generally being less abundant and more unstable, are necessary for certain cellular identities, notably the pluripotent identity. In cancers, aberrant expression of these non-coding transcripts has been observed. A majority of the genetic variants associated with diseases or human traits are located in these non-coding regions.

In my thesis work, we reanalyzed over 900 experiments targeting RNA Polymerase II, the enzyme responsible for generating these transcripts. I was able to identify more than 180,000 regions bound by RNA polymerase II in the intergenic genome, thus likely transcribed, and identify in which tissues they were active. We also analyzed the transcriptional signal at these regions in nearly 29,000 RNA-seq experiments from ENCODE, GTEx, and TCGA. In cancer data from TCGA, this allowed for the identification of new genomic regions that could serve as markers whose expression is associated with the tumor state of the tissue or the patient's survival. This work also enabled the development of methods for analyzing genomic data that work with a low signal and non-coding regions, which have been implemented in a Python package, Muffin.

Keywords: RNA Polymerase II, Genomics, Bioinformatics, Genomic Regulation, Enhancer, ChIP-seq

Remerciements

Cette thèse ne serait pas ce qu'elle est aujourd'hui sans la contribution de nombreuses personnes, et elle a été l'occasion de rencontrer des gens formidables que je tenais à remercier ici.

Tout d'abord, je tiens à remercier Benoît Ballester pour m'avoir encadré pendant ces presque quatre ans de thèse et de stage, et sans qui ma thèse n'aurait pas été possible. Tu as toujours été disponible et impliqué dans ce projet, et j'ai beaucoup appris grâce à toi. En dehors du cadre strictement professionnel, tu es quelqu'un de très sympathique et avec de grandes qualités humaines, ça a toujours été sympa de taper la discute pendant les après-midi sans motivation !

Je remercie également Lionel Spinelli, et ses petites questions techniques qui font qu'on adore le détester, qui a pris un peu du peu de temps qu'il a de libre chaque semaine pour contribuer à faire avancer ce projet, sans lui ce travail n'aurait jamais été aussi abouti. Tu m'as toujours soutenu et bien conseillé, et j'ai également énormément appris grâce à toi.

Je remercie les membres du Jury de thèse, Sarah Djebali, Charles Lecellier, Andrée Delahaye-Duriez et Salvatore Spicuglia, pour avoir accepté mon invitation, et les remercie d'avance d'avoir pris le temps d'évaluer ce gros pavé, et d'avoir fait le déplacement à Marseille. Je remercie également les personnes qui ont fait partie de mes comités de suivi de thèse et ont permis de guider ce projet : Lionel Spinelli (encore), Pierre Milpied, Andreas Zanzoni, Elisabeth Rémy, Charbel Souaid et Julie Bavais.

Je remercie aussi le TAGC pour m'avoir chaleureusement accueilli pendant ces quatre ans et avec qui j'ai pu partager ces fameuses beer sessions. Plus particulièrement, je remercie les doctorants et stagiaires avec qui j'ai probablement joué plusieurs milliers de parties de président, avec entre autres : Jean-Christophe, mon voisin de bureau et acolyte de buffet, que je remercie pour avoir partagé ces moments de glande et de perte de motivation, même si nous ne sommes pas d'accord sur le chauffage ; Fayrouz, mon ancienne voisine de bureau que je remercie pour les assiettes de gâteaux, même si nous ne sommes pas d'accord sur le chauffage (non, ce n'est pas moi le problème, il fait bon dans les autres bureaux) ; Julie, qui ne pourra pas rester tard à la soirée, et membre de la team goûter (quand elle est là) ; Pauline, merci pour les cookies ; et Pauline, à qui je décerne la palme du meilleur pot de thèse ; Mathieu, le sudiste qui parle trop fort ; Juliette, qui mange trop de chocolat et qui

a été là depuis le début du stage ; Loréna, de retour et membre assidue de la team goûter ; Gaëlle, membre à temps partiel de la team goûter ; Élise, ma voisine radine avec qui j'ai du endurer un terrible supplice ; Thomas, le faux doctorant ; Hélléa, la fausse Pauline ; et Marie, qui était quelqu'un de super même si elle frappait les gens, et qui n'a malheureusement pas pu finir sa thèse. Enfin, je remercie ma famille et mes amis qui ont toujours été là pour me soutenir durant cette période et avec qui j'ai passé de bons moments en dehors de la thèse.

Table des matières

Affidavit 2

Liste de publications et participation aux conférences 3

Résumé 4

Abstract 5

Remerciements 6

Table des matières 8

Table des figures 11

Liste des acronymes 13

Avant-propos 16

Introduction 18

1 La régulation de l'expression du génome chez l'Homme 18
- 1.1 Le génome : support de l'information phénotypique . . . 18
- 1.2 La transcription de l'ADN . . . 20
- 1.3 La régulation de la transcription dépendante de l'ARN Polymérase II . 23
  - 1.3.1 La région promotrice . . . 23
  - 1.3.2 Initiation, élongation et pause proximale . . . 24
  - 1.3.3 Une initiation divergente de la transcription . . . 26
  - 1.3.4 La terminaison de la transcription . . . 26
  - 1.3.5 L'ARN Polymérase II . . . 27
  - 1.3.6 Les régions régulatrices distales du génome . . . 29
  - 1.3.7 La conformation spatiale de la chromatine . . . 31
  - 1.3.8 Les séquences répétées . . . 34
  - 1.3.9 Les modifications épigénétiques . . . 36
- 1.4 Fonctions des ARNs non-codants transcrits par l'ARN Polymérase II . . 40
- 1.5 Dysfonctionnements de la régulation dans les maladies et phénotypes humains . . . 42

2 Séquençage de l'ADN 44
- 2.1 Séquençage Sanger et premier séquençage du génome humain . . . 44

2.2 Séquençage par synthèse Illumina, dit de seconde génération (NGS) . 47
2.2.1 Introduction . . . 47
2.2.2 Fragmentation et ajout d'adaptateurs . . . 47
2.2.3 Amplification par Bridge PCR . . . 48
2.2.4 Séquençage par synthèse . . . 48
2.2.5 Analyse bioinformatique du séquençage . . . 49
2.2.6 Limitations et séquençage de troisième génération . . . 51
2.3 Génotypage et GWAS (Genome-Wide Association Studies) . . . 53
2.3.1 Génotypage, déséquilibre de liaison et imputation . . . 53
2.3.2 Héritabilité . . . 53
2.3.3 Modèle additif d'association génotype-phénotype . . . 54
2.3.4 Stratified LD-score regression . . . 54

3 Approches de séquençage fonctionnel à haut débit 56
3.1 Introduction . . . 57
3.2 Identification du génome transcrit : RNA-seq . . . 58
3.2.1 Protocole expérimental simplifié . . . 58
3.2.2 Traitement bioinformatique des données . . . 59
3.2.3 Limitations . . . 60
3.3 Identification des départs de transcription : CAGE . . . 62
3.3.1 Protocole expérimental simplifié . . . 62
3.3.2 Traitement bioinformatique des données . . . 63
3.3.3 Limitations . . . 63
3.4 Identification du génome en cours de transcription : NET-seq, PRO-seq et dérivés . . . 64
3.5 Identification des régions ouvertes de l'ADN : DNase-seq, FAIRE-seq et ATAC-seq . . . 66
3.5.1 Protocole expérimental simplifié . . . 66
3.5.2 Traitement bioinformatique des données . . . 68
3.5.3 Limitations . . . 69
3.6 Identification des régions de fixation de protéines à l'ADN : ChIP-seq et dérivés . . . 70
3.6.1 Protocole expérimental simplifié . . . 70
3.6.2 Traitement bioinformatique des données . . . 72
3.6.3 Limitations et évolutions . . . 73
3.7 Identification de séquences à effet régulateur : MPRAs et STARR-seq . . . 74
3.7.1 Protocole expérimental et analyse bioinformatique . . . 74
3.7.2 Limitations . . . 76
3.8 Séquençage par cellule unique, dit "single-cell" . . . 77
3.8.1 Protocole expérimental simplifié . . . 77
3.8.2 Analyse bioinformatique du séquençage . . . 78
3.8.3 Limitations et évolutions . . . 81

4 Open Data et banques de données génomiques **83**
4.1 Données Ouvertes ou Open Data, reproductibilité et réutilisabilité des données . . . 83
4.1.1 Méthodes ouvertes . . . 84
4.1.2 Données ouvertes . . . 84
4.2 Formats de données génomiques . . . 86
4.2.1 Le format FASTQ . . . 86
4.2.2 Le format FASTA . . . 86
4.2.3 Les formats SAM et BAM . . . 87
4.2.4 Le format BED . . . 87
4.3 Banques de données génomiques . . . 89
4.3.1 La base de données GEO . . . 89
4.3.2 Le consortium ENCODE (ENCyclopedia Of Dna Elements) . . . 89
4.3.3 Le consortium Genotype-Tissue Expression (GTEx) . . . 90
4.3.4 Le projet The Cancer Genome Atlas (TCGA) . . . 90
4.3.5 Le projet ReMap . . . 90
4.3.6 Le projet UK BioBank . . . 91

Résultats **92**

5 Caractérisation de la transcription intergénique aux sites de liaison de l'ARN polymérase II dans les tissus normaux et cancéreux **92**
5.1 Article . . . 92
5.2 Étude de l'occupation de la Pol II et de l'activité transcriptionelle pour les transposons . . . 143
5.3 Etude de la transcription intergénique par séquençage 10X . . . 146

6 Généralisation des méthodes développées **147**
6.1 Identification de modules de régulation pour la base de données ReMap 148
6.2 Travail préliminaire sur l'étude combinatoire de Facteurs de Transcription 160
6.3 Muffin : un package python contenant des outils génériques pour l'analyse de données de séquençage fonctionnel . . . 163

Discussion et perspectives **179**

Bibliographie **183**

Table des figures

1.1 Organisation hiérarchique de la chromatine . . . 20
1.2 Epissage de l'ARN. . . . 22
1.3 Complexe de préinitiation. . . . 23
1.4 Motifs fréquemment rencontrés au niveau des TSS et leur position. . . . 24
1.5 Initiation de la transcription. . . . 24
1.6 Schéma de l'élongation de la Pol II. . . . 25
1.7 Pause proximale de l'ARN Polymérase II. . . . 25
1.8 Transcription divergente. . . . 26
1.9 Terminaison de la transcription. . . . 27
1.10 Structure de l'ARN Polymérase II. . . . 28
1.11 Phosphorylation de l'ARN Polymérase II. . . . 29
1.12 Mécanismes de fonctionnement des enhancers. . . . 30
1.13 Extrusion de boucle. . . . 32
1.14 Conformation de la chromatine et Hi-C. . . . 33
1.15 Classification des transposons. . . . 35
1.16 Méthylation de l'ADN . . . 37
1.17 Hexamère d'histone et modifications N-terminales. . . . 38
1.18 Marques d'histones. . . . 39

2.1 Séquençage Sanger. . . . 45
2.2 Adaptateurs de séquençage Illumina. . . . 47
2.3 Séquençage Illumina. . . . 49
2.4 Assemblage du génome. . . . 50
2.5 Mapping de reads. . . . 51

3.1 RNA-seq. . . . 59
3.2 CAGE. . . . 62
3.3 Comparaison des signaux PRO-seq, GRO-cap et CAGE. . . . 65
3.4 ATAC-seq. . . . 67
3.5 Protocole Chromatin ImmunoPrecipitation Sequencing (ChIP-seq) simplifié. . . . 71
3.6 Signaux de ChIP-seq. . . . 73
3.7 MPRAs et STARR-seq. . . . 75
3.8 Séquençage 10X. . . . 78
3.9 Projection UMAP des transcriptomes de 2700 cellules Peripheral Blood Mononuclear Cell (PBMC)s, analysé par l'outil Seurat. . . . 81

4.1 Fichier FastQ. . . . 86

4.2 Fichier Fasta. 86
4.3 Fichier SAM. 87
4.4 Fichier BED. 88

5.1 Enrichissement en éléments répétés parmi les 180 000 pics consensus. 143
5.2 Enrichissement en éléments répétés parmi les pics consensus actifs dans les cellules embryonnaires. 144
5.3 Enrichissement en éléments répétés parmi les pics consensus Pol II marqueurs dans les cancers. 145
5.4 Visualisation UMAP du jeu de données 10k PBMCs de 10X genomics. . 147

6.1 Identification de Cis Regulatory Module (CRM)s. 149
6.2 Visualisation des CRMs. 161

Liste des acronymes

ADN
Acide désoxyribo-nucléique. 16, 18–20

ADNc
ADN complémentaire. 77, 78

ARN
Acide ribo-nucléique / RNA en anglais. 16, 19, 20, 58, 62, 64

ATAC-seq
Assay for Transposase-Accessible Chromatin with highthroughput sequencing. 66, 89

CAGE
Cap Analysis of Gene Expression. 16, 62, 64, 182

ChIP-seq
Chromatin ImmunoPrecipitation Sequencing. 11, 17, 28, 39, 70, 89, 90, 92, 148

CNV
Copy Number Variation. 43

CRM
Cis Regulatory Module. 12, 148, 149

DE
Différentiellement Exprimés. 59, 60

DNase-seq
DNase I hypersensitive sites sequencing. 66, 89, 161

ENA
European Nucleotide Archive. 90

ENCODE
Encyclopedia of DNA Elements. 16, 17, 21, 90

eQTL
expression Quantitative Trait Loci. 90

eRNA
enhancer RNA. 31, 40

FAIRE-seq
Formaldehyde-Assisted Isolation of Regulatory Elements and sequencing. 66

FANTOM5
Functional Anotation Of The Mammalian Genome. 16, 21, 31

GEM
Gel bead in EMulsion. 77, 78

GEO
Gene Expression Omnibus. 17, 89, 90

GRO-seq
Global Run-On sequencing. 64

GSEA
Gene Set Enrichment Analysis. 60, 68

GWAS
Genome Wide Association Study. 43, 53, 54

IP
Immuno-Précipitation. 70, 71

LDSC
LD-score regression. 55

lncRNA
longs ARNs non-codants. 31, 144

MPRA
Massively Parallel Reporter Assay. 74

NET-seq
Native Elongating Transcription sequencing. 64

PBMC
Peripheral Blood Mononuclear Cell. 11, 81, 146, 163

PCA
Analyse en Composantes Principales. 60, 68, 80

PCR
Polymerase Chain Reaction. 45, 48

PIC
Complexe de préinitiation. 23, 29

Pol II
ARN Polymérase II. 17, 20, 23, 28, 29, 31, 36, 42, 63–65, 70, 92, 161, 179, 182

Poly-A
Poly-Adénylée. 26, 77

PRO-seq
Precision nuclear run-on sequencing. 16, 28, 64

PWM
Position Weight Matrix. 24

RIP-seq
RNA ImmunoPrecipitation Sequencing. 40, 73

RNA-seq
séquençage de l'ARN. 17, 57, 89

RT
Reverse Transcriptase. 77

SAM
Sequence Alignment Map. 87

S-LDSC
Stratified LD-score regression. 55

SNP
Single Nucleotide Polymorphism. 53

STARR-seq
Self-Transcribing Active Regulatory Region sequencing. 74

TAD
Topologically Associating Domain. 32, 33

TCGA
The Cancer Genome Atlas. 90, 180

TES
Sites de fin de transcription. 42

TF
Facteur de Transcription. 23, 25, 29, 31, 36, 42, 57, 66, 70, 91

TFBS
Sites de Fixation de Facteur de Transcription. 29, 57, 68, 72

TSS
site de départ de transcription. 23, 24, 42, 62, 65

TT-seq
Transient Transcriptome sequencing. 64

UMI
Unique Molecular Identifier. 77, 146

Avant-propos

Le projet de séquençage de l'Acide désoxyribo-nucléique (ADN) du génome humain, terminé en 2003, soit près de 20 ans avant la rédaction de ce manuscrit, a révolutionné la génomique, et de façon plus générale, l'étude des maladies chez l'être humain. Alors que le point d'intérêt central de la recherche génomique se portait sur les gènes et leurs mutations codantes, les gènes se sont révélés couvrir une partie du génome encore plus petite qu'attendue. En effet, seulement 1% à 2 % du génome code pour des protéines. Le reste du génome, anciennement incorrectement qualifié de "junk-DNA" ou "ADN poubelle", était déjà connu pour abriter des éléments régulateurs influençant l'expression des gènes ainsi que quelques Acide ribo-nucléique / RNA en anglais (ARN) non-codants. Cependant, en dehors de quelques exemples épars, le génome non-codant était très mal annoté ; et l'on se rendait compte qu'il occupait une place bien plus importante qu'attendu, alors que l'évolution des espèces tend à effacer tout matériel superflu.

Le projet Encyclopedia of DNA Elements (ENCODE)[46] a été lancé fin 2003, avec pour but d'identifier et caractériser tous les éléments fonctionnels du génome, codants ou non-codants. Couplé à une réduction drastique du coût du séquençage de l'ADN, ainsi que la mise en place de protocoles de séquençage dits "fonctionnels", cela a permis de générer une énorme quantité de données de séquençage à analyser par des approches bioinformatique, et d'identifier des éléments régulateurs à haut débit dans différents tissus. Ces approches à haut débit ont permis de découvrir que le génome non-codant ne jouait pas seulement un rôle de régulation direct par l'ADN, mais était en réalité fréquemment transcrit même s'il ne codait pas pour des protéines. Étonnamment, il a été découvert les régions régulatrices à effet dit "enhancer", augmentant l'expression des gènes proches, semblaient être transcrites lorsqu'elles sont actives.

Aujourd'hui, le rôle de chacun de ces transcrits non-codant reste assez mal connu, et leur identification reste assez difficile de par leur instabilité, leur courte taille, ainsi que leur quantité généralement bien plus faible que celle des gènes. Cependant, des projets tels que Functional Anotation Of The Mammalian Genome (FANTOM5) ont permis d'identifier une large quantité de départs de transcription ainsi que d'enhancers transcrits, même si la méthode utilisée, Cap Analysis of Gene Expression (CAGE), reste biaisée et sélectionne plutôt des transcrits abondants et stables. Depuis, des méthodes sensibles telles que le Precision nuclear run-on sequencing (PRO-seq) et ses dérivés ont été développées. Cependant, il existe relativement peu de données disponibles utilisant ces technologies. D'autre part,

d'autres travaux ont mis en évidence une expression aberrante des transcrits non-codants dans diverses maladies, notamment dans les cancers.

Dans un premier temps, j'introduirai les différents mécanismes régulationnels du génome liés à la transcription non-codante ; puis le séquençage de l'ADN, appliqué au séquençage du génome humain, et par la suite de façon plus précise aux méthodes de séquençage fonctionnelles. Enfin, je présenterai les concepts liés aux données ouvertes, ainsi que les principales banques de données génomiques que j'ai utilisé au cours de mes travaux.

Au cours de ma thèse, j'ai utilisé des données de ChIP-seq ciblant l'ARN Polymérase II (Pol II), l'enzyme responsable de la transcription de ces transcrits. Brièvement, cette technologie permet d'identifier quelles régions du génome sont fixées par la Pol II. Cette approche permet de s'affranchir de la relative instabilité de ces transcrits non-codants, même s'il s'agit d'une approche indirecte et donnant peu d'information sur la couverture du transcrit. Afin de n'étudier que des régions non-codantes, nous avons exclu les exons ainsi que les introns, car il n'est pas possible de savoir avec cette approches si l'enzyme est en cours de transcription sur le gène, ou sur un élément non-codant intronique. En utilisant les données publiques de la base de données Gene Expression Omnibus (GEO) et du consortium ENCODE, j'ai ainsi réanalysé uniformément et intégré plus de 1100 expériences de ChIP-seq ciblant la sous-unité POLR2A de la Pol II. Après des contrôles qualité, 906 expériences ont été conservées. En utilisant et adaptant des outils statistiques éprouvés ainsi qu'en développant de nouvelles approches, j'ai pu établir un atlas de 181 546 régions potentiellement transcrites par la Pol II dans les régions intergéniques du génome. Cet atlas a pu être annoté en étudiant dans quelles expériences, et donc quels tissus ou cellules la Pol II est fixée. Par la suite, j'ai étudié le niveau de transcription au niveau de chacune de ces régions potentiellement transcrites dans près de 29 000 jeux de données de séquençage de l'ARN (RNA-seq), notamment sur des jeux de données étudiant les cancers. En compensant un signal faible par un grand nombre d'échantillons et en utilisant cet atlas Pol II comme a priori sur les régions a étudier, cela a permis d'identifier des régions non-codantes dont l'expression est associée à l'état tumoral du tissu ainsi qu'à la survie du patient. De plus, les méthodes développées dans mes travaux peuvent être réutilisées par le biais d'un package python simple d'utilisation.

Introduction

1 La régulation de l'expression du génome chez l'Homme

Sommaire

1.1 Le génome : support de l'information phénotypique 18
1.2 La transcription de l'ADN 20
1.3 La régulation de la transcription dépendante de l'ARN Polymérase II . 23
1.3.1 La région promotrice 23
1.3.2 Initiation, élongation et pause proximale 24
1.3.3 Une initiation divergente de la transcription 26
1.3.4 La terminaison de la transcription 26
1.3.5 L'ARN Polymérase II 27
1.3.6 Les régions régulatrices distales du génome 29
1.3.7 La conformation spatiale de la chromatine 31
1.3.8 Les séquences répétées 34
1.3.9 Les modifications épigénétiques 36
1.4 Fonctions des ARNs non-codants transcrits par l'ARN Polymérase II . . 40
1.5 Dysfonctionnements de la régulation dans les maladies et phénotypes humains 42

1.1 Le génome : support de l'information phénotypique

Le génome est l'ensemble complet de l'ADN d'un organisme. Il contient toutes les informations nécessaires pour construire et maintenir cet organisme. L'ADN est composé de deux chaînes antiparallèles en double hélice [189], formées par des nucléotides. Chaque nucléotide est composé d'un sucre, d'un groupe phosphate et d'une base azotée. Ces quatre bases azotées sont l'adénine (A), la cytosine (C), la

guanine (G) et la thymine (T), et l'ordre de ces nucléotides encode l'information génétique. Dans la structure en double hélice de l'ADN, les bases azotées de chaque chaîne sont appariées à celles de la chaîne complémentaire. L'adénine s'apparie avec la thymine, et la cytosine s'apparie avec la guanine. Ces appariements sont rendus possibles grâce à la formation de liaisons hydrogène entre les bases complémentaires. Cette propriété est essentielle lors de la réplication de l'ADN pendant la division cellulaire. En effet, lors de ce processus, les deux brins de l'ADN sont séparés par l'Hélicase, et différentes ADN Polymérases synthétisent le brin complémentaire pour chaque brin originel. Au final, le génome entier est dupliqué, et chaque cellule fille reçoit une copie du génome, comportant un brin originel et un brin répliqué [18].

L'ADN du génome humain comporte plus de trois milliards de paires de bases : s'il était étiré, il mesurerait près de 2 mètres de long. Dans la cellule, l'ADN est stocké dans le noyau, et compacté sous forme de chromatine, à l'aide de protéines histones. Ces dernières forment des nucléosomes, composés de huit histones (2 exemplaires des histones H2A, H2B, H3 et H4), autour desquels l'ADN s'enroule, formant l'euchromatine. Par la suite, cette chromatine peut être condensée davantage avec l'aide de l'histone H1, permettant de créer une fibre de chromatine compacte de 30nm, formant l'hétérochromatine, très peu accessible par d'autres protéines (Figure 1.1). [207] L'ADN sert donc de support dont l'information est duplicable et peut se propager. Cependant, pour produire une cellule fonctionnelle, ce code génétique doit être interprété pour générer de nouvelles molécules interagissant entre elles et avec l'environnement : à l'aide de mécanismes complexes, l'ADN est transcrit en ARNs, certains étant par la suite traduits en protéines.

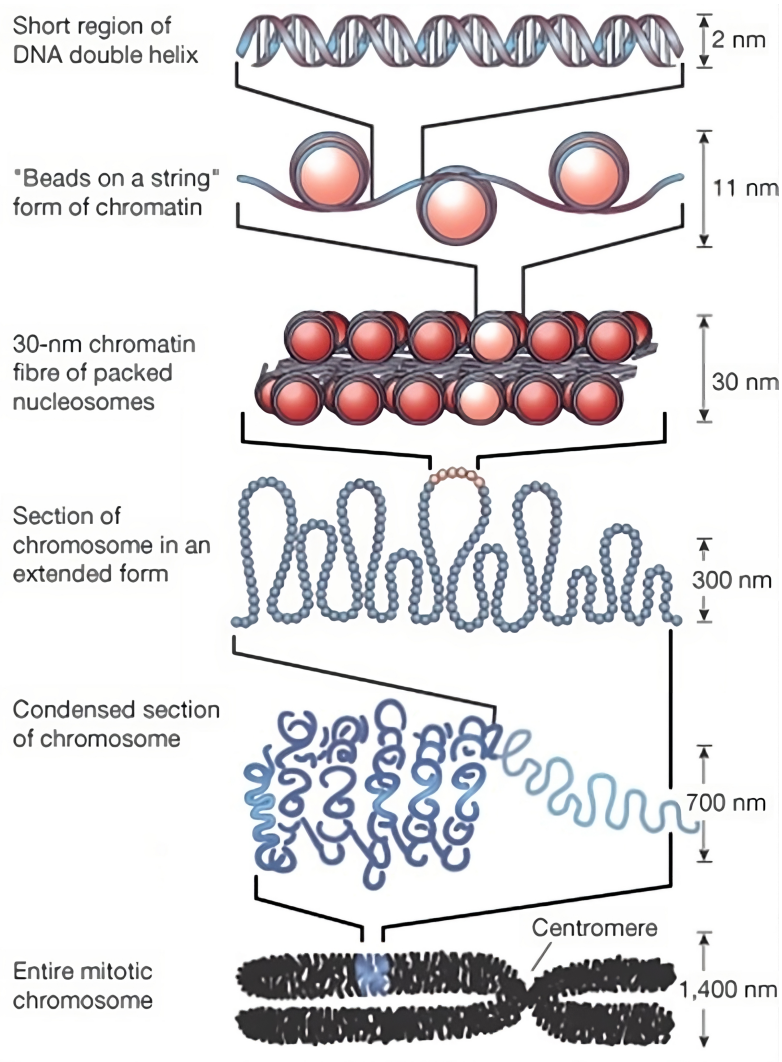


FIGURE 1.1 – Organisation hiérarchique de la chromatine. Édité de "Controlling the double helix", Felsenfeld & Groudine, Nature, 2003 [59].

1.2 La transcription de l'ADN

La transcription est le processus par lequel un ARN est généré à partir des bases complémentaires d'un brin d'ADN matrice. L'ARN est chimiquement proche de l'ADN, avec quelques distinctions : le sucre utilisé est le ribose au lieu du désoxyribose, l'ARN est monocaténaire (chez l'humain), et la thymine est remplacée par l'uracile. Cette transcription est catalysée par une ARN polymérase. Chez les mammifères, il y a trois ARN polymérases, notées ARN Polymérase I, II et III. L'ARN Polymérase I transcrit les ARN ribosomiques, tandis que l'ARN Polymérase III transcrit les ARN de transfert. Ces ARN sont cruciaux pour la cellule car ils permettent la traduction des ARNs messagers en protéines. L'ARN Polymérase II (Pol II) est en charge de la transcription de tous les gènes codant pour des protéines ainsi que de nombreux

transcrits non codants.

Bien que la quantité d'ARN produite par la Pol II soit inférieure à celle des deux autres ARN polymérases, représentant moins de 5% des ARNs totaux [71], la diversité de ces transcrits est vaste. Le projet GENCODE [61] a référencé 252 835 transcrits différents chez l'humain, dont 89 067 codent pour des protéines. Toutefois, la base de données GENCODE n'est pas exhaustive, se concentrant principalement sur des transcrits longs avec une grande confiance en leur existence. D'autres initiatives, comme LNCipedia [183], cherchent à cataloguer les transcrits non codants mentionnés dans la littérature, souvent omis de GENCODE. Les travaux des consortiums FANTOM5 [5] et ENCODE [57] ont suggéré que la majorité de notre génome pourrait être transcrite, en s'appuyant des protocoles de séquençage à haut débit ciblant l'ARN, tels que le RNA-seq (Section 3.2) ou le CAGE (Section 3.3).

La Pol II produit divers types de transcrits. Certains peuvent être polyadénylés, en particulier les ARNs messagers qui sont exportés hors du noyau; cette étape augmente également leur stabilité. L'ARN, étant moins stable que l'ADN, a une durée de vie allant de quelques minutes à quelques jours. Les transcrits peuvent également subir un épissage (Figure 1.2), où les séquences introniques sont éliminées, ne laissant que les séquences exoniques codant pour une protéine, ainsi que quelques régions non-traduites ou UTR en amont et en aval du transcrit[94]. Ces transcrits peuvent subir un épissage alternatif, où différents exons sont conservés pour un même transcrit initial; on parle alors d'isoformes de transcrits. De manière surprenante, certains ARNs non codants peuvent subir ce processus d'épissage sans être traduits en protéines. La taille des ARNs varie fortement, allant d'une vingtaine de nucléotides pour les micro-ARNs à 2,2 millions de nucléotides pour le gène de la dystrophine, avant épissage[65].

L'ARN Polymérase II est donc responsable de la transcription d'une extrême variété de transcrits. Cependant, il n'est pas nécessaire pour une cellule de produire tous ces transcrits en même temps à un instant précis. Si certains transcrits, dits "housekeeping", sont essentiels pour le fonctionnement basique de la cellule (comme ceux codant pour les ADN ou ARN polymérases), d'autres permettent la spécialisation cellulaire, le contrôle du développement ou l'interaction avec l'environnement. Il est donc crucial pour l'organisme de moduler l'activité transcriptionnelle de l'ARN Polymérase II pour spécialiser ses cellules, orchestrer son développement, s'adapter à l'environnement et économiser de l'énergie.

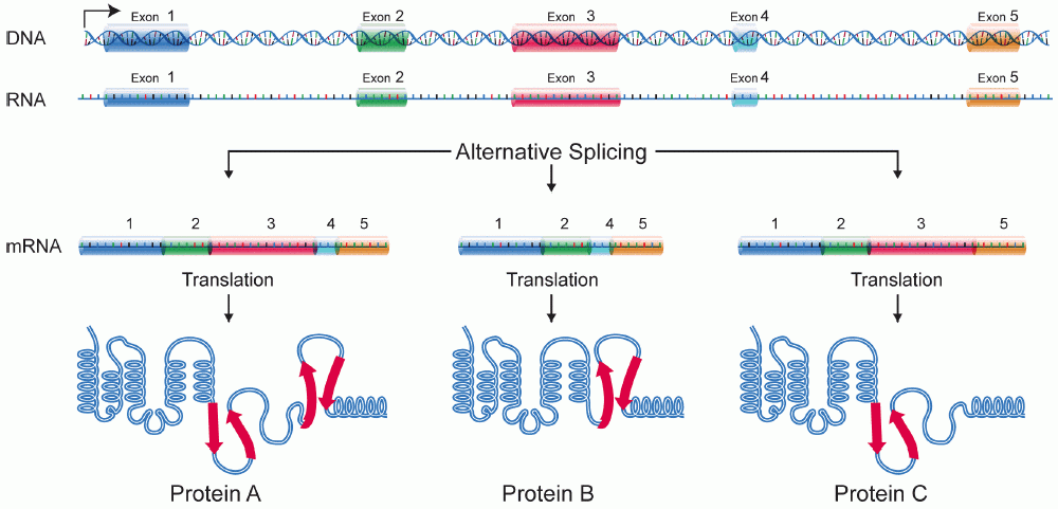


FIGURE 1.2 – Epissage de l'ARN. L'ADN est transcrit en ARN, maturé et épissé, puis transcrit en protéines. Provient du National Human Genome Research Institute, Public Domain [78, 82]

1.3 La régulation de la transcription dépendante de l'ARN Polymérase II

1.3.1 La région promotrice

Un élément clé de la régulation de la transcription est l'assemblage de la machinerie transcriptionnelle au niveau de la région dite "promotrice" du transcrit (1.3). Ceci est suivi par l'initiation de la transcription au niveau du site de départ de transcription (TSS), qui représente la première base d'ADN transcrite en ARN. Cette région contient des motifs nucléotidiques reconnus par des protéines appelées Facteur de Transcription (TF). Les TFs sont des protéines se fixant à l'ADN et étant capables de moduler l'activité transcriptionnelle. Certains de ces TF sont dits généraux, on parle dans ce cas de GTFs, ils sont au nombre de six : TFIIA, TFIIB, TFIID, TFIIE, TFIIF, et TFIIH. Ces derniers sont nécessaires pour recruter la Pol II afin de former le Complexe de préinitiation (PIC) et le stabiliser [179]. A ce PIC minimal peuvent s'ajouter plus d'une centaine de protéines, dont le complexe mediator, capables de moduler l'activité transcriptionnelle de la Pol II en fonction de signaux cellulaires et de la composition du PIC. [72].

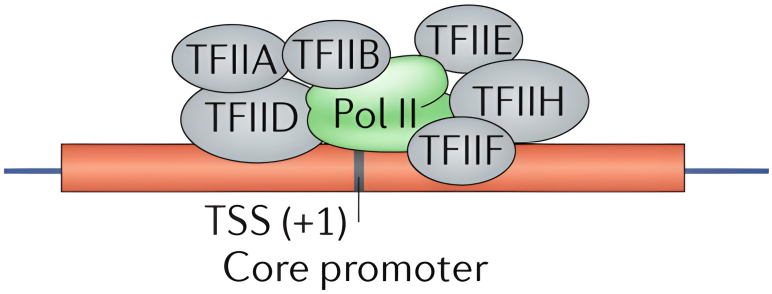


FIGURE 1.3 – Complexe de préinitiation minimal positionné au niveau du TSS. Adapté de Haberle & Stark, Nature Reviews Molecular Cell Biology, 2018 [72].

Avec l'analyse à grande échelle du génome, de nombreuses études ont démontré l'absence d'un promoteur canonique chez les eucaryotes. Bien que certains motifs, comme la TATA box, soient couramment identifiés dans les promoteurs des gènes eucaryotes à des positions spécifiques par rapport au TSS, il n'y a pas de séquence consensus strictement requise pour l'initiation de la transcription chez l'humain. Il semble exister une vaste diversité de combinaisons de motifs d'ADN capables d'initier la transcription, qui ne semble pas pouvoir être expliquée par des modèles simples. Des modèles récents utilisant des approches de Deep Learning semblent pouvoir prédire les sites d'initiation de la transcription identifiés par CAGE (voir Section 3.3) [10, 93, 156], mais la compréhension des mécanismes d'initiation de la transcription basés sur la séquence d'ADN reste un enjeu majeur de la génomique.

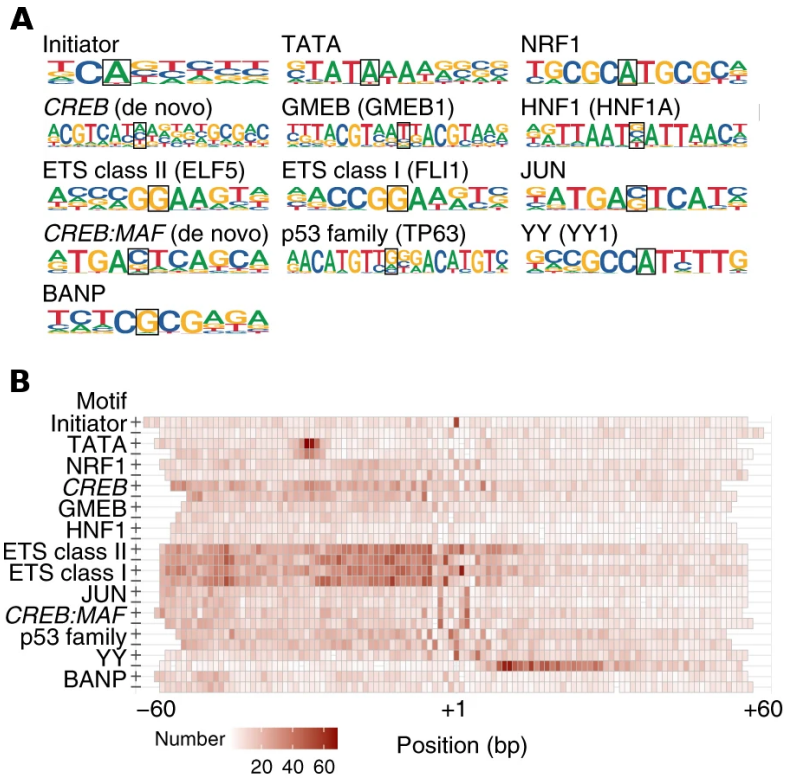


FIGURE 1.4 – Motifs fréquemment rencontrés au niveau des TSS et leur position relative. **A.** Logos de séquence des Position Weight Matrix (PWM) des motifs. **B.** Position relative des motifs par rapport au TSS. Adapté de Sahu et al., Nature Genetics, 2022 [156].

1.3.2 Initiation, élongation et pause proximale

A l'aide du complexe de préinitiation, la double hélice d'ADN se sépare pour former une courte "bulle de transcription" (Figure 1.5 et 1.6) permettant à l'ARN Polymérase II d'initier la transcription du premier nucléotide d'ARN au niveau du TSS et entrer en phase d'élongation : l'enzyme allonge nucléotide par nucléotide la molécule d'ARN en cours de synthèse avec la base correspondante au brin d'ADN matrice [33].

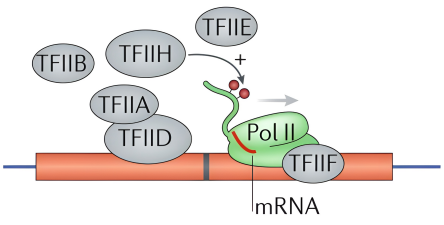


FIGURE 1.5 – Initiation de la transcription. La Pol II quitte le TSS et commence à transcrire un ARN. Les billes rouges correspondent à des phosphorylations de la Pol II (abordées dans la section 1.3.5). Adapté de Haberle & Stark, Nature Reviews Molecular Cell Biology, 2018 [72].

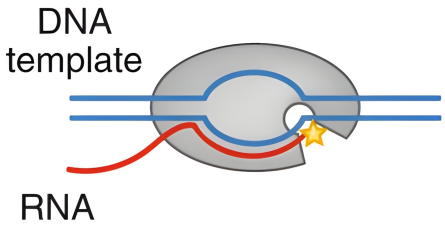


FIGURE 1.6 – Schéma de l'élongation de la Pol II et formation de la bulle transcriptionnelle. L'étoile correspond à un nouveau nucléotide ajouté. Adapté de Conaway & Conaway, Nature Structural & Molecular Biology, 2022 [40].

Cependant, chez les métazoaires, il a été observé une pause de la Pol II 30 à 50 bases après l'initiation dans de nombreux gènes(Figure 1.7). Cette pause est un élément clef de la régulation de l'expression du transcrit naissant, car elle permet de moduler l'expression en fonction de la propension de la transcription à être mise en pause. Ainsi, il a été montré que certains gènes étaient très peu exprimés malgré une très forte concentration de débuts de transcrits et de Pol II juste après leur région promotrice. Ce mécanisme est souvent présent au niveau des gènes nécessitant une réponse rapide à un stimuli extérieur, par exemple les gènes liés au stress cellulaire ou les gènes développementaux. [42]

Si l'on sait que certaines protéines, telles que NELF, DSIF ou P-TEFb régulent respectivement positivement et négativement la pause de la Polymérase, les mécanismes causant cette pause restent assez mal connus. Il semblerait que plusieurs facteurs influent sur la propension d'un transcrit à posséder cette pause proximale au promoteur : premièrement, les TFs recrutés au niveau du promoteur présents dans le complexe de préinitiation ; et deuxièmement la composition de la séquence transcrite. Il semblerait qu'un taux de GC élevé augmente le taux de pause, de par la formation de "R-loops", qui sont des appariements entre un ADN et un ARN tout deux à simple brin, ralentissant la transcription de la Pol II. [1] Par la suite, la transcription peut soit reprendre suite au recrutement de P-TEFb par un facteur de transcription, soit être avortée.

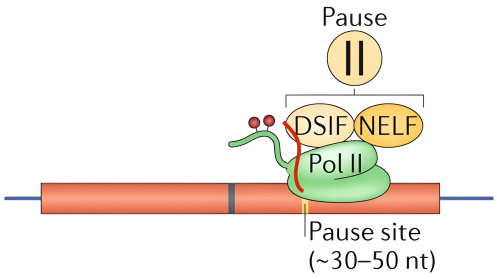


FIGURE 1.7 – Pause proximale de l'ARN Polymérase II. Adapté de Haberle & Stark, Nature Reviews Molecular Cell Biology, 2018 [72].

1.3.3 Une initiation divergente de la transcription

L'utilisation de protocoles de séquençage à haut débit ciblant l'ARN (voir Section 3.2) a permis d'identifier un grand nombre de transcrits partageant une même région promotrice, mais étant transcrit dans des directions divergentes (l'un en 5'->3' et l'autre 3'->5', voir Figure 1.8) [44, 165, 166], soulevant la question des mécanismes liés à ce partage de promoteurs. Le projet FANTOM5, à l'aide du séquençage CAGE, a identifié une transcription très souvent bidirectionelle au niveau des enhancers (décrits dans la prochaine section). Le protocole GRO-cap (voir Section 3.4), une méthode sensible permettant de détecter efficacement les transcrits naissants, a notamment permis de découvrir que la transcription s'initiait de façon bidirectionelle chez l'humain, que ce soit au niveau des promoteurs ou des enhancers, produisant une paire de transcrits stable/instable, stable/stable ou instable/instable [43]. Cette transcription s'initie principalement aux bordures d'une région sans nucléosome, facilement accessible par les TFs et la machinerie transcriptionnelle [166, 199].

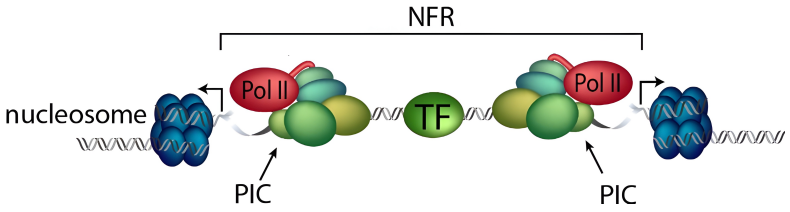


FIGURE 1.8 – Modèle d'initiation divergente de la transcription en bordure de NFR (Nucleosome Free Region). Adapté de Seila et al., Cell Cycle, 2009 [166].

1.3.4 La terminaison de la transcription

S'il agit d'un évènement clef dans le processus transcriptionnel, les mécanismes de terminaison de la transcription ne sont pas parfaitement définis. Dans le cas des transcrits à queue Poly-Adénylée (Poly-A), mieux étudiés, nos connaissances actuelles mènent à un modèle dit "Torpedo/Allosterique" [53] (Figure 1.9). Brièvement, dans ce modèle, la séquence du signal de polyadénylation serait reconnue par la Pol II et son complexe protéique. La Pol II ralentit sa vitesse de transcription, et un clivage de l'ARN synthétisé s'effectue après le signal Poly-A [202], libérant le transcrit. La seconde partie du transcrit clivé, toujours en cours de transcription par la Pol II, est dégradé dans la direction 5'->3' par l'enzyme XRN2, qui possède un point d'entrée 5'-phosphate par la coupure. Une fois que XRN2 atteint la Pol II, cette dernière induit la terminaison de la transcription et se détache de l'ADN [191]. La Pol II peut continuer de transcrire plus de 1 000 à 2 000 nucléotides, dégradés par XRN2, avant son arrêt.

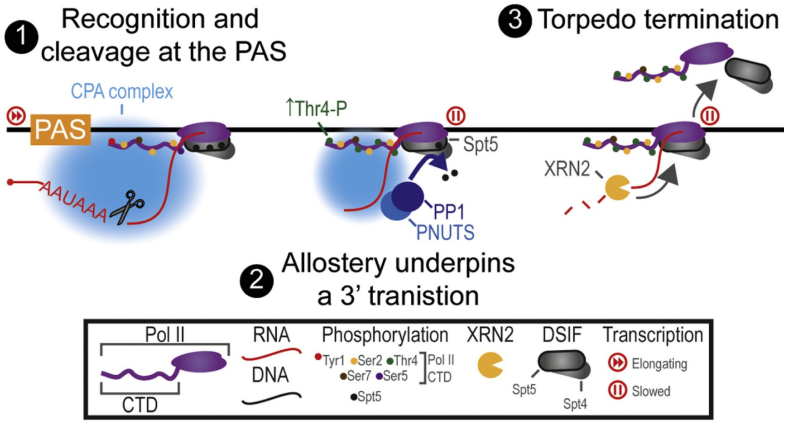


FIGURE 1.9 – Modèle "Torpedo/Allostérique" de terminaison de la transcription. CPA : Cleavage and Poly-Adenylation. Adapté de Eaton & West, Trends in Genetics, 2020 [54].

Le mécanisme conduisant à l'utilisation de sites de terminaison alternatifs est régulé par les protéines SCAF4 et SCAF8 (et potentiellement d'autres protéines), se fixant sur l'ARN en amont des sites de polyadénylation, empêchant le clivage et la polyadénylation du transcrit, et donc le processus de terminaison de la transcription [70].

1.3.5 L'ARN Polymérase II

Chez l'Humain, cette enzyme est composée de 12 sous-unités (Figure 1.10) nomées de RPB1 à RBP12 (protéine) ou POLR2A à POLR2L (nom du gène correspondant). Sa plus grosse sous-unité, RPB1 ou POLR2A, fait partie du domaine de liaison à l'ADN. Son domaine C-terminal possède chez l'Humain 52 répétitions d'un heptapeptide Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7 pouvant être phosphorylé à des positions précises [16, 196].

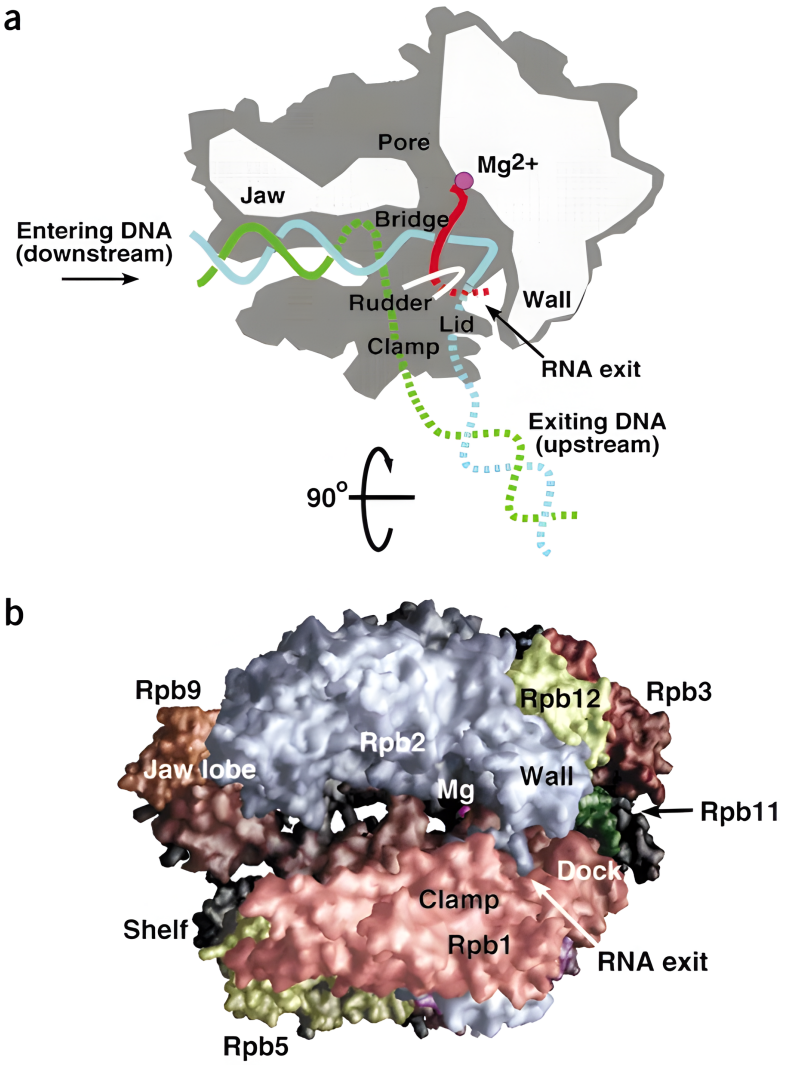


FIGURE 1.10 – Structure de l'ARN Polymérase II. **a.** Vue du dessus schématique de l'ARN Polymérase II. **b.** Vue de côté de l'ARN Polymérase II et de ses principales sous-unités. Adapté de Hahn, Nature Structural & Molecular Biology, 2004 [73].

L'état de phosphorylation de ce domaine C-terminal est caractéristique de l'état de la Pol II : une absence de phosphorylation correspond à un état de pré-initiation; une phosphorylation sur le cinquième résidu de la répétition (Ser5P ou S5P) correspond à une Pol II ayant initié la transcription mais en état de pause; des phosphorylation concomitantes sur les résidus Tyr1P-Ser7P sont caractéristiques d'une Pol II en cours d'élongation; et enfin une phosphorylation Ser2P-Thr4P indique une terminaison de la transcription (Figure 1.11). Les positions sur le génome des complexes Pol II peut être identifiée par PRO-seq (Section 3.4) ou ChIP-seq (Section 3.6). Dans le cas du ChIP-seq, il est également possible de cibler un de ses état de phosphorylation.

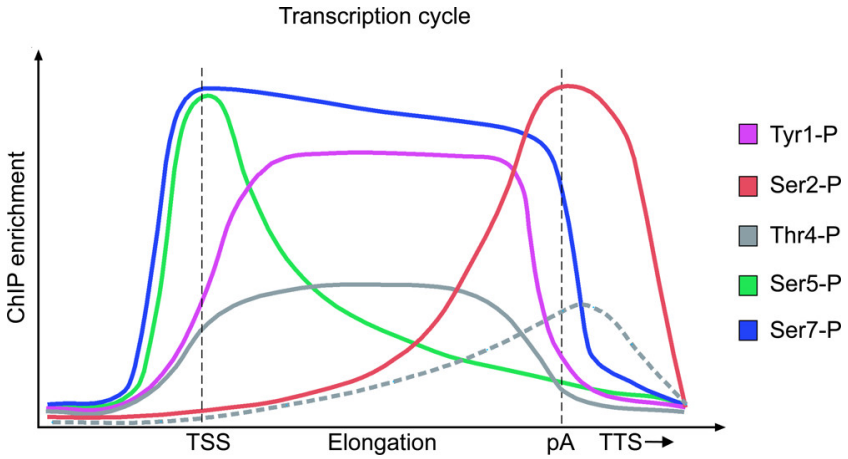


FIGURE 1.11 – Phosphorylation et cycle de transcription de l'ARN Polymérase II. Niveaux de phosphorylation de la Pol II observés par ChIP-seq selon la position sur le transcrit, chez la levure. Des résultats similaires sont observés chez les mammifères, à l'exception du niveau de Thr4-P. La ligne Thr4-P en pointillés correspond aux résultats observés chez l'humain. Adapté de Eick & Geyer, Chem. Rev. 2013 [56].

1.3.6 Les régions régulatrices distales du génome

Si la régulation de la transcription s'effectue en partie au niveau du promoteur et du transcrit en lui même, certaines séquences d'ADN sont capables de réguler l'expression des gènes à longue distance de la séquence promotrice. Ces régions du génome peuvent soit augmenter le niveau d'expression des gènes, et sont dites dans ce cas-ci "enhancers" [170] ; inversement, elles peuvent réprimer l'expression des gènes, et sont alors dites "silencers" [141]. Il a été observé que ces séquences peuvent agir jusqu'à un million de paires de bases du promoteur régulé.

Le mécanisme précis de fonctionnement des enhancers n'est pas parfaitement défini. Cependant, les régions enhancers sont capables de recruter des protéines TFs de par la présence de Sites de Fixation de Facteur de Transcription (TFBS) dans leur séquence. A l'aide du complexe CTCF-Cohesin, celles-ci sont rapprochées et mises en contact du promoteur en formant des "boucles" dans l'ADN [3] (Figure 1.12). Plusieurs mécanismes peuvent expliquer le fonctionnement des enhancers. L'augmentation de la concentration en TFs à proximité du promoteur permet probablement de faciliter le recrutement de ces TFs et GTFs au niveau du promoteur. D'autre part, ces TFs peuvent interagir avec le PIC, la Pol II et notamment avec le complexe mediator afin de moduler l'activité de la Pol II, et ce à différentes étapes du processus transcriptionnel [72, 192]. Les silencers, moins bien étudiés, semblent suivre des mécanismes similaires au enhancers, mais recrutent des TFs répresseurs de transcription. De façon étonnante, de nombreux silencers semblent également pouvoir générer une activité enhancer, et le passage d'une fonction à l'autre serait

dépendant du contexte cellulaire [164].

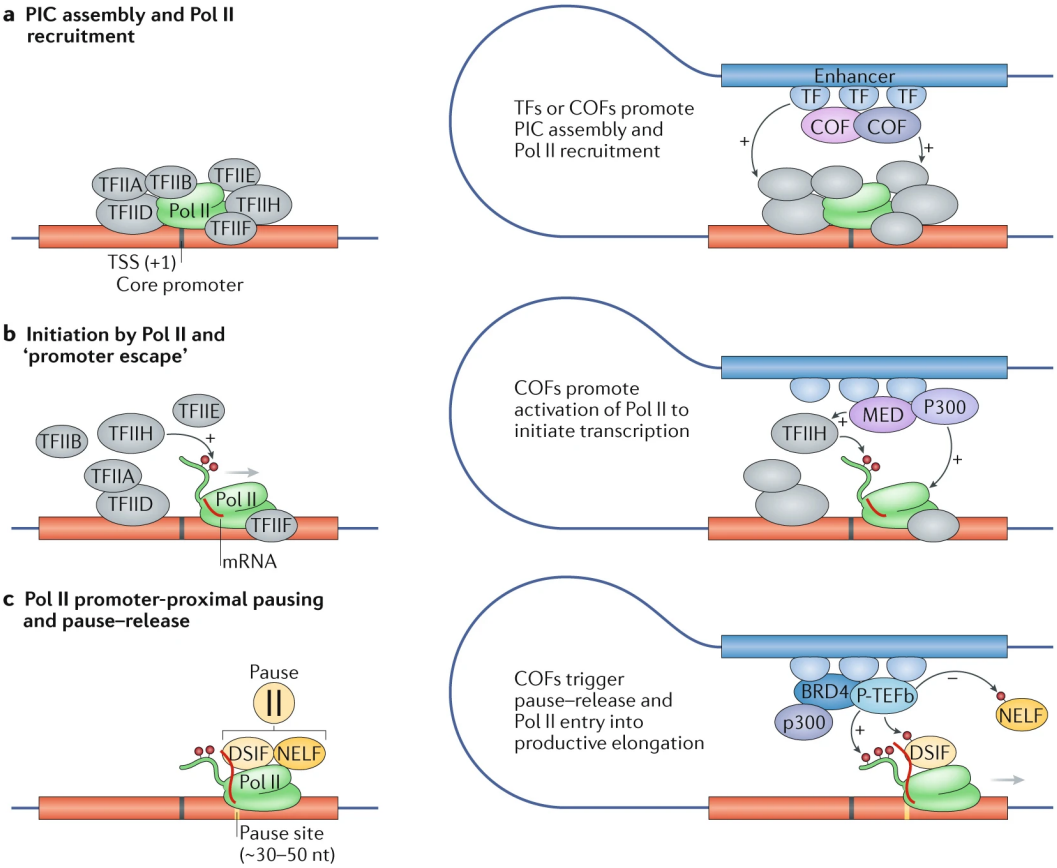


FIGURE 1.12 – Mécanismes possibles de fonctionnement des enhancers. Selon les TFs et Cofacteurs (COFs) recrutés, les enhancers peuvent réguler différents points de la transcription. Adapté de Haberle & Stark, Nature Reviews Molecular Cell Biology, 2018 [72].

Souvent, les transcrits possédant des fonctions similaires et devant être co-exprimés sont positionnés de façon proche dans le génome, afin de partager des éléments régulateurs similaires. De même, les enhancers sont parfois localisés en groupe dans le génome, formant des "super-enhancers" recrutant une très forte concentration de TFs [150]. Le mode de fonctionnement des enhancers ou silencers semble être principalement additif et modulaire : plusieurs enhancers agissant sur le même promoteur vont additionner leurs effets de façon indépendante l'un de l'autre [156]. On dit aussi qu'ils sont modulaires : une fois actifs, leur activité semble dépendre relativement peu du promoteur cible, et ils peuvent être ainsi déplacés ou dupliqués pour agir sur de nouveaux promoteurs, offrant potentiellement un levier évolutif pour moduler l'expression des gènes sans modifier ceux-ci [156, 164]. De même, il a été observé que les promoteurs de gènes étaient actifs dans un grand nombre de types cellulaires, alors que l'activation des enhancers est en général bien plus spécifique [131] : les enhancers permettent de moduler précisément l'expression des gènes dans

différents types cellulaires en conservant un même promoteur.

Les enhancers sont également capables de recruter la Pol II, et de façon étonnante, la plupart des enhancers actifs d'une cellule semblent transcrits, générant des enhancer RNA (eRNA), dont le niveau d'expression semble corréler avec l'activité enhancer [95, 158]. Ces eRNAs sont pour la plupart très courts, d'une cinquantaine de bases, et sont issus d'une transcription fréquemment bidirectionnelle [5, 95]. Certains eRNAs, au contraire, sont des longs ARNs non-codants (lncRNA), pouvant parfois être polyadénylés et épissés [160]. De plus, certains promoteurs de gènes codant pour des protéines montrent une activité enhancer, et sont nommés ePromoters [129]. La transcription des eRNAs a permis au consortium FANTOM5 d'identifier plus de 65 000 potentiels enhancers transcrits [5] à l'aide du protocole CAGE (voir Section 3.3). Des travaux récents suggèrent un gradient entre enhancers et promoteurs de par leur faculté à recruter des TFs et à initier la transcription d'une façon plus ou moins directionnellement stable [132].

Les interactions entre éléments régulateurs (enhancers ou silencers) et promoteurs sont elles-même régulées : les éléments régulateurs ont leur domaine d'activité limité à certains promoteurs.

1.3.7 La conformation spatiale de la chromatine

Afin de permettre aux éléments régulateurs du génome d'interagir entre eux, la conformation spatiale de l'ADN doit être altérée. Cela permet d'une part de mettre les couples enhancer/silencer-promoteurs en contact, mais également de délimiter le domaine de régulation de ces enhancers/silencers. La construction de ces "boucles" est effectuée en grande partie par le complexe Cohesin, couplée à la protéine CTCF [176], par un mécanisme d'extrusion de boucle [45] (Figure 1.13).

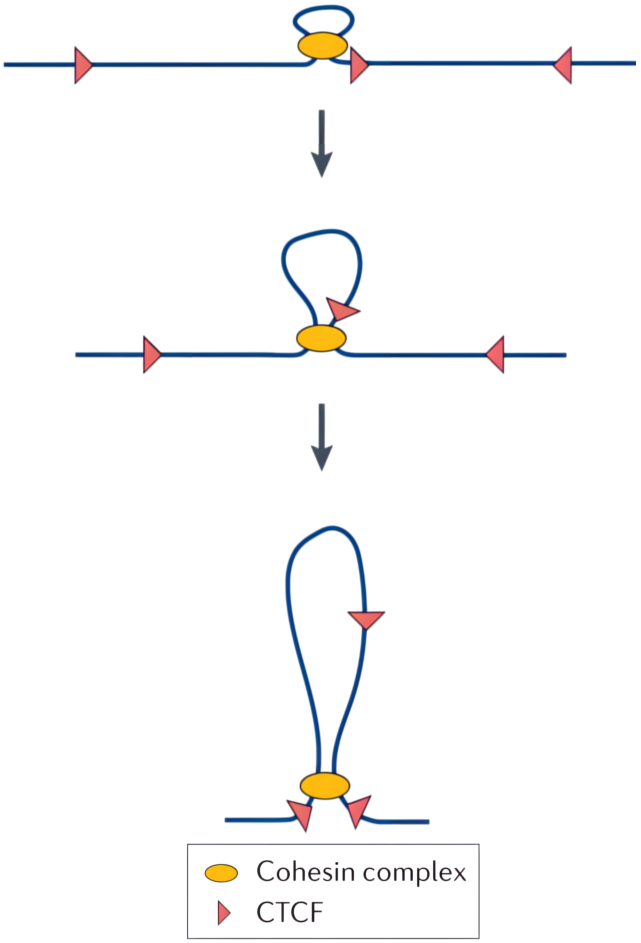


FIGURE 1.13 – Extrusion de boucle par le complexe CTCF-cohesin. La cohesin extrude une boucle d'ADN, jusqu'à être arrêtée par deux complexes CTCF orientés. Adapté de Davidson & Peters, Nature reviews molecular cell Biology, 2021 [45].

La topologie du génome est hiérarchique : elle s'organise premièrement en chromosomes ayant peu d'interactions entre eux; puis ces chromosomes sont divisés en compartiments, et enfin de manière plus fine en Topologically Associating Domain (TAD), voire en sous-TAD (Figure 1.14). De façon générale, un même compartiment ou TAD à une tendance à englober une chromatine dans un état similaire, et des gènes codant pour des protéines avec des fonctions biologiques similaires, indiquant une co-régulation de la région [17]. La conformation de la chromatine peut être étudiée à l'aide de protocoles de séquençage Hi-C, permettant de quantifier à l'échelle du génome la proximité spatiale qu'ont les régions du génomes entre elles dans une condition biologique donnée [111].

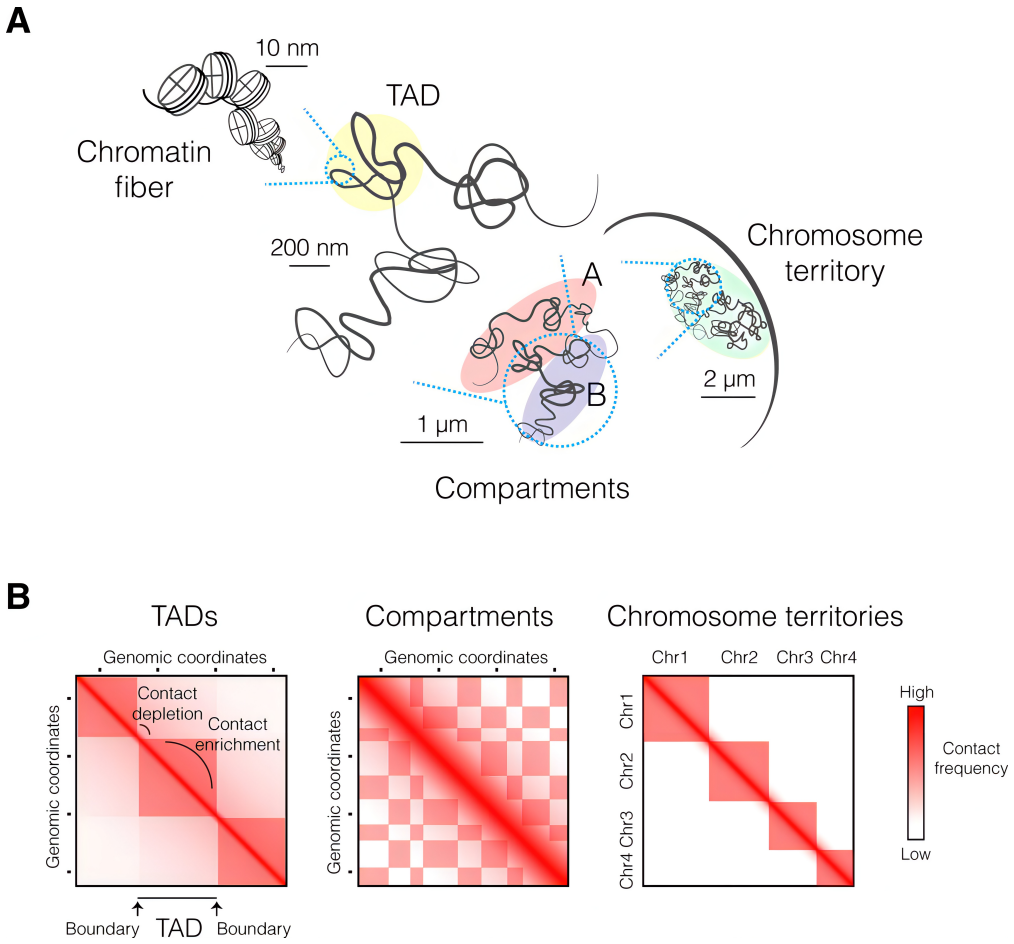


FIGURE 1.14 – **A.** Vue schématique du repliement des chromosomes à l'intérieur du noyau. La couche la plus fine du repliement de la chromatine se situe au niveau de l'association ADN-histone, formant des nucléosomes organisés en une fibre de chromatine. La chromatine est emballée à des densités de nucléosomes différentes en fonction de la régulation génique et se replie en domaines d'intéractions préférentielles, appelés TADs. À l'échelle chromosomique, la chromatine est séparée en compartiments "A" actifs et "B" réprimés en matière d'interactions, reflétant les contacts préférentiels entre les régions de chromatine présentant les mêmes caractéristiques épigénétiques. Les chromosomes individuels occupent leur propre espace à l'intérieur du noyau. **B.** Représentation schématique des cartes Hi-C, reflétant les différentes hiérarchies du repliement de la chromatine. Les TADs apparaissent sous forme de carrés le long de la diagonale, enrichis en interactions et séparés par des zones de déplétion des contacts délimitées par les limites des TAD. À l'échelle chromosomique, les interactions à longue distance de la chromatine forment un motif caractéristique de compartiments A et B mutuellement exclus. Enfin, les interactions intrachromosomiques sont surreprésentées par rapport aux contacts interchromosomiques. Adapté de Szabo et al., Science Advances, 2019 [176].

1.3.8 Les séquences répétées

Plus de la moitié du génome humain est composée de séquences répétées. Parmi celles-ci la majorité sont des transposons (voir Table 1.1 et Figure 1.15) . Les éléments transposables sont des éléments répétés représentant près de la moitié du génome humain, et ont originellement possédé un promoteur et des séquences codantes leur donnant la faculté de se déplacer ou se dupliquer dans le génome. Il a été montré qu'une partie de ces éléments transposables a permis de générer de nouveaux éléments régulateurs promoteurs, enhancer ou silencer, ou à les déplacer [36]. Si les mutations qu'ils induisent sont potentiellement délétères pour l'organisme, ce qui a causé l'inactivation quasi-complète de leur activité de transposition au fil de l'évolution, les éléments transposables ont fortement contribué fortement à refaçonner la régulation des gènes pour l'humain et les métazoaires en général. Ils contribuent également à générer de nombreux ARNs non-codants [24], certains étant essentiels pour l'embryogénèse [148, 159]. La structure et les différentes familles de transposons sont décrites dans la Figure 1.15 ci-dessous.

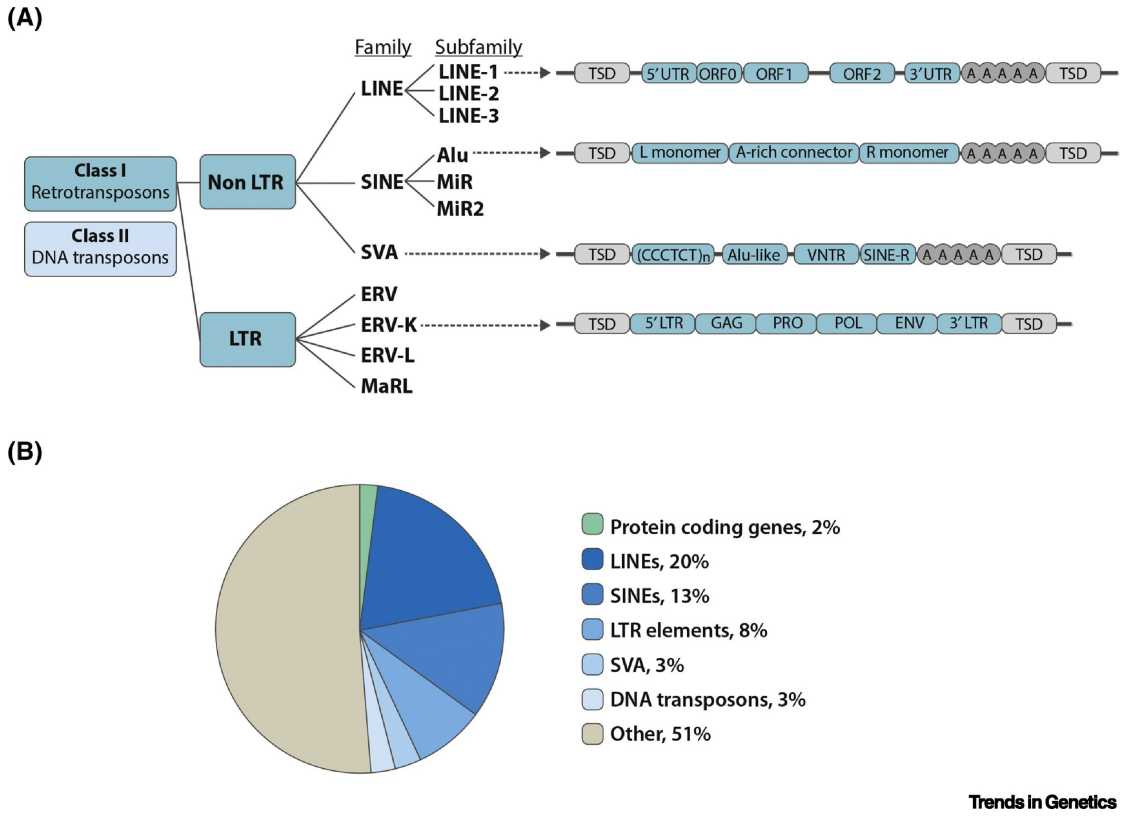


FIGURE 1.15 – **A.** Classification et structure générale des éléments transposables (TE). Les éléments complets de la famille des Long interspersed Nuclear Element (LINE)-1 contiennent deux cadres de lecture ouverts, ce qui en fait la seule famille autonome active sans Long Terminal Repeat (non-LTR) chez les humains : l'ORF1 code pour une protéine de liaison à l'ARN et l'ORF2 pour une protéine ayant des activités de transcription inverse. De plus, les LINE-1 codent pour un ORF0 récemment découvert, spécifique aux primates, avec un promoteur de la région 5' non traduite (UTR) en sens et en antisens. Les éléments Alu contiennent deux monomères séparés par un connecteur riche en A en tant que 5' UTR et une queue poly-A dans le 3' UTR. En outre, ils contiennent deux promoteurs pour l'ARN polymérase III dans leur région 5'. Les ERV-K complets, le reste du groupe autonome LTR, contiennent deux LTR séparés par les gènes viraux gag, pro, pol et env. **B.** Le génome humain est constitué d'au moins 47 % de TE, tandis que les gènes codant pour des protéines ne représentent que 2 %. Adapté de Jonsson et al., Trends In Genetics, 2020 [88].

De plus, les centromères et télomères sont principalement composés de séquences dites satellites ou microsatellites répétées en tandem, répétant des motifs précis de 5 à 171 paires de bases parfois plusieurs centaines de fois. Ce type de répétitions simples ou de microsatellites d'un court motif de moins d'une dizaine de paires de

bases apparaissent également ailleurs dans le génome et représentent une grande fraction des éléments répétés. Enfin, les séquences à faible complexité, possédant un contenu en AT ou GC très élevé (plus de 85% dans RepeatMasker [173]) sont également considérées comme séquences répétées.

TABLEAU 1.1 – Couverture des éléments répétés en paires de bases et en fraction des éléments répétés pour le génome T2T. Données issues de [140].

Type	**Repeat Bases (Mbp)**	**Percentage of Repeats (%)**
Percentage of repeats	1647.81	100
Long interspersed nuclear elements	631.64	38.33
Short interspersed nuclear elements	390.27	23.68
Long terminal repeats	269.91	16.38
Satellite	150.42	9.13
DNA	109.35	6.64
Simple repeat	77.69	4.71
Low complexity	6.44	0.39
Retroposon	4.65	0.28
rRNA	1.71	0.10

Ces éléments répétés sont le plus souvent identifiés par alignement de séquence par rapport à une base de données de référence de séquences consensus d'élément répétés, le plus souvent en utilisant l'outil RepeatMasker sur la base de données RepBase [13].

1.3.9 Les modifications épigénétiques

La composition en transcrits d'une cellule, dit transcriptome, lui donne à terme ses fonctions biologiques, et permet de caractériser son type cellulaire et la réponse à son environnement. Cependant, afin d'obtenir une variété de transcriptomes distincts, l'information génétique de régulation de la transcription doit être interprétée de façon différente en fonction du contexte cellulaire et environnemental. Afin de moduler l'activité transcriptionnelle et régulatrice, l'ADN et la chromatine subissent des modifications chimiques appelées couramment marques ou modifications épigénétiques, influençant l'accès des TF et de la Pol II à l'ADN.

Les cytosines de l'ADN peuvent subir un processus nommé méthylation, qui consiste en un ajout d'un groupement méthyle sur leur cinquième base carbonée (5-méthylcytosine), voir Figure 1.16. Cette méthylation apparaît fréquemment dans le contexte d'un dinucléotide CpG, qui est l'enchainement d'un nucléotide cytosine

suivi d'une guanine sur l'un des deux brins. Lors de la réplication de l'ADN, cela permet la conservation de la méthylation par la méthylation du dinucléotide CpG complémentaire, par le biais d'une méthyltransférase[134]. Chez les mammifères, jusqu'à 80% des CpG possèdent une méthylation de leur cytosine. Bien que la fréquence du dinucléotide CpG soit inférieure à celle attendue par un mélange aléatoire des bases du génome humain, certaines régions du génome, nomées îlots CpG, possèdent une concentration très forte en CpG. Ces îlots CpG sont très fréquemment observés au niveau des promoteurs et des éléments régulateurs du génome. De manière générale, la méthylation est associée à une répression de la transcription au niveau des promoteurs et enhancers, tandis qu'elle est plutôt importante dans le corps des gènes transcrits. Celle-ci affecte la reconnaissance des motifs d'ADN par les facteurs de transcription. Certains TFs sont capables de reconnaître spécifiquement des motifs méthylés, que leur fonction soit répressive, ou de façon plus rare et contre-intuitive, promotrice de transcription [69].

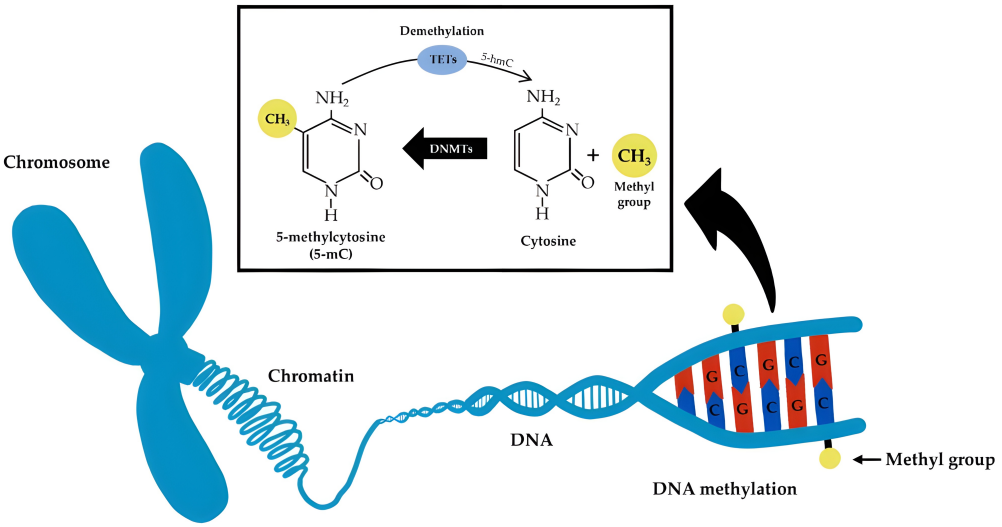


FIGURE 1.16 – Représentation schématique simplifiée de la méthylation de l'ADN. DNMTs : DNA methyltransferase. TETs : Ten-Eleven Translocation methylcytosine dioxygenases. Adapté de Valente et al., Nanomaterials, 2023 [182].

Les protéines histones, compactant l'ADN, peuvent également subir des modifications affectant la compaction de l'ADN et donc son accessibilité. Les histones possèdent un long domaine N-terminal dépassant du nucléosome subissant fréquemment des modifications chimiques, et ce particulièrement pour les histones H3 et H4 : on parle de marques d'histones (voir Figure 1.17). Plusieurs modifications peuvent s'effectuer sur les résidus d'acides aminés, à l'aide d'enzymes et de complexes protéiques ; principalement la méthylation, l'acétylation et la phosphorylation. Une nomenclature a été développée pour noter les modifications des histones : premièrement le nom de l'histone, suivi de l'acide aminé modifié dans son appelation à simple lettre, ainsi que sa position dans la protéine par rapport au domaine N-terminal, puis le type de

modification [12]. Par exemple, H3K27Ac correspond à une acétylation de la lysine en position 27 de l'histone H3.

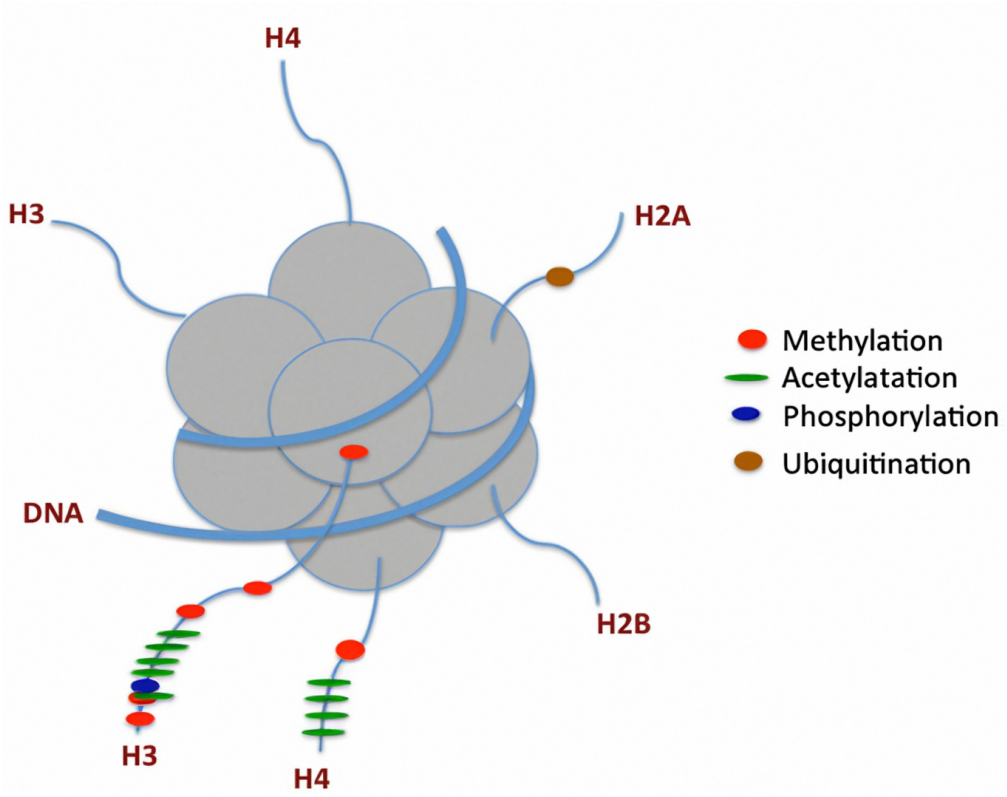


FIGURE 1.17 – Hexamère d'histone et modifications N-terminales. Adapté de Nayan et al., New Frontiers in Research : Reproductive Epigenetics, 2015 [137].

Cela permet une grande diversité dans la modification des histones. Cependant, certaines marques sont spécifiques à certains éléments de régulation et à leur fonction (voir Figure 1.18). Par exemple, la marque H3K27Ac est principalement associée aux enhancers et promoteurs actifs, H3K4me3 est principalement présente au niveau des promoteurs actifs, et H3K27me3 est une marque caractéristique de la répression de la transcription et favorise la formation d'hétérochromatine. [6]. D'autre part, H3K79me2 est typiquement recontrée aux niveau des gènes en cours de transcription et des ARN polymérases II en cours d'élongation. La modification des histones a deux effets principaux : elle peut altérer la structure des nucléosomes, influençant ainsi l'accessibilité de l'ADN à la machinerie de transcription et aux facteurs de transcription. De plus, certaines protéines peuvent se lier à la chromatine et reconnaître des marques spécifiques, agissant de diverses manières, notamment en compactant l'ADN, ou au contraire facilitant le déplacement des nucléosomes pour exposer la séquence d'ADN sous-jacente, et en recrutant des activateurs ou des répresseurs transcriptionnels[207].

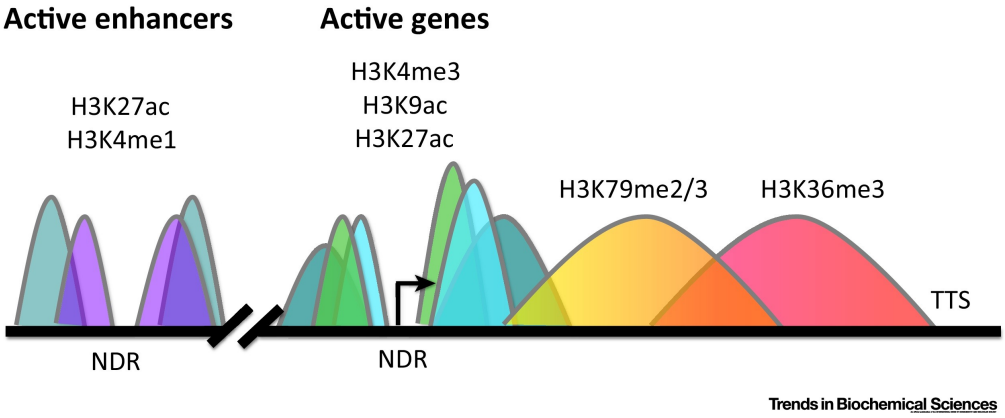


FIGURE 1.18 – Marques d'histones fréquemment présentes au niveau des enhancers, promoteurs, et transcrits en cours d'élongation. NDR : Nucleosome Depleted Region. Adapté de Gates et al., Trends in Biochemical Sciences, 2017 [64].

La combinaison des modifications épigénétiques de l'ADN serait alors potentiellement caractéristique de son état actuel ; on parle de code épigénétique. Le projet ROADMAP [20], initié en 2008, a étudié la combinaison de plus de 30 marques épigénétiques (méthylation, accessibilité de la chromatine et marques d'histones) dans plus de 111 types cellulaires à l'échelle du génome (nommés épigénomes), obtenues par ChIP-seq (voir Section 3.6), DNase-seq (voir Section 3.5), RNA-seq(voir Section 3.2), et par séquençage au bisulfite (pour la méthylation). Chaque épigénome a été segmenté en 15 à 25 catégories fonctionnelles (par exemple : enhancer, transcription, repression...) en fonction des combinaisons de marques épigénétiques observées à une base d'ADN donnée [100], à l'aide de l'outil ChromHMM [58].

Cependant, il est crucial de garder à l'esprit que les cellules et les organismes sont des systèmes extrêmement complexes et dynamiques. Les activations et désactivations successives des éléments de régulation par le biais de modifications épigénétiques, dans un ordre et un timing précis, jouent un rôle essentiel dans des processus tels que l'embryogenèse et la construction d'un organisme complet à partir d'une seule cellule initiale [62]. Parallèlement, tout dysfonctionnement dans la régulation de la transcription peut entraîner des malformations et des maladies chez les êtres humains.

1.4 Fonctions des ARNs non-codants transcrits par l'ARN Polymérase II

Historiquement, l'ARN a été longtemps perçu, à l'exception des ARN ribosomaux et de transfert, comme un simple messager intermédiaire entre l'ADN et les protéines. Toutefois, avec la découverte des transcrits non-codants et leur identification à grande échelle, révélant un nombre dépassant largement celui des ARN codant pour des protéines, il est devenu évident qu'un grand nombre d'ARNs remplissent des rôles intrinsèques, allant au-delà de leur fonction de simple messager [107]. Les ARNs présentent diverses caractéristiques notables, telles que leur conformation spatiale qui dépend de leur séquence [184], des motifs reconnus par des protéines qui se lient à l'ARN [153], et de nombreux ARNs précédemment annotés comme non-codants semblent avoir de courts cadres ouverts de lecture, permettant la production de micropeptides [124]. De plus, les ARNs à simple brin peuvent interagir avec d'autres ARNs ou ADNs à simple brin par appariement de bases complémentaires. Ces propriétés confèrent aux ARNs la capacité d'interagir avec diverses protéines et molécules cellulaires, ce qui élargit considérablement leur potentiel d'intervention dans de nombreux processus biologiques, sans pour autant coder pour une protéine.

Les longs ARNs non-codants polyadénylés, et donc stables peuvent potentiellement intéragir avec des éléments situés à longue distance de leur loci de transcription. Il a été montré par exemple qu'ils peuvent se lier au complexe de répression de transcription Polycomb, et plus de 9000 ARNs ont été identifiés comme capables de se lier au complexe PRC2 de la protéine Polycomb chez la souris par RNA ImmunoPrecipitation Sequencing (RIP-seq) [204]. Certains lncRNAs sont également capables de promouvoir l'expression de certains gènes [143] et auraient un effet enhancer. La chromatine purifiée, chez l'homme, comporte également près de deux fois plus d'ARN que d'ADN associé à la chromatine, renforçant l'hypothèse d'une régulation dépendante de l'ARN [147]. Ces ARNs peuvent également avoir des effets régulateur post-transcriptionnels [185]. Certaines de ces propriétés ont également été retrouvées dans des transcrits codant pour des protéines[99, 185].

Dans le cas des eRNAs plus courts, et plus instables, ceux-ci agiraient préférentiellement en cis, et plusieurs mécanismes semblent être identifiés pour ces ARNs. Plusieurs études récentes tendent à montrer qu'ils interagiraient avec différentes protéines du "hub enhancer-promoteur" : premièrement, ceux-ci pourraient aider à stabiliser le complexe Cohesin [108, 160] ; deuxièmement, ceux-ci seraient capables de relancer la Pol II en état de pause, en inhibant NELF [162] and contribuant à activer P-TEFb [205] ; ils pourraient également interagir avec les TFs et la machinerie transcriptionnelle recrutée au niveau du promoteur afin de moduler l'activité de la Pol II, ou maintenir un état épigénétique propice à la transcription [90]. Il est également possible que la transcription au niveau des enhancers soit en partie issue de bruit transcriptionnel de par la forte présence de TFs et co-facteurs au niveau des

enhancers. Dans ce cas de figure, l'activité transcriptionnelle pourrait contribuer à maintenir une chromatine ouverte au niveau des enhancers [76].

1.5 Dysfonctionnements de la régulation dans les maladies et phénotypes humains

Un malfonctionnement des mécanismes de régulation de la transcription va causer des changements en concentration de transcrits dans la cellule sous différentes formes : cela peut causer des variations en concentration de transcrits, l'utilisation de TSS et Sites de fin de transcription (TES) alternatifs, l'utilisation d'isoformes alternatifs des transcrits, et l'expression de transcrits devant normalement être non-exprimés. Ces altérations du transcriptome perturbent le fonctionnement de la cellule en raison du rôle intrinsèque des ARN et du protéome cellulaire, ce qui conduit finalement au développement de maladies. Des perturbations de la régulation de la transcription ont été observées dans de nombreuses affections, notamment les cancers, les maladies inflammatoires et auto-immunes, les troubles neurologiques, le diabète et les pathologies cardiovasculaires [106].

Ces maladies sont fréquemment liées à des mutations, ou à des variants génétiques. Les mutations sont des substitutions, délétions ou ajouts de paires de bases ; ou a plus grande échelle des duplications ou réarangements chromosomiques. Celles-ci peuvent être germinales, donc héréditaires et toucher toutes les cellules d'un individu, ou somatiques, apparaissant spontanément dans une cellule puis transmettant son information génétique à ses descendantes. Les variants, sont des variations génétiques répertoriées et étant relativement fréquentes dans la population (typiquement plus d'un pourcent de la population). Dans les tumeurs, les mutations sont somatiques car ne touchant que certaines cellules, même si des facteurs héréditaires augmentent la susceptibilité pour certains cancers [86].

Les dysfonctionnements régulationnels peuvent se manifester en premier lieu au niveau des protéines TFs, qui régulent l'expression d'un grand nombre de transcrits, en se fixant au niveau des enhancers, silencers ou promoteurs. Des mutations altérant une protéine TF peuvent réduire son activité de régulation, voire conduire à une perte complète de fonction. Cela aboutit à de grands dysfonctionnements au niveau de la cellule, et de l'organe. Par exemple, dans les cancers, la protéine MYC, régulant l'élongation de la Pol II, est fréquemment amplifiée, et induit une accélération du cycle cellulaire, et donc de la propagation des cellules tumorales [112] ; dans les maladies auto-immunes, la protéine AIRE est souvent mutée. Celle-ci régule l'expression des ARNs ectopiques dans le thymus, permettant une sélection négative des Lymphocytes T ciblant les protéines du soi [2].

Des mutations ou variants situées dans les régions régulatrices du génome vont au contraire affecter l'expression d'un seul ou de quelques transcrits à la fois, et ont donc en général un impact plus faible sur le phénotype. Ces variations génétiques sont considérées principalement comme des facteurs de susceptibilité. Dans les cancers, l'altération des protéines impliquées dans la réparation de l'ADN conduit à

une forte accumulation de mutations ponctuelles et de duplication (Copy Number Variation (CNV)) au niveau des enhancers, impactant au final l'expression d'un grand nombre de gènes [34]. Dans l'étude de l'impact des variations héréditaires du génome sur les phénotypes humains, à travers des Genome Wide Association Study (GWAS) (voir Section 2.3), il a été montré que la large majorité des variants associés à des traits ou maladies sont localisés dans des régions régulatrices non-codantes [187]. Ces variants ont également pu être associés à l'expression des gènes pour pouvoir confirmer leur aspect fonctionnel dans la régulation [115].

2 Séquençage de l'ADN

Sommaire

2.1 Séquençage Sanger et premier séquençage du génome humain 44
2.2 Séquençage par synthèse Illumina, dit de seconde génération (NGS) . 47
2.2.1 Introduction 47
2.2.2 Fragmentation et ajout d'adaptateurs 47
2.2.3 Amplification par Bridge PCR 48
2.2.4 Séquençage par synthèse 48
2.2.5 Analyse bioinformatique du séquençage 49
2.2.6 Limitations et séquençage de troisième génération 51
2.3 Génotypage et GWAS (Genome-Wide Association Studies) 53
2.3.1 Génotypage, déséquilibre de liaison et imputation 53
2.3.2 Héritabilité 53
2.3.3 Modèle additif d'association génotype-phénotype 54
2.3.4 Stratified LD-score regression 54

2.1 Séquençage Sanger et premier séquençage du génome humain

La première étape vers la compréhension du code génétique chez l'humain est la construction d'un génome de référence, permettant de connaître l'ordre exact des nucléotides de chaque chromosome. Cependant, bien que des variations rares et individuelles ainsi que des millions de polymorphismes (variations relativement fréquentes) existent dans la population [52], l'immense majorité du génome est identique d'un individu à l'autre. De ce fait, n'importe quel humain peut offrir un génome de référence acceptable [21]. Le projet de séquençage du génome humain a débuté en 1990 et est un projet international regroupant 6 pays : les Etats-Unis, le Royaume-Uni, la France, le Japon, la Chine et l'Allemagne. Le projet de séquençage du génome humain a utilisé des échantillons de sang provenant d'une dizaine d'individus d'origines ethniques différentes, et le génome de référence actuel est en réalité un patchwork de génomes de plusieurs individus [103, 188].

Le séquençage du génome humain a utilisé une technologie de séquençage dite de première génération : le séquençage Sanger [157]. Celui-ci utilise une méthode dite de terminaison de chaîne (Figure 2.1). Celle-ci requiert un ADN simple brin à

séquencer, une amorce ajoutée à cet ADN simple brin, une ADN Polymérase, des nucléotides ATCG normaux, et des di-deoxynucléotides triphosphate (ddNTPs), ceux-ci arrêtant la synthèse du brin complémentaire au niveau de ce nucléotide. L'ADN à séquencer est au préalable amplifié par Polymerase Chain Reaction (PCR), les amorces d'ADN sont ajoutées afin de pouvoir initier la replication, qui va s'effectuer jusqu'à ajouter aléatoirement un ddNTP, puis l'ADN est dénaturé et purifié de façon à récupérer les ADN simple brins. Les ddNTPs étant en concentration bien plus faible que les nucléotides "normaux", ce processus permet d'obtenir au moins une molécule de taille différente pour chaque base à séquencer. Les molécules subissent ensuite une migration par électrophorèse. Dans le protocole originel, quatre colonnes d'un gel en polyacrylamide sont utilisées pour chaque nucléotide : en regardant les différentes bandes de migration, il est possible d'identifier chaque molécule en fonction de sa taille, et donc la position de chaque nucléotide. Dans le protocole récent, les nucléotides sont marqués par fluorescence, et la séquence est identifiée en étudiant le chromatogramme d'une électrophorèse capillaire, permettant de réduire les coûts et d'automatiser le protocole ainsi que l'analyse [168].

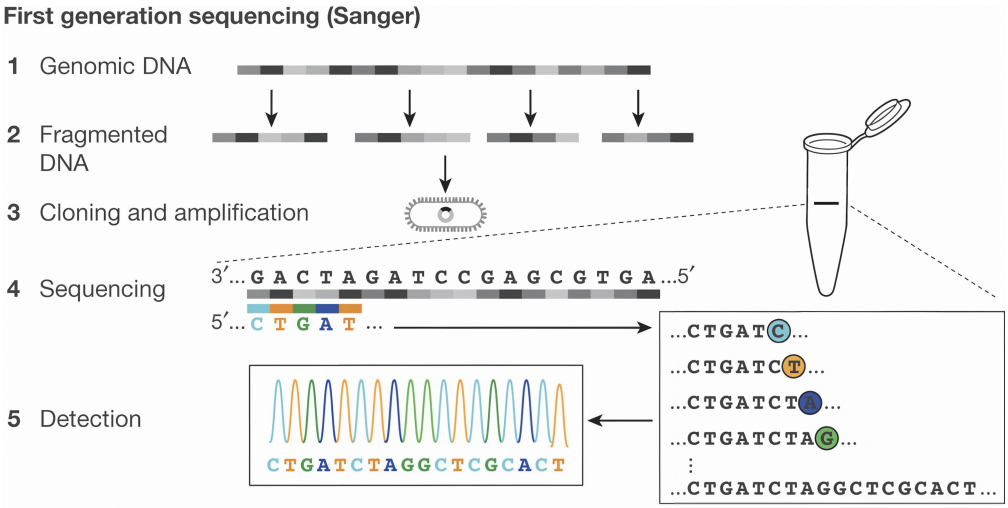


FIGURE 2.1 – Illustration du fonctionnement du séquençage Sanger. Adapté de Shendure et al., Nature, 2017 [168].

Le séquençage Sanger possède toutefois une limitation quand à la taille de la séquence à analyser : plus la séquence est grande, plus la probabilité de répliquer l'ADN jusqu'au bout de la molécule est faible et donc le signal perd en intensité. Il est donc difficile de séquencer des fragments de plus de 1 000 bp. Afin de séquencer le génome entier d'un organisme complexe tel que l'Humain, une approche dite de "shotgun sequencing" doit être utilisée [4] : le génome étudié est fragmenté en segments d'ADN de taille séquençable, puis les séquences identifiées sont agrégées à l'issue d'un processus nommé assemblage. De par la très grande quantité de données à traiter, cette étape fait appel à une approche informatique : l'assembleur étudie les recoupements entre les millions de fragments séquencés afin de les placer les uns par

rapport aux autres et reconstituer la séquence originelle.

Le projet de séquençage du génome humain a duré près de 13 ans, et coûté au total plus de 2,7 millards de dollars de l'époque [168]. La réduction drastique des coûts et du temps nécessaire au séquençage a été essentielle pour sa démocratisation dans les laboratoires de recherche. Le séquençage Sanger, bien que de très bonne qualité et peu coûteux pour de courtes séquences, a été supplanté par de nouvelles technologies de séquençage à haut débit dites de seconde génération (NGS : Next Generation Sequencing).

2.2 Séquençage par synthèse Illumina, dit de seconde génération (NGS)

2.2.1 Introduction

Au cours des années 2000, plusieurs nouvelles technologies de séquençage de seconde génération, bien moins coûteuses et bien plus rapides à effectuer ont émergé, avec notamment : le pyroséquençage [122] et le séquençage "Ion Torrent" de 454 Life Sciences, la méthode SOLiD de Applied Biosystems, et le séquençage par synthèse d'Illumina [19]. C'est la technologie de "séquençage par synthèse" d'Illumina qui s'est imposée et qui est toujours aujourd'hui en quasi monopole dans ces approches de seconde génération, en obtenant le meilleur compromis entre coût, qualité et longueur des reads, et temps de préparation [151]. Cette méthode s'effectue en quatre étapes : fragmentation et ajout d'adaptateurs, amplification, séquençage, et enfin l'analyse des séquences produites [19].

2.2.2 Fragmentation et ajout d'adaptateurs

Dans cette étape, souvent appelée préparation des librairies, l'ADN préalablement purifié subit une fragmentation, suivi d'un ajout de séquence à chaque extrémité des fragments d'ADN. Cette séquence contient trois éléments, de l'extérieur vers l'intérieur du fragment : une région de fixation qui va permettre à l'ADN simple brin de s'appareiller aux oligos présents sur la "Flow Cell", où se déroulement les réactions chimiques ; un ADN dit "barcode", va permettre d'identifier un échantillon en cas de séquençage concomitant de plusieurs échantillons (afin de réduire les coûts) ; et le dernier permet l'appariement d'une amorce essentielle pour le séquençage en lui-même (Figure 2.2).



FIGURE 2.2 – Schéma d'un adaptateur de séquençage Illumina. Les séquences P5 et P7 permettent l'appariement à la flow cell, les index permettent l'identification de plusieurs échantillon en cas de séquençage simultané, les séquences Rd1 et Rd2 permettent l'appariement de primers permettant à la Polymérase d'initier la réplication, l'insert ADN correspond à la séquence d'ADN à séquencer. Provient de la documentation Illumina [80].

Deux méthodes permettent de réaliser cette étape, premièrement la méthode dite de "Tagmentation" : une enzyme transposase clive aléatoirement l'ADN purifié et ajoute les adaptateurs à chaque extrémité. Une deuxième approche par sonication consiste à fragmenter l'ADN par ultrason, puis ajouter les adaptateurs à l'aide d'une

ADN polymérase et d'une ADN ligase. Selon la quantité d'ADN disponible, plusieurs cycles de PCR peuvent avoir lieu à la fin de cette étape.

2.2.3 Amplification par Bridge PCR

Dans un premier temps, l'ADN est dénaturé en ADN simple brin, et les librairies peuvent alors s'hybrider sur les oligos de surface de la Flow Cell grâce aux adaptateurs, puis une ADN Polymérase synthétise le brin complémentaire. L'ADN est dénaturé, la Flow Cell est nettoyée, pour ne garder que le brin nouvellement synthétisé, attaché à l'oligo attaché lui-même à la Flow Cell. Chaque brin peut alors s'hybrider aux oligos de la Flow Cell à chaque extrémité. Chaque brin d'ADN forme alors un pont, puis une ADN Polymérase synthétise le brin complémentaire. L'ADN est ensuite dénaturé en ADN simple brin, répliqué et ainsi de suite. On parle ici de "Bridge PCR", dont l'intérêt est d'amplifier de façon locale l'ADN dans la Flow Cell et de former des spots ou clusters de copies de la molécule initiale (voir 2.3.4).

2.2.4 Séquençage par synthèse

Par la suite, l'ADN est de nouveau dénaturé en ADN simple brin, et dans un premier temps, seul le brin "forward" (ou 5'->3') est conservé par clivage, puis la Flow Cell est lavée. Une amorce de séquençage est hybridée sur les brins fixés à la Flow Cell, puis une ADN Polymérase est introduite avec des nucléotides modifiés marqués par fluorescence. L'ADN polymérase ajoute le premier nucléotide complémentaire du fragment séquencé. Cependant, ces nucléotides modifiés portent également un groupement bloquant la suite de la réplication. Des spots lumineux apparaissent alors sur la surface de la Flow Cell, dont la couleur correspond au nucléotide ajouté. Chacun d'eux correspond à une des molécule d'ADN originelle, amplifiée localement de nombreuses fois. A l'aide d'un capteur optique, le séquenceur prend une image de la Flow Cell. La Flow Cell est alors lavée, le groupement fluorescent ajouté à l'ADN est retiré, ainsi que le groupement bloquant, permettant à la Polymérase de continuer la réplication (voir 2.3.5).

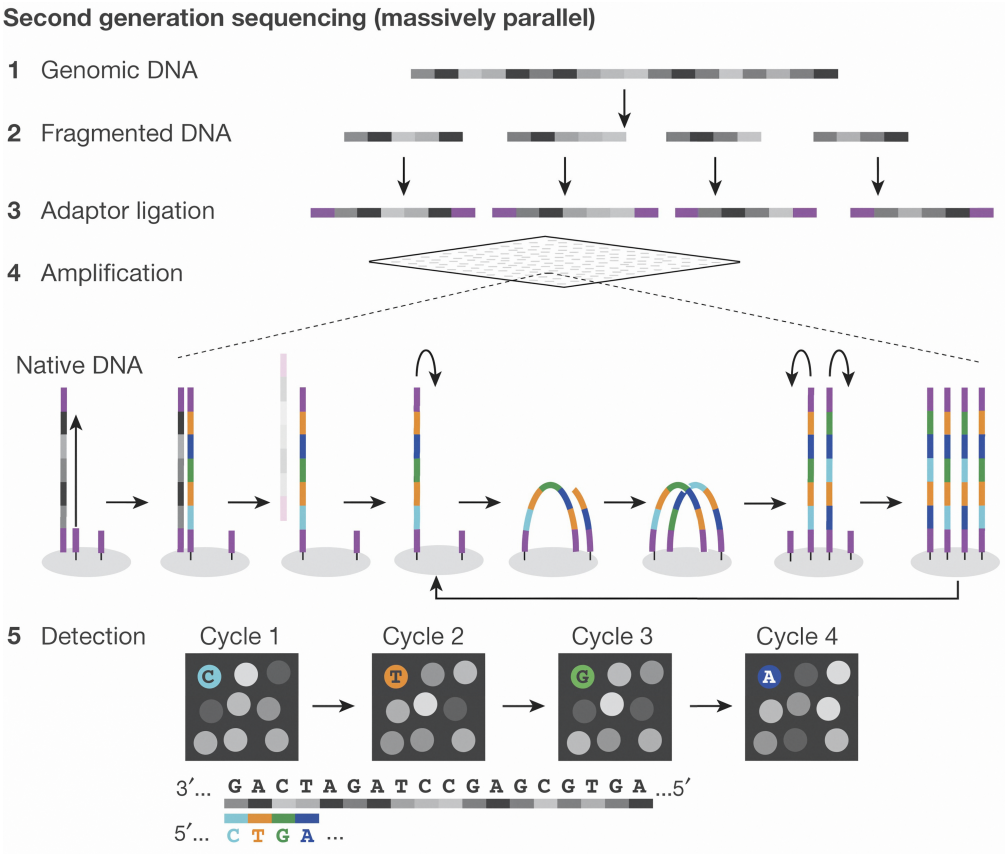


FIGURE 2.3 – Illustration du fonctionnement du séquençage par synthèse Illumina. Adapté de Shendure et al., Nature, 2017 [168].

Le suivi de la coloration de chaque spot sur la Flow Cell permet alors de reconstituer une séquence, et ce processus est répété jusqu'à obtenir une longueur de chaîne de nucléotide lue ou "read" suffisante. Chaque cycle de réaction s'opérant sur plusieurs millions de spots simultanément, cette technologie permet donc de séquencer très rapidement et à coût modéré des génomes complexes.

2.2.5 Analyse bioinformatique du séquençage

Cette technologie produit des millions de reads (ou séquences lues), et une notation de qualité de chaque base en fonction de la qualité de la lecture du spot, principalement affectée par un problème de "déphasage" à l'intérieur d'un cluster d'amplificats, car les nucléotides modifiés ne sont en réalité pas parfaitement bloquants. Il se peut donc que parfois, il y ait plusieurs nucléotides insérés au lieu d'un seul pour une ou plusieurs des molécules d'ADN amplifiées d'un cluster à chaque cycle de séquençage, ce qui créée un décalage dans la lecture. Au fur et à mesure des réactions, la fraction des ADNs déphasés augmente et baisse la qualité de lecture en fin de read [151]. La première étape est donc une phase dite de "trimming" : les séquences des adaptateurs sont retirées des reads, tout comme les bases de

trop faible qualité en fin de read, et de même, les reads de trop mauvaise qualité globale doivent également être enlevés. Plusieurs outils tels que Trimmomatic [23] ou Cutadapt [123] peuvent réaliser cette étape.

Selon le type d'analyse, les reads peuvent être assemblés à l'aide d'un assembleur, tel que SPAdes [11], dans le but d'obtenir un génome de novo (Figure 2.4). Cependant, si le coût du séquençage est faible, plusieurs contraintes limitent la qualité du génome obtenu et sont discutées dans la prochaine sous-section.

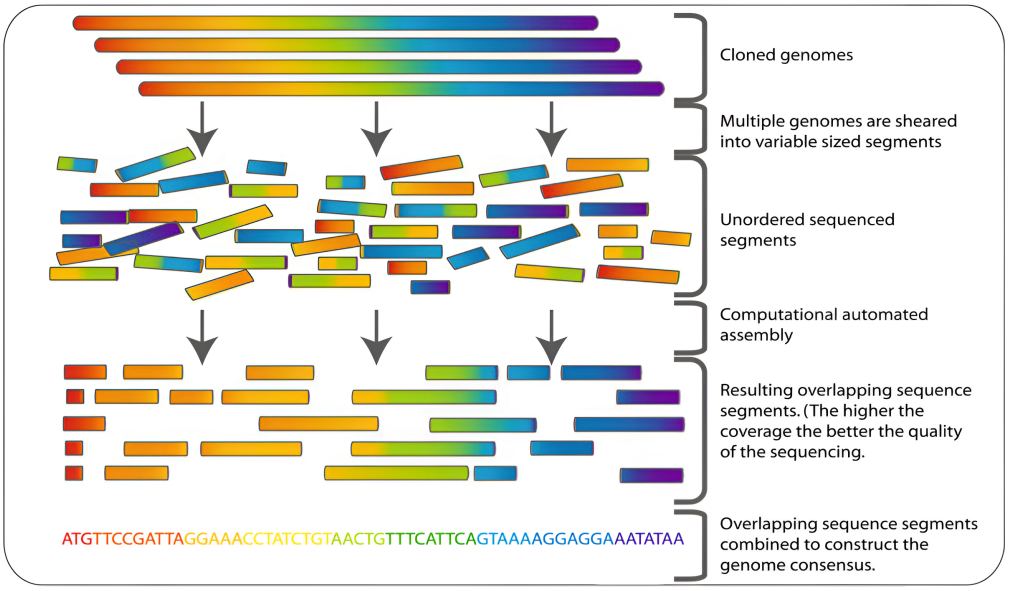


FIGURE 2.4 – Assemblage du génome. Adapté de Commins et al., Biological Procedures Online, 2009 [39].

Les reads obtenus peuvent être également alignés sur un génome de référence. Brièvement, cette étape consiste à trouver l'emplacement d'origine du read sur ce génome, sans procéder à l'assemblage (Figure 2.5). Cette étape est essentielle dans les protocoles de séquençage fonctionnels (abordés dans le prochain chapitre), afin d'utiliser ces reads en tant que compteurs moléculaires sur un génome de référence.

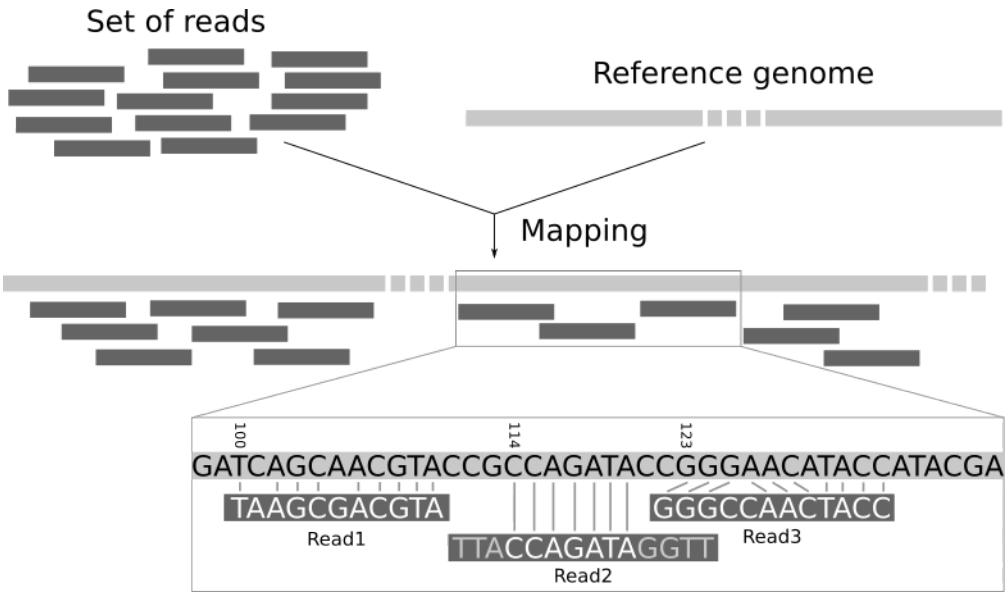


FIGURE 2.5 – Alignement de reads. Adapté de [195] (Auteurs : Joachim Wolff, Bérénice Batut et Helena Rasche).

2.2.6 Limitations et séquençage de troisième génération

Cependant, le séquençage par synthèse d'Illumina limite la taille maximale des fragments d'ADN qu'il est possible de séquencer à environ 300 paires de bases, principalement à cause du problème de déphasage inhérent à la technologie. Cela est problématique dans certaines portions de génome fortement répétées, notamment au niveau des centromères et télomères des chromosomes, où il est impossible d'assembler les fragments correctement. De même, il est dans ce cas difficile de placer les fragments d'ADN sur un génome de référence, on parle de reads "multi-mapping" car pouvant être alignés à différents emplacements sur le génome. Enfin, cette approche nécessite de nombreuses phases de PCR, pouvant créer des biais de séquence, notamment au niveau de séquences riches en GC [66].

Des nouvelles approches émergentes, dites "long read" ou de troisième génération, permettent de s'affranchir de ces limitations [114]. Elles permettent de générer des reads pouvant mesurer des dizaines de milliers voire millions de paires de bases, ainsi que de s'affranchir de la PCR et des biais de séquence qu'elle cause. Cependant, jusqu'à récemment, celles-ci possédaient un fort taux d'erreurs par base séquencée, mais des progrès au niveau de la chimie ainsi que sur le traitement informatique des données ont permis d'améliorer grandement la précision de l'identification des bases. De même, ces technologies restent plus coûteuses que le séquençage par synthèse, bien que leur coût soit raisonnable et en décroissance. Ainsi, l'avènement de ces nouvelles technologies ont permis de finaliser le séquençage d'un génome humain entier [140]. Deux principales technologies "long read" coexistent aujourd'hui : la technologie SMRT de Pacific Biosciences, ainsi que la technologie à Nanopores

d'Oxford Nanopore Technologies.

2.3 Génotypage et GWAS (Genome-Wide Association Studies)

Le projet de séquençage du génome humain a permis d'obtenir un génome de référence. Cependant, le génome de chaque individu est légèrement différent l'un de l'autre, et ces petites différences peuvent potentiellement expliquer certains traits ou phénotypes communs humains tels que la taille, ou des facteurs de susceptibilité pour certaines maladies.

2.3.1 Génotypage, déséquilibre de liaison et imputation

Suite à la finalisation d'un premier génome de référence, plusieurs projets, dont le projet 1000 génomes, ont séquencé des milliers de génomes dans plusieurs populations afin d'identifier des polymorphismes, c'est à dire des variations relativement fréquentes entre individus. Cela a permis d'identifier plus de 15 millions de Single Nucleotide Polymorphism (SNP) présents dans plus d'un pourcent des individus, représentant plus de 90% de la variabilité génétique humaine[52]. Afin de réduire les coûts de génotypage, des microarrays sont fréquemment utilisés en lieu et place d'un séquençage complet du génome pour identifier les allèles des SNPs portés par les individus.

De plus, il n'est pas nécessaire de génotyper chaque SNP : en effet, les allèles des SNPs sont corrélés entre eux par un phénomène appelé déséquilibre de liaison (LD). Celui-ci est causé par divers facteurs, avec en particulier la dérive génétique et la recombinaison génétique [172]. Ainsi, en exploitant les corrélations entre allèles, il est possible d'imputer les SNP manquants, réduisant davantage le coût du génotypage. Ainsi, le génotypage d'une fraction des SNPs permet de caractériser en grande partie le génome d'un individu à faible coût, et donc de génotyper de grandes cohortes. Cela permet d'identifier des SNPs associés avec des traits ou maladies à travers des GWAS [138].

2.3.2 Héritabilité

L'héritabilité au sens large H^2 est une mesure de la fraction de la variabilité du phénotype expliquée par la variabilité génétique, l'autre fraction étant expliquée par l'environnement. Elle comprend les effets additifs, et les interactions entre allèles (dominances et interactions épistatiques entre allèles) [197]. Formellement :

$$H^2 = V_G / V_P$$

La variabilité expliquée par le génotype peut être divisée en trois sous-catégories : la variabilité expliquée par les effets additifs des allèles, la variabilité expliquée par les effets de dominance entre allèles, et enfin les interactions épistatiques entres allèles

[32].

$$V_G = V_A + V_D + V_I$$

L'héritabilité au sens restreint $h^2 = V_A/V_P$ ne comprend que les effets additifs des allèles. Dans une majorité de traits complexes et multifactoriels (par exemple, la taille ou le poids), les effets additifs représentent une grande fraction de l'héritabilité au sens large. A l'opposé, les traits ou maladies rares sont généralement causés par un seul ou quelques allèles récessifs rares. [197]

2.3.3 Modèle additif d'association génotype-phénotype

Le modèle additif suppose une indépendance entre les effets de chaque SNP, où l'effet de chacun serait purement additif sur le phénotype étudié [30] :

$$\bar{y_i} = \sum_{j=1}^{J} \beta_j x_{i,j} + \epsilon_i$$

Où $\bar{y_i}$ est le phénotype prédit de l'individu i par le modèle linéaire, $x_{i,j}$ le génotype de l'individu i pour le SNP j, β_j le coefficient (régressé) d'association entre l'allèle majoritaire et le phénotype, et ϵ_i une variable aléatoire sur l'erreur du modèle linéaire. Il est alors possible de dériver une p-value d'association entre l'allèle et le phénotype pour chaque SNP j à partir de la t-value du coefficient β_j du modèle linéaire. Le calcul des interactions épistatiques entre variants est potentiellement très coûteux en termes de calcul de par le très grand nombre de variants, et la complexité quadratique d'évaluer chaque paire de variants.

2.3.4 Stratified LD-score regression

En l'absence de biais et sous l'assomption d'indépendance des SNPs, l'héritabilité au sens restreint est définie par la somme des carrés des coefficients β_j du modèle additif. Cependant, le déséquilibre de liaison cause une corrélation entre les coefficients des SNPs corrélés : cela gonfle le nombre de SNPs significativement associés avec le trait (car étant corrélés avec le "vrai" SNP causal), ainsi que l'héritabilité estimée h^2 [30]. Ainsi, les coefficients "réels" β_j ne sont pas observables, et à l'issue d'une GWAS, les coefficients $\hat{\beta_j}^{GWAS}$ observés sont [30] :

$$\hat{\beta_j}^{GWAS} = s_j + \sum_{k=1}^{J} \beta_k r_{x_{i,j},x_{i,k}} + \epsilon_j$$

Avec $r_{x_{i,j},x_{i,k}}$ la corrélation entre les allèles des SNPs x_j et x_k, et s_j un biais potentiel causé par une stratification de population ou des liens de parenté.
Bulik-Sullivan et al. [30] ont dérivé, à partir de l'équation précédente :

$$E(\chi_j^2) = 1 + N\alpha + \frac{Nh_{SNP}^2}{M} l_j$$

$$l_j = \sum_{k=1}^{J} r^2_{x_{i,j},x_{i,k}}$$

Avec l_j le score LD du SNP j, h^2_{SNP} l'héritabilité du trait étudié basée sur les SNPs, N le nombre d'individus étudiés, M le nombre de SNPs et α une mesure de facteurs confondants. χ^2_j correspond au carré de la t-value d'association (ou z-score, puisqu'il est possible d'utiliser l'approximation normale car le nombre d'individu est généralement élévé) pour le SNP j.
L'équation peut être reformulée pour mettre évidence la régression linéaire simple à résoudre (LD-score regression (LDSC)) :

$$E(\chi^2_j) = a + bNl_j$$

Cela permet d'identifier les facteurs confondants ($a = 1 + N\alpha$) et l'héritabilité au sens restreint ($b = h^2_{SNP}/M$). Cette méthode est facilement applicable car elle ne nécessite que le nombre d'individus génotypés, les déséquilibres de liaison pour une population donnée et les z-scores ou p-values de chaque SNP testé pour une GWAS donnée.

L'héritabilité au sens restreint peut également être calculée pour un sous-ensemble de régions génomiques C (notée h^2_C) et donc pour le sous-ensemble de SNPs contenu dans ces régions : on parle de Stratified LD-score regression (S-LDSC). Le ratio $\frac{h^2_C/M_C}{h^2/M}$ mesure alors un enrichissement en héritabilité h^2 pour ce sous-ensemble génomique précis. Ainsi, il a été montré que les régions enhancer possédaient un fort enrichissement en héritabilité au sens restreint pour plusieurs traits, dont la taille ou la schizophrénie [60].

3 Approches de séquençage fonctionnel à haut débit

Sommaire

3.1 Introduction . . . 57
3.2 Identification du génome transcrit : RNA-seq . . . 58
3.2.1 Protocole expérimental simplifié . . . 58
3.2.2 Traitement bioinformatique des données . . . 59
3.2.3 Limitations . . . 60
3.3 Identification des départs de transcription : CAGE . . . 62
3.3.1 Protocole expérimental simplifié . . . 62
3.3.2 Traitement bioinformatique des données . . . 63
3.3.3 Limitations . . . 63
3.4 Identification du génome en cours de transcription : NET-seq, PRO-seq et dérivés . . . 64
3.5 Identification des régions ouvertes de l'ADN : DNase-seq, FAIRE-seq et ATAC-seq . . . 66
3.5.1 Protocole expérimental simplifié . . . 66
3.5.2 Traitement bioinformatique des données . . . 68
3.5.3 Limitations . . . 69
3.6 Identification des régions de fixation de protéines à l'ADN : ChIP-seq et dérivés . . . 70
3.6.1 Protocole expérimental simplifié . . . 70
3.6.2 Traitement bioinformatique des données . . . 72
3.6.3 Limitations et évolutions . . . 73
3.7 Identification de séquences à effet régulateur : MPRAs et STARR-seq . . . 74
3.7.1 Protocole expérimental et analyse bioinformatique . . . 74
3.7.2 Limitations . . . 76
3.8 Séquençage par cellule unique, dit "single-cell" . . . 77
3.8.1 Protocole expérimental simplifié . . . 77
3.8.2 Analyse bioinformatique du séquençage . . . 78
3.8.3 Limitations et évolutions . . . 81

3.1 Introduction

L'avènement de protocoles de séquençage dits "fonctionnels", couplé à la démocratisation du séquençage de seconde génération à bas coût, ont permis de découvrir plusieurs millions d'éléments fonctionnels dans le génome : nouveaux transcrits, nouveaux éléments régulateurs, nouveaux TFs et TFBS. Dans ces protocoles, plusieurs étapes en amont de la préparation de la librairie à séquencer permettent de capturer des séquences d'ADN selon leur propriétés à étudier : par exemple, à l'aide du RNA-seq, il est possible de quantifier la transcription de l'ADN du génome. Les reads de séquençage servent alors de compteurs moléculaires. Bien qu'il existe un très grand nombre de techniques différentes, ce chapitre vise à présenter les protocoles les plus fréquemment utilisés et nécessaires à mes travaux, leur protocole expérimental simplifié, leur analyse bioinformatique, ainsi que leurs limitations.

3.2 Identification du génome transcrit : RNA-seq

L'objectif du RNA-seq est d'identifier les séquences transcrites du génome en séquençant et alignant l'ADN complémentaire des ARNs sur un génome de référence, ou dans un but d'annotation, par assemblage de novo [186]. L'avènement du RNA-seq a permis de découvrir de nouveaux transcrits et exons et de délimiter précisément leur limites sur le génome de référence [135]. Le RNA-seq permet également de comparer le transcriptome de plusieurs conditions de façon quantitative avec une plus grande sensibilité que les alternatives par microarrays [135, 186]. C'est aujourd'hui l'un des protocoles de séquençage les plus populaires avec l'ATAC-seq, de par l'intérêt de l'information qu'il génère ainsi que son coût et de la simplicité relative de son protocole. Le projet recount répertorie ainsi plus de 750 000 expériences RNA-seq réanalysées uniformément [193].

3.2.1 Protocole expérimental simplifié

Bien qu'il existe plusieurs variations du protocole expérimental, celui-ci suit de façon générale ces quatres étapes (Figure 3.1) [175] :

1. Isolation et purification de l'ARN.
2. Enrichissement pour certains types d'ARNs.
   - Les ARNs ribosomaux représentant près de 80% de la quantité totale d'ARN, ces derniers vont occuper une grande partie des reads obtenus après séquençage, au dépend des autres transcrits, ce qui va réduire leur détectabilité. Ceux-ci peuvent être retirés en grande partie à l'aide d'oligos de capture (courtes séquences synthétiques d'ADN simple brin complémentaires à la séquence d'intérêt). De même, une étape d'enrichissement pour les ARNs messagers est fréquemment effectuée à l'aide d'oligos-dT ciblant la queue Poly-A de ces transcrits. A l'aide d'une électrophorèse, il est également possible de filtrer les ARNs par leur taille.
3. Sonication et obtention d'ADN complémentaire par Transcriptase Reverse.
4. Séquençage des ADNs complémentaires.

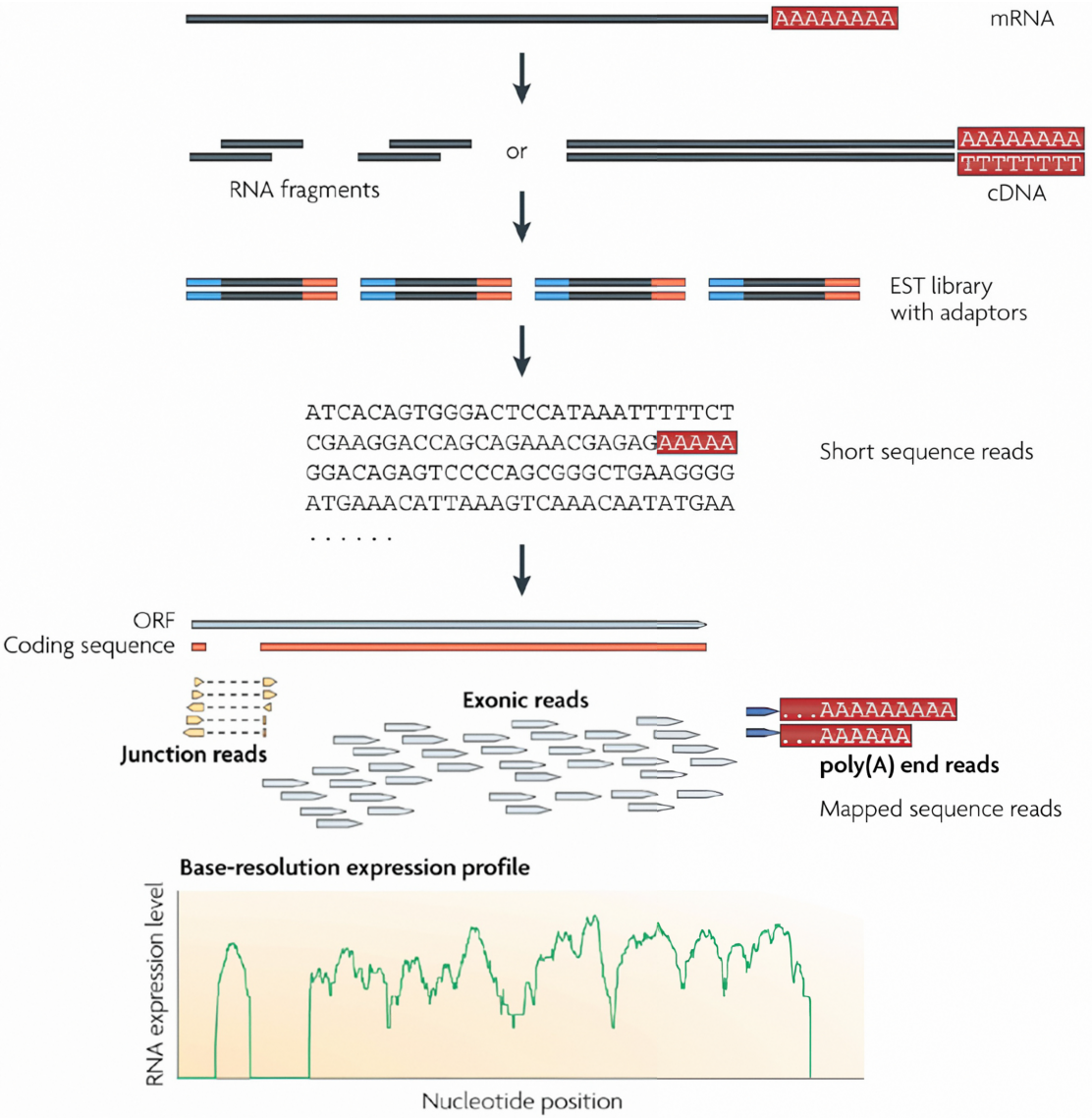


FIGURE 3.1 – Illustration du protocole de séquençage RNA-seq. Adapté de Wang et al., Nature Reviews Genetics, 2009 [186].

3.2.2 Traitement bioinformatique des données

Après séquençage, les reads sont alignés sur un génome de référence, et les régions les plus séquencées correspondent alors aux séquences d'ADN dont la transcription est la plus abondante [186]. Le plus fréquemment, le RNA-seq est utilisé pour comparer une ou plusieurs conditions biologiques avec plusieurs réplicats, et identifier les transcrits Différentiellement Exprimés (DE) [175]. Dans ce cas, les étapes du traitement bioinformatique du RNA-seq sont généralement les suivantes [41] :

1. Alignement par un aligneur détectant l'épissage des transcrits, tel que STAR [50].
2. Quantification du nombre de reads sur les éléments génomiques, via Feature-

Counts [110], et obtention d'une table de comptage de reads échantillon/gène.

- Typiquement, les éléments génomiques d'intérêt sont les exons des mRNAs, référencés dans GENCODE [61] ou Refseq.
- Ces deux premières étapes peuvent également combinées en utilisant des outils tels que Salmon [145] ou Kallisto [26], afin d'obtenir des comptages sur un transcriptome de référence.

3. Normalisation du nombre de reads de chaque échantillon.
   - Cela permet de corriger la différence de profondeur de séquençage, causée par le nombre total de reads ainsi que la différence de composition en transcrits entre les différentes conditions et réplicats. La méthode la plus fréquemment utilisée est la méthode dite de "Médiane des Ratios" de DESeq2 [116].
4. Visualisation à l'aide de Heatmap ou par Analyse en Composantes Principales (PCA).
   - Cela permet d'inspecter visuellement les données, afin d'identifier des groupes de gènes potentiellement DE entre les conditions, ou l'existence de variables confondantes.
5. Identification des transcrits DE.
   - Le nombre de transcrits à tester étant très grand, des méthodes de correction de tests multiples, telle que la méthode de Benjamini-Hochberg, doivent être utilisées afin de réduire le nombre de faux positifs. Cependant, cela réduit grandement la puissance statistique, dans des protocoles possédant très souvent de faibles effectifs de test. Des outils spécialisés, tels que DESeq2 [116] ou limma [155], permettent de maximiser la puissance statistique du test d'expression différentielle, dans un premier lieu par l'utilisation un test spécialisé (via un modèle linéaire généralisé de Binomiale Négative pour DESeq2, et une transformation par logcpm suivie d'un t-test modéré pour limma), et également en utilisant l'information contenue dans les autres gènes afin d'affiner l'estimation de variabilité.
6. Identification des processus biologiques potentiellement affectés par enrichissement fonctionnel ou Gene Set Enrichment Analysis (GSEA).
   - Cette étape permet d'identifier parmi les gènes DE, si il y a une surreprésentation d'annotation de groupes de gènes (par exemple, s'il y a plus de gènes impliqués dans l'immunité qu'attendu), et donc d'inférer les processus biologiques affectés par la condition. g :Profiler est un outil populaire pour réaliser cette tache [152].

3.2.3 Limitations

Une des limitations du RNA-seq est qu'il est relativement difficile d'identifier des transcrits courts, peu exprimés et peu stables, donc principalement les transcrits

non-codants. Il y a trois raisons principales à cela : premièrement, les transcrits longs occupent une plus grande partie des reads de séquençage au détriement de transcrits plus courts, ce qui peut être amélioré avec le protocole CAGE ou RAMPAGE (décrit dans le prochain chapitre) ; les transcrits codants sont également plus abondants ; enfin, les ARNm, plus stables, ont une durée de vie plus longue dans la cellule et s'accumulent après leur transcription, ce qui les rend plus facilement détectables. Ce dernier point est adressé dans les protocoles ciblant les transcrits naissants (voir Section 3.4).

3.3 Identification des départs de transcription : CAGE

L'objectif du protocole CAGE est d'identifier à la paire de base près les sites d'initiation de la transcription sur un génome de référence. Le principal intérêt de cette approche est de limiter le séquençage au seul début des transcrits : cela permet d'identifier précisément le site d'initiation ainsi que d'éviter l'enchevêtrement de différents transcrits rendant l'identification des TSS difficile [169]. De plus, cela permet de ne pas "gaspiller" des reads de séquençage sur le reste du corps des gènes, ce qui rend la méthode beaucoup plus sensible pour les transcrits ou TSS peu exprimés.

3.3.1 Protocole expérimental simplifié

Le protocole est assez similaire au RNA-seq, mais rajoute une étape de digestion et de capture [177] (Figure 3.2) :

1. Isolation et purification de l'ARN.
2. (Facultatif) Enrichissement en ARN PolyA.
3. Capture des ARNs possédant une coiffe 5' méthylguanosine par biotinylation.
4. Ajout d'adaptateurs et obtention d'ADN complémentaire par Transcriptase Reverse.
5. Digestion par enzyme de restriction afin de conserver les 20 à 27 premières paires de bases, selon l'enzyme utilisée.
6. Séquençage des ADNs complémentaires.

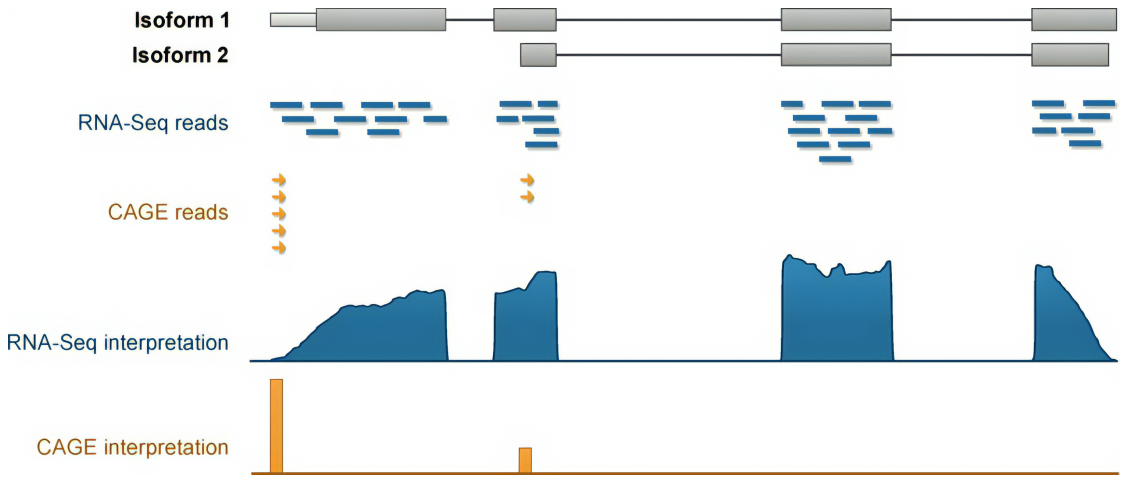


FIGURE 3.2 – Comparaison des signaux issus de séquençage RNA-seq et CAGE. Le CAGE permet d'identifier deux TSS distincts sur cet exemple. Adapté de Yu et al., Nucleic Acids Research, 2015 [200].

3.3.2 Traitement bioinformatique des données

Le traitement bioinformatique de ces données est sensiblement différent, car le protocole CAGE est généralement utilisé en l'absence d'éléments génomiques de référence. Ceux-ci sont identifiés de novo avec l'aide d'un "peak-caller", identifiant les régions du génome avec une forte concentration de reads, formant des pics par inspection visuelle des données sur un navigateur de génome [5]. Les premières étapes de l'analyse bioinformatique sont donc :

1. Alignement par un aligneur, tel que Bowtie2 [105].
   - Il n'y a pas besoin d'identifier les évènements de splicing sur de si courtes séquences.
2. Identification des TSS via Peak Calling, par exemple en utilisant Reclu ou Paraclu [142].
3. Dans le cas d'une étude différentielle, le traitement bioinformatique peut à partir de cette étape suivre le protocole de RNA-seq présenté dans le chapitre précédent, en quantifiant le signal contenu dans les pics.

3.3.3 Limitations

Toutefois, cette méthode tend à sélectionner des transcrits plutôt stables, car séquençant l'entièreté de la cellule et donc en grande partie des transcrits stables possédant une durée de vie de plusieurs heures. Des méthodes plus sensibles ciblent uniquement les transcrits en cours de transcription par la Pol II, telles que le PRO-seq, le GRO-seq ou le NET-seq. Ces protocoles possèdent des variantes PRO-cap et GRO-cap, séquençant également uniquement la partie 5' des transcrits à la façon du CAGE.

3.4 Identification du génome en cours de transcription : NET-seq, PRO-seq et dérivés

L'objectif de ces protocoles dérivés du RNA-seq est d'identifier uniquement les transcrits en cours de transcription sur un génome de référence, afin de pouvoir capturer et séquencer plus efficacement les ARNs instables. Le protocole Native Elongating Transcription sequencing (NET-seq) [37] simplifié est le suivant :

1. Immuno-Precipitation de la Pol II couplée à l'ARN naissant.
2. Isolation et purification de l'ARN obtenu.
3. Ajout d'adaptateurs et obtention d'ADN complémentaire par Transcriptase Reverse.
4. Séquençage des ADNs complémentaires.

Une des limitations de ce protocole est le manque de rendement de l'étape d'Immuno-Precipitation de la Pol II couplée à l'ARN, créant du bruit et nécessitant davantage de profondeur de séquençage. Des protocoles marquant directement l'ARN permettent un signal de meilleure qualité, tels que le Global Run-On sequencing (GRO-seq) [44], le PRO-seq [101] et le Transient Transcriptome sequencing (TT-seq) [163]. Ces protocoles diffèrent par le type de marquage utilisé, résultant en des signaux différents :

GRO-seq : La cellule est incubée avec des nucléotides U marqués 5-bromouridine 5'-triphosphate (Br-UTP). La Pol II synthétise les transcrits avec ces nouveaux nucléotides, qui sont isolés par Immuno-Precipitation avec des anticorps anti-Br-UTP [44].

PRO-seq : La cellule est incubée avec des nucléotides modifiés portant une Biotine et bloquant la continuation de la transcription. La Pol II synthétise les transcrits avec ces nouveaux nucléotides, qui sont isolés par capture par Streptavidine, protéine possédant une affinité très forte et très spécifique pour la biotine, améliorant la qualité du signal par rapport à une Immuno-Precipitation. Ici, la Pol II ne peut insérer qu'un seul nucléotide, et le PRO-seq permet donc de connaître au nucléotide près la position des complexes d'élongation [101].

TT-seq : La cellule est incubée avec des nucléotides U modifiés portant un groupement souffré (4-thiouridine ou 4sU). La Pol II synthétise les transcrits avec ces nouveaux nucléotides. Par la suite, une réaction chimique permet de substituer les groupement souffrés par une biotine. Les ARNs marqués sont isolés par capture par Streptavidine, ce qui permet une capture complète et très sensible des ARNs naissants [163].

Ces protocoles peuvent être modifiés pour obtenir un séquençage en 5' similaire au CAGE. On parle alors de GRO-cap [43], NET-CAGE [77] ou PRO-cap [121]. Ces protocoles peuvent identifier les départs de transcription avec une grande sensibilité (Figure 3.3).

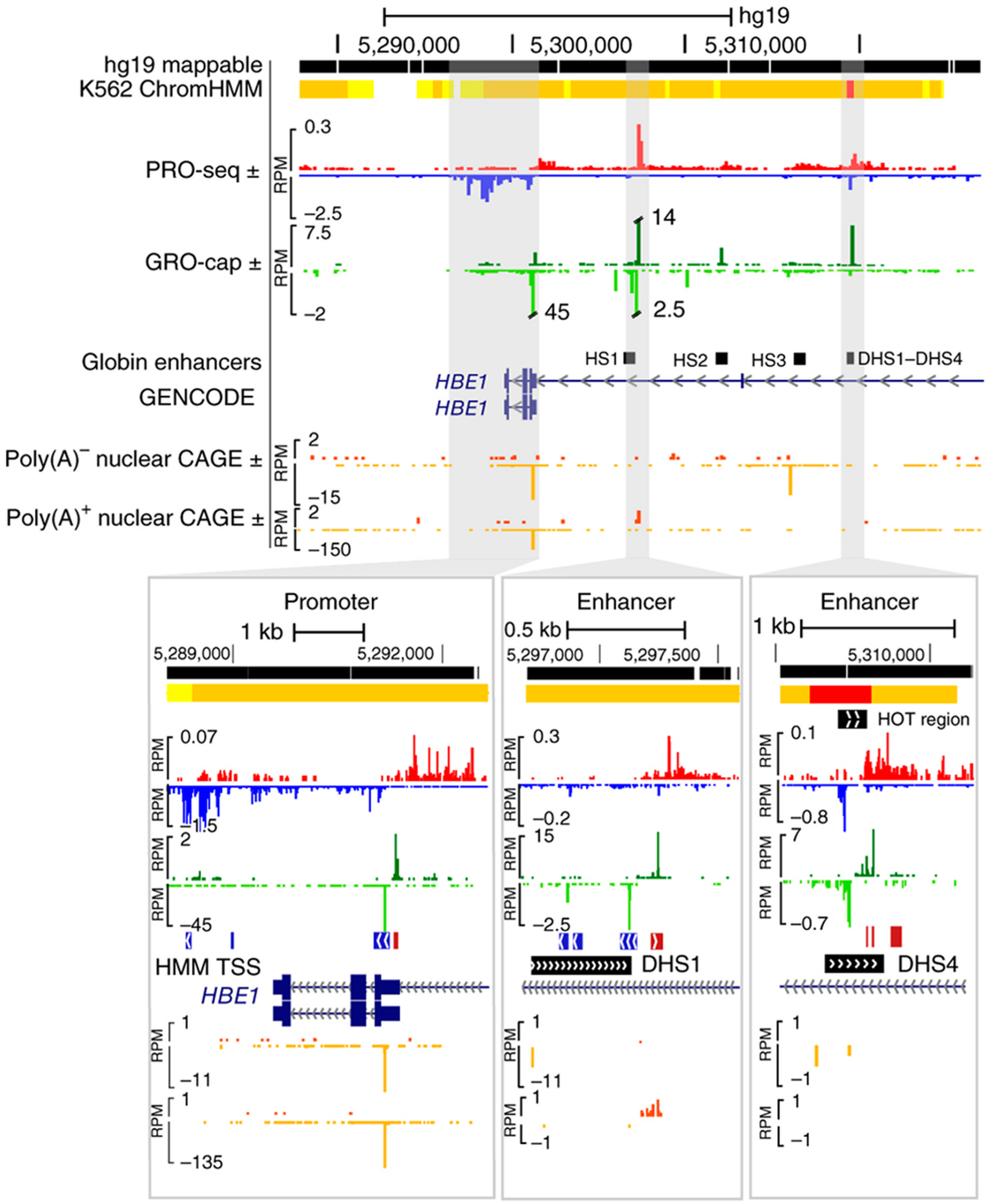


FIGURE 3.3 – Visualisation dans un navigateur de génome des signaux de PRO-seq, GRO-cap et CAGE (avec et sans sélection Poly-A). Le PRO-seq et GRO-cap identifient des transcrits instables indetectables par CAGE. Le GRO-cap identifie précisement les TSS, tandis que le PRO-seq permet d'identifier la position des Pol II, fréquemment pausées 50bp après le TSS . Adapté de Core et al., Nature Genetics, 2014 [43].

3.5 Identification des régions ouvertes de l'ADN : DNase-seq, FAIRE-seq et ATAC-seq

Les régions régulatrices actives du génome sont en général dans un état de chromatine peu condensé, avec des nucléosomes déplacés, afin de permettre aux protéines TFs et à la machinerie transcriptionelle d'accéder à l'ADN. On parle alors "d'ouverture" ou "d'accessibilité" de la chromatine. Identifier les régions ouvertes de l'ADN permet ainsi d'identifier des éléments régulateurs potentiellement actifs dans une condition biologique donnée. Les protocoles DNase I hypersensitive sites sequencing (DNase-seq) [25], Formaldehyde-Assisted Isolation of Regulatory Elements and sequencing (FAIRE-seq) [171] et Assay for Transposase-Accessible Chromatin with highthroughput sequencing (ATAC-seq) [27] visent donc à identifier ces régions ouvertes et à quantifier le degré d'accessibilité. Les protocoles DNase-seq et ATAC-seq sont les deux protocoles les plus populaires et seront abordés ici. Le protocole DNase-seq a été massivement utilisé par ENCODE et a permis d'obtenir un atlas annoté des régions hypersensibles à la DNase à travers une analyse intégrative de 733 expériences [131]. Au cours des dernières années, le protocole ATAC-seq est cependant devenu de loin le plus populaire de par ses avantages : il est bien moins coûteux en termes de temps de préparation et nécessite bien moins de matériel biologique pour obtenir un signal suffisant [28].

3.5.1 Protocole expérimental simplifié

Brièvement, le protocole DNase-seq simplifié suit ces étapes [174] (Figure 3.4) :

1. Lyse des cellules et isolation des noyaux.
2. Digestion de l'ADN par la DNase I.
   - Cette enzyme est une endonuclease créant des coupures double brin aléatoires dans la séquence d'ADN. Cependant, ces coupures sont plus fréquentes dans les régions ouvertes et non protégées par un nucléosome, qui sont facilement accessibles par l'enzyme.
3. Ajout d'un adaptateur en 5' et digestion par MmeI.
   - L'adaptateur contient le site de l'enzyme de restriction, qui coupe 20 paires de bases en aval du motif.
4. Ajout des adaptateurs de séquençage.
5. Filtrage de la taille des fragments par électrophorèse.
   - Cela permet de sélectionner uniquement les fragments possédant 86 paires de bases : les adaptateurs plus la séquence de 20bp.
6. Séquençage.

Le protocole ATAC-seq utilise une transposase hyper-active insérant au passage les inserts de séquençage au lieu d'une digestion par DNase, réduisant le nombre d'étapes et la perte de matériel génétique [27] :

1. Lyse des cellules et isolation des noyaux.
2. Tagmentation par une Transposase Tn5 hyperactive.
   - La transposase Tn5 coupe l'ADN de façon aléatoire, et ajoute un tag de 9 paires de bases ainsi que des adaptateurs de séquençage. Cependant, ces coupures sont plus fréquentes dans les régions ouvertes et non protégées par un nucléosome, qui sont facilement accessibles par l'enzyme.
3. (Facultatif) Sélection de la taille des fragments par éléctrophorèse.
   - Élimine les fragments trop longs.
4. Amplification par PCR et séquençage.

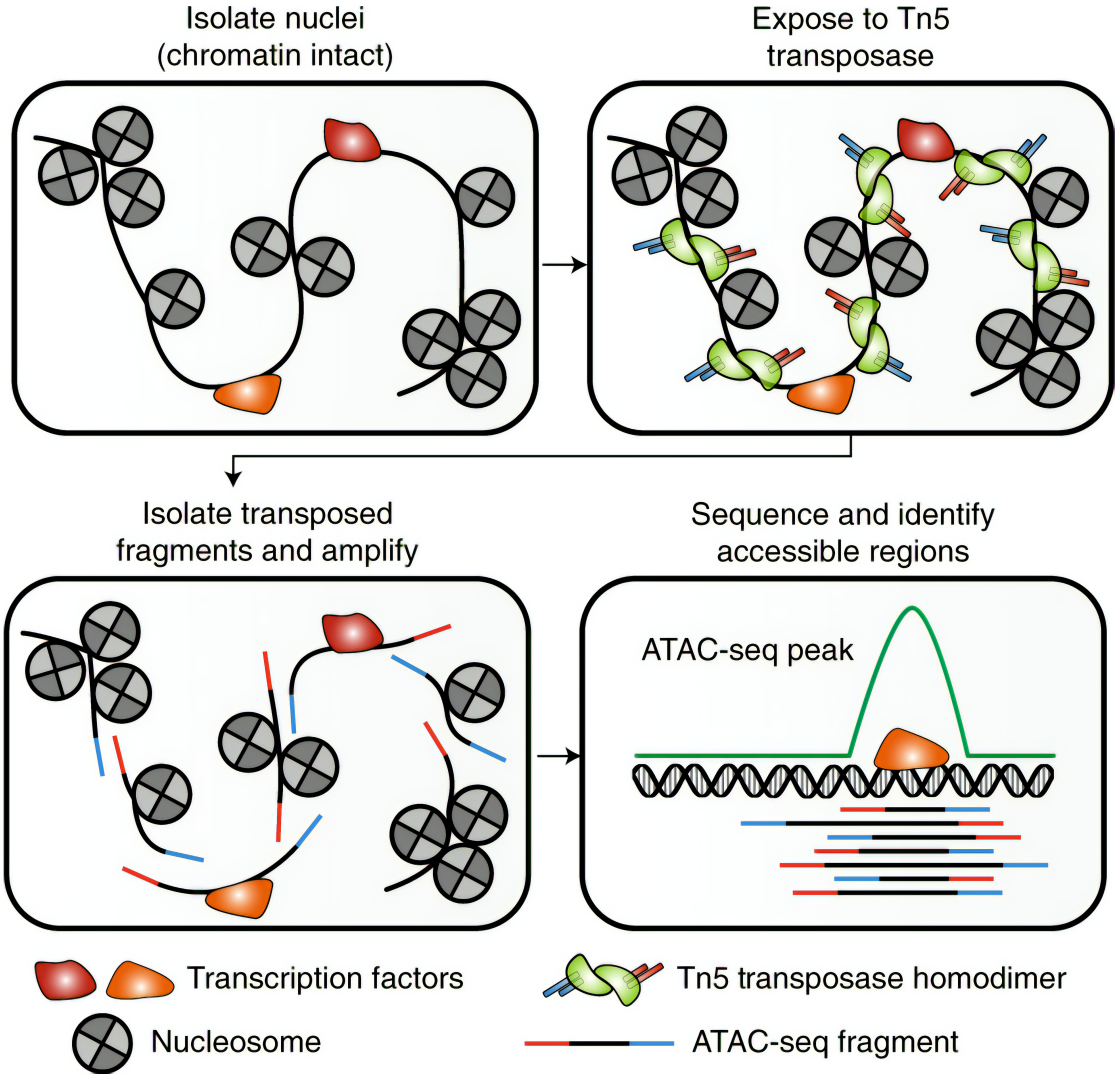


FIGURE 3.4 – Schéma simplifié du protocole ATAC-seq. Adapté de Grandi et al., Nature Protocols, 2017 [68].

3.5.2 Traitement bioinformatique des données

Les données de séquençage issues de ces protocoles forment des "pics" courts de reads positionnés sur les régions ouvertes de l'ADN. L'analyse des données des méthodes est assez similaire entre les deux approches [9, 198] :

1. Alignement par un aligneur de séquence, tel que Bowtie2 [105].
2. Identification des régions ouvertes via Peak Calling, par exemple en utilisant MACS2 [203].
   - Certains Peak Callers sont spécialisés pour les données de DNase-seq, tels que Hotspot [180] ; ou par exemple HMMRATAC [178] pour des données d'ATAC-seq.
3. (Optionnel) Identification de TFBS.
   - Les TFs réduisent l'accessibilité à l'ADN pour les enzymes de clivage, résultant en un signal plus faible dans des régions autrement ouvertes.
4. Contrôle qualité du signal.
   - Plusieurs métriques permettent de vérifier la réussite du protocole, la plus utilisée étant la fraction de reads située dans les pics (FRiP) ou les promoteurs [9].

Dans le cas d'une étude comparative entre plusieurs conditions ou tissus, le traitement bioinformatique peut à partir de cette étape suivre un protocole proche de celui utilisé pour le RNA-seq présenté dans le chapitre précédent [198] :

1. Identification de pics consensus entre expériences.
   - Cela permet d'identifier les régions de référence, concaténées à travers plusieurs expériences. Un même pic peut être déplacé ou changer de taille de quelques dizaines de paires de bases due à la variabilité entre réplicats.
2. Quantification du nombre de reads sur les pics consensus, via FeatureCounts [110].
3. Normalisation du nombre de reads de chaque échantillon, via DESeq2 [116].
   - Certaines approches suggèrent d'utiliser une fenêtre coulissante sur tout le génome au lieu de pics consensus pour évaluer l'expression différentielle, telles que csaw [118].
4. Visualisation à l'aide de Heatmap ou par PCA.
5. Identification de régions différentiellement accessibles entre les conditions, par exemple en utilisant DESeq2 [116].
6. Identification des processus biologiques potentiellement affectés par enrichissement fonctionnel ou GSEA.
   - Dans ce cas de figure, les régions ouvertes ne sont pas préalablement annotées. Des outils tels que GREAT [128] permettent d'associer ces régions à des gènes, qui eux sont annotés, afin de réaliser un enrichissement fonctionnel.
7. Identification de motifs de TFBS enrichis [75].
   - Cette étape permet d'identifier quelles sont les potentielles protéines régulatrices d'un sous ensemble de régions ouvertes d'intérêt.

3.5.3 Limitations

Les enzymes utilisées dans ces deux protocoles ont des biais de séquence, et ont tendance à cliver plus efficacement certains motifs, et ce particulièrement pour la transposase qui possède une forte affinité pour les régions riches en GC [198]. De plus, si l'ouverture de la chromatine permet d'identifier des régions potentiellement fonctionnelles, elle ne donne pas plus d'informations sur la fonction biologique réelle de l'élément régulateur identifié. Les protocoles ChIP-seq, ainsi que les MPRAs, permettent d'obtenir davantage d'informations sur la fonction biologique des éléments régulateurs.

3.6 Identification des régions de fixation de protéines à l'ADN : ChIP-seq et dérivés

Le ChIP-seq permet d'identifier le site de fixation de protéines à l'ADN, principalement les TFs et les histones [15, 87]. Ces dernières peuvent de plus être sélectionnées selon leur marquage, permettant d'observer l'état épigénétique de la chromatine sur l'entièreté du génome. Le ChIP-seq a été utilisé notamment pour identifier les marques d'histones et l'occupation de la Pol II à travers le génome dans des cellules lymphoïdes, ce qui a permis de montrer que la Pol II était présente au niveau de promoteurs de gènes très peu exprimés et probablement pausée au niveau de ces gènes [15]. Le ChIP-seq a également montré une forte occupation des enhancers par la Pol II [158]. L'intégration de plusieurs marques d'histones identifiées par ChIP-seq a également permis de caractériser l'état de le chromatine dans plusieurs types cellulaires, à travers le projet ROADMAP [100].

3.6.1 Protocole expérimental simplifié

Le protocole expérimental simplifié du ChIP-seq est le suivant [144] (Figure 3.5) :

1. Fixation au Formaldéhyde.
   - Le formaldéhyde permet de transformer les liaisons hydrogènes entre les protéines et l'ADN en liaison covalentes, renforçant leur stabilité.
2. Lyse du noyau.
3. Fragmentation de l'ADN.
   - Souvent effectuée par sonication.
4. Immuno-Précipitation (IP) à l'aide d'un anti-corps ciblant la molécule d'intérêt.
   - Cette étape récupère l'ADN fragmenté lié à la protéine d'intérêt.
5. PCR et ajout d'adaptateurs.
6. Séquençage.

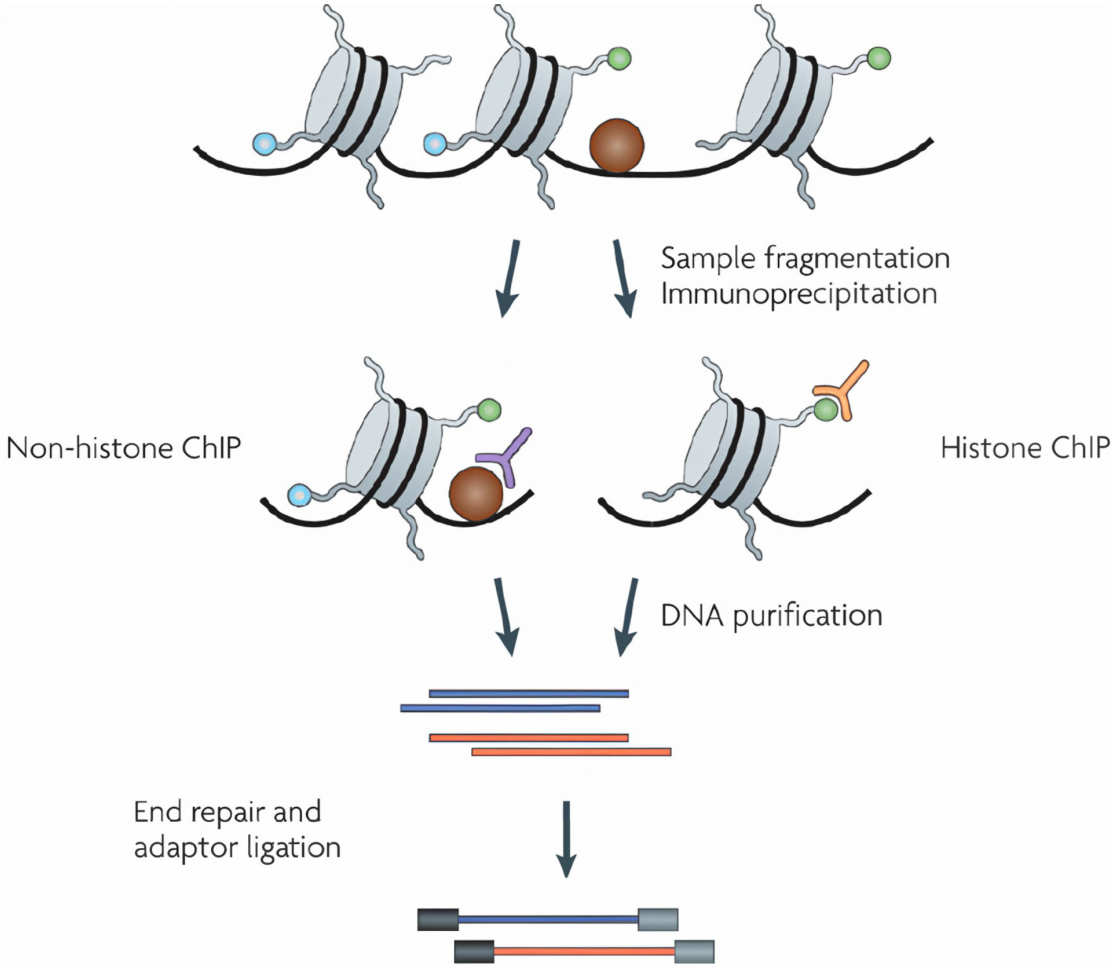


FIGURE 3.5 – Protocole ChIP-seq simplifié. Adapté de Park et al., Nature Reviews Genetics, 2009 [144].

Le protocole ChIP-seq s'accompagne fréquemment d'un contrôle nommé "Input", utilisant le protocole précédent mais sans l'étape d'Immuno-Precipitation [144]. Cela permet de prendre en compte les biais de séquence causés par la fixation et la PCR, ainsi que la fragmentation qui est inégale et affectée par l'état de la chromatine ainsi que la composition de la séquence d'ADN. Deux autres types de contrôles ont également été proposés : en réalisant une réaliser une IP sans anticorps, et en réalisant une IP en utilisant un anticorps dirigé contre l'immunoglobuline E, protéine connue pour ne pas intéragir avec la chromatine [104]. Ces deux contrôles ont été abandonnés car plus coûteux et demandant trop de matériel biologique pour obtenir une quantité d'ADN suffisante pour le séquençage, même s'ils possèdent l'avantage de contrôler les biais liés à l'immunoprécipitation. De même, l'Immuno-Précipitation est une étape avec un rendement assez faible ; selon l'affinité de l'anticorps et la réussite du protocole, 30 à 90% des reads de séquençage peuvent correspondre à du bruit du fond [109]. L'identification des régions actives est alors un problème de signal sur bruit, et donc d'enrichissement de reads IP sur reads Input.

3.6.2 Traitement bioinformatique des données

1. Alignement des reads par Bowtie2.
2. Normalisation du signal.
   - La normalisation du signal cherche à établir le zéro entre l'Input et les régions sans signal, mais bruitées de l'IP. Souvent, la normalisation équilibre le nombre de reads des deux librairies. Cependant, cela ignore la proportion de signal contenue dans l'IP, réduisant la sensibilité du Peak Calling. Des méthodes telles que NCIS [109] et SES [48] permettent d'estimer la proportion de signal et obtenir un Peak Calling plus sensible.
3. Alignement des tags et Peak Calling.
   - Les fragments obtenus par sonication sont longs et ne peuvent pas être couverts entièrement par un séquençage à reads courts Illumina, qui ne peut séquencer que les extrémités de chaque fragment. Ainsi, le signal est décalé sur les brin séquencés forward et reverse [203].
   - Le Peak Calling identifie les régions avec un enrichissement de signal par rapport à l'input. En l'absence d'Input, certains peak callers peuvent estimer le bruit de fond de façon locale à l'aide d'heuristiques. L'outil le plus populaire pour cette étape est MACS2 [74, 104, 203].
4. (Optionnel) Identification de motifs de TFBS enrichis [75].
   - Cela permet d'identifier les motifs de fixation du TF considéré ou des co-régulateurs.

Ici aussi, une étude d'occupation différentielle de l'ADN par les TFs ou marques d'histones entre deux conditions peut être menée. Cependant, inclure l'Input dans une telle analyse n'est pas trivial, et la plupart des approches actuelles utilisent un protocole similaire à l'ATAC-seq pour l'étude différentielle de pics et ignorent l'input, en supposant que les biais de séquence sont comparables d'un échantillon à l'autre [55, 118].

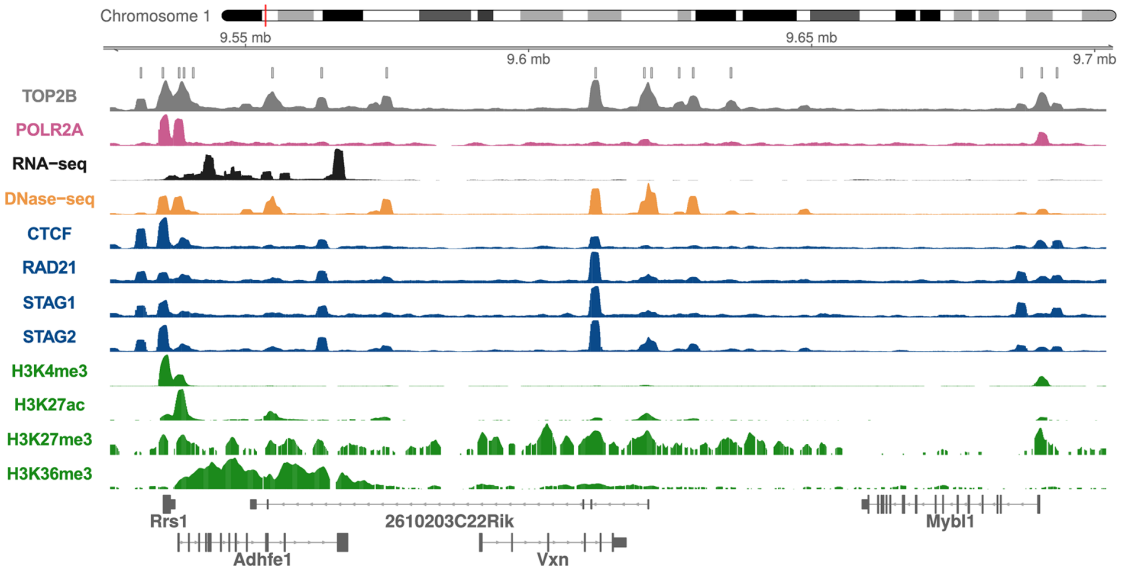


FIGURE 3.6 – Visualisation dans un navigateur de génome de signaux de RNA-seq, DNase-seq et ChIP-seq ciblant différentes protéines. Gris, Rose, Bleu et Vert : ChIP-seq. Adapté de Martinez-Garcia et al., PLOS Computational Biology, 2021 [125].

3.6.3 Limitations et évolutions

Le ChIP-seq requiert une fixation au formaldéhyde, molécule fortement toxique pouvant altérer la biologie de la cellule, et biaisant les résultats observés [14]. Il est également très dépendant de l'affinité et de la spécificité de l'anticorps utilisé [144]. Des variantes du protocole ChIP-seq existent, telles que le ChIP-exo, qui rajoute une étape de digestion par une exonucléase : cette étape permet de réduire la taille des fragments récupérés et conserver uniquement le motif de fixation lié à la protéine, tout en digérant de plus une grande partie de l'ADN bruité non-lié à une protéine [154]. Le RIP-seq est l'équivalent du ChIP-seq pour les molécules d'ARN, et permet d'identifier à quels ARNs et à quelle position une protéine d'intérêt se fixe. Le protocole CUT and Tag permet d'éviter la fixation au formaldéhyde et la phase d'immuno-précipitation, en recrutant une transposase à l'aide d'un anticorps ciblé contre la protéine d'intérêt [91]. La transposase coupe alors l'ADN préférentiellement autour de la protéine d'intérêt, et insère les adaptateurs de séquençage de façon similaire à l'ATAC-seq. Ce dernier protocole, relativement récent, à pour double avantage d'être moins coûteux, plus rapide à mettre en place et d'offrir un bien meilleur rapport signal sur bruit, même s'il hérite des biais de séquence inhérents à la transposase [92].

3.7 Identification de séquences à effet régulateur : MPRAs et STARR-seq

L'objectif de ces protocoles est d'identifier à grande échelle des séquence régulatrices parmi des séquences candidates ou à l'échelle du génome entier. Les Massively Parallel Reporter Assay (MPRA)s [130, 146] sont des adaptations des protocoles de gène rapporteur au méthodes de séquençage à haut débit, permettant de tester un grand nombre d'éléments régulateur à la fois, en associant une séquence à tester à un barcode. Le Self-Transcribing Active Regulatory Region sequencing (STARR-seq) est une variante des MPRA, où le barcode n'est pas nécessaire car la séquence d'intérêt se transcrit elle même [8]. Les MPRAs sont particulièrement indiqués pour l'étude exhaustive de l'effet des variants ou mutations dans un nombre limité d'élément régulateurs [81, 130, 146], tandis que le STARR-seq permet d'évaluer l'activité d'un grand nombre de régions candidates ou l'entièreté du génome. Le STARR-seq a d'abord été appliqué au génome de la drosophile, plus court et demandant moins de profondeur de séquençage [7].

3.7.1 Protocole expérimental et analyse bioinformatique

Le protocole général est le suivant [63] (Figure 3.7) :

1. Préparation ou synthèse des fragments d'ADN à tester.
2. (MPRAs) Ajout d'un barcode en bout de séquence et séquençage.
   - Le séquençage permet d'identifier quelles séquences sont associées à quel barcode.
3. Construction des plasmides et clonage.
   - Plusieurs constructions peuvent être utilisées selon le protocole.
4. Transfection ou infection par vecteur viral.
   - Insère les plasmides dans les cellules hôtes.
5. (MPRAs) Dans les cellules, les éléments fonctionnels altèrent la production des transcrits de GFP + barcodes.
6. (STARR-seq) Dans les cellules, les éléments fonctionnels altèrent la production des transcrits de GFP + eux-mêmes.
7. Séquençage de l'ADN des plasmides et des ARNm produits par les plasmides.
   - L'ADN des plasmides sert d'Input, de façon similaire au ChIP-seq. Le ratio de la quantité d'ARNm sur l'ADN input sera proportionnel à l'effet régulateur de la séquence.

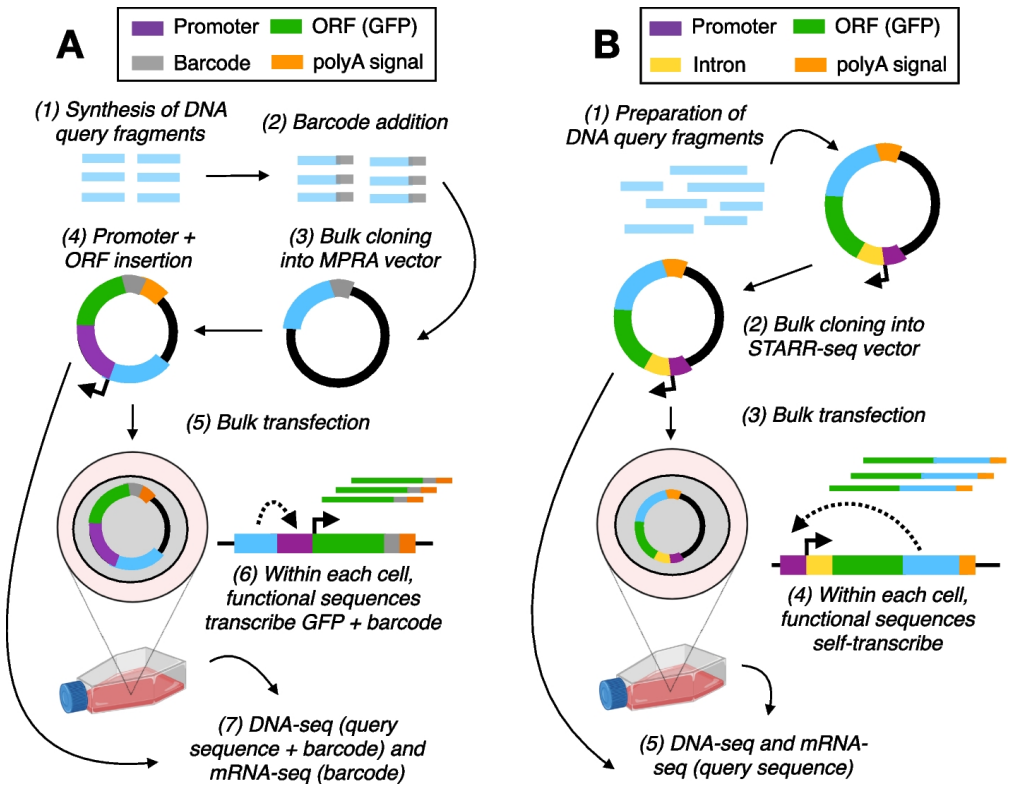


FIGURE 3.7 – Etapes expérimentales simplifiées des protocoles MPRA et STARR-seq. Tiré de Gallego Romero & Lea, BMC Genome Biology, 2023 [63].

L'analyse bioinformatique du MPRA consiste à évaluer le niveau d'enrichissement des barcodes du vecteur, en comparant la quantité de barcode sous forme d'ADN sur la quantité de barcodes sous forme d'ARNm produite, puis en reliant le barcode à la séquence d'origine. Pour cela, les séquences sont filtrées selon leur taille afin de ne conserver que le transcrit d'intérêt (GFP + barcode + queue polyA). Le STARR-seq suit un protocole similaire dans le cas d'un séquençage ciblé de régions génomiques d'intérêt ; dans le cas d'un crible du génome entier de novo, les reads d'input (ADN) et de mRNA des séquences d'intérêt peuvent être alignés sur un génome de référence et suivre un protocole proche du ChIP-seq, afin d'identifier des pics de signal correspondant à des régions enhancers, mais également potentiellement à des régions silencer [79].

Dans le cadre de la détection des enhancers, le STARR-seq à l'avantage d'être moins coûteux en temps de préparation et en budget, en évitant l'étape de synthèse des fragments, d'ajout des barcodes et de leur séquençage. De plus, en s'affranchissant des barcodes, il permet de tester une plus grande variété de séquences [63].

3.7.2 Limitations

Ces approches ont toutefois des limites techniques. Premièrement, les éléments ne sont pas testés dans leur contexte génomique : il n'interagissent pas avec d'autres éléments distaux ni le contexte épigénétique d'origine, et l'élément régulateur n'intéragit pas avec son vrai promoteur [81]. Deuxièmement, l'utilisation de la transfection ou d'un vecteur à lentivirus tend à activer des pathways de réponse immunitaire et biaise les activitées régulatrices observées [136]. Enfin, les résultats semblent divergents selon les différentes constructions de plasmides utilisées [98].

3.8 Séquençage par cellule unique, dit “single-cell”

Les protocoles décrits précédemment sont appliqués à l’échelle d’un tissu entier, ou d’un groupement de cellules, on parle de séquençage "Bulk". De nouvelles techniques permettent d’appliquer ces protocoles de séquençage de façon individuelle à chaque cellule, et donc d’observer l’hétérogénéité cellulaire des tissus. Par exemple, dans l’étude d’une maladie, il est alors possible d’identifier des variations de composition cellulaire au sein d’un tissu, ainsi que des variations de transcriptome pour certains types cellulaires. La technologie la plus populaire actuellement est développée par 10X Genomics, et l’on parle de séquençage 10X par abus de language, bien que cette approche utilise le séquençage Illumina ou puisse être adaptée à d’autres séquenceurs. Celle-ci permet d’appliquer le protocole d’ATAC-seq, CUT&TAG ou de RNA-seq à l’échelle d’une seule cellule, et ce pour des milliers à dizaines de milliers de cellules simultanément, avec un taux de capture de plus de 50% des cellules [206].

3.8.1 Protocole expérimental simplifié

Brièvement, cette technologie isole chaque cellule unique avec une bille de gel unique dans une "bulle" de solution, appellée Gel bead in EMulsion (GEM) contenant les réactifs et enzymes nécessaires au protocole de séquençage fonctionnel (Reverse Transcriptase (RT), Transposases, Polymérases...). Ces GEM sont dans une solution d’huile, ce qui crée une émulsion et permet d’isoler chaque cellule avec sa bille. Chaque bille de gel contient des oligonucléotides. Ceux-ci comprennent, dans le cas du scRNA-seq, de la position 5’ à 3’ (voir Figure 3.8de) : les oligos complémentaires de la flow cell illumina, un "barcode" de 14bp identique pour tous les oligos de la bille de gel (et donc de la cellule associée), les amorces de séquençage, un Unique Molecular Identifier (UMI) de 10bp aléatoire, et enfin un oligo-dT, permettant la capture des ARNs Poly-A et servant également d’amorce pour la RT. C’est dans les GEMs que s’effectue les réactions, dont l’appariement des ARNs avec les oligos et la Transcription Inverse. Les reads de séquençage Illumina étant relativement courts, cette approche ne capture donc qu’une partie en 3’ des transcrits Poly-A [206].

Afin d’améliorer la détectabilité du peu de molécules d’ADN complémentaire (ADNc) d’intérêt présent dans un GEM, celles-ci doivent être amplifiées par un grand nombre de cycles de PCR. Cela cause des biais d’amplification en fonction de la séquence de la molécule, et une augmentation de la variabilité du nombre de duplications d’une même molécule [85]. L’ajout d’un UMI aux fragments originels permet de compter le nombre de molécules présentes initialement avant amplification [97].

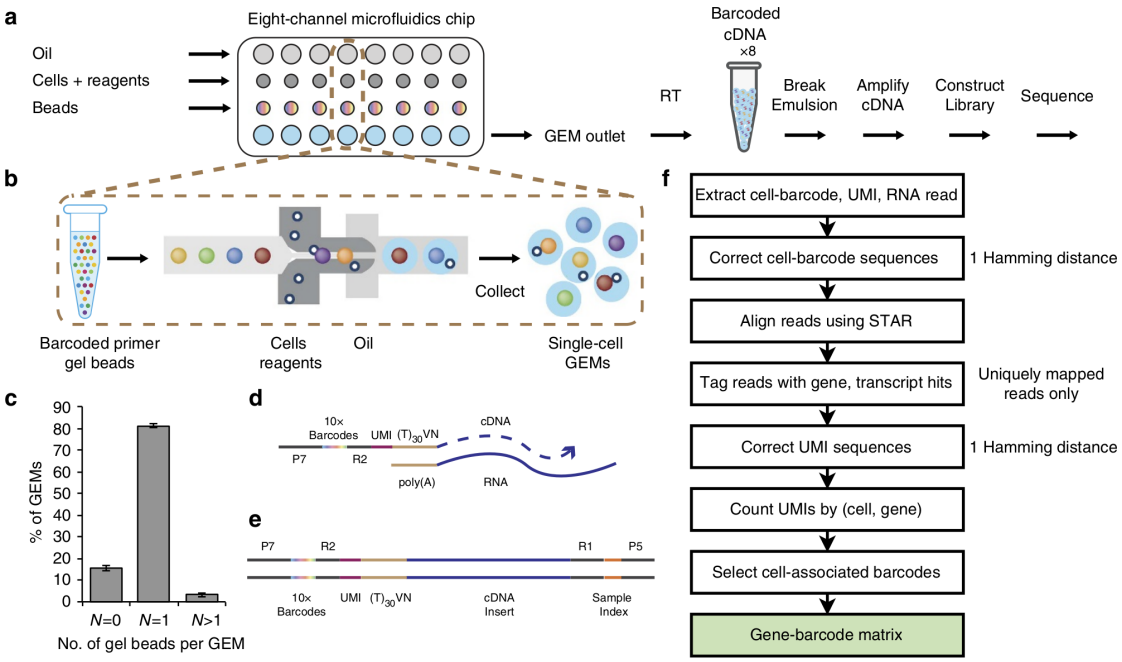


FIGURE 3.8 – Illustration du fonctionnement du séquençage scRNA-seq 10X. **a.** Processus expérimental. **b.** Appariement des billes de gel avec une cellule unique dans une émulsion (GEM). **c.** Nombre de billes de gel par GEM. **d.** Représentation schématique des séquences contenues dans les billes de gel. **e.** Représentation schématique de la librairie terminée avant séquençage en fin de réaction. **f.** Etapes préliminaires du traitement bioinformatique des données. Adapté de Zheng et al., Nature Communications, 2017 [206].

3.8.2 Analyse bioinformatique du séquençage

Ce protocole de séquençage requiert une attention particulière quand au traitement des données par informatique [206] :

1. Séparation par read des barcodes cellulaires, UMIs, et séquence d’ADNc.
2. Correction des barcodes cellulaires.
   - Les nombreuses étapes de PCR causent des erreurs de réplication. Sur les 270 millions de combinaisons de barcodes cellulaire possibles, seulement 750 000 sont en réalité utilisées. Cela permet de corriger dans la majorité des cas les barcodes erronés par un barcode correct.
3. Alignement des reads d’ADNc par STAR [50].
4. Tag de chaque read par le gène ou transcrit touché.
5. Correction des UMIs.
   - Les nombreuses étapes de PCR causent des erreurs de réplication. Il y a également bien plus de molécules d’ARN que d’UMIs valides. Cependant,

pour un même transcrit et une même cellule, la probabilité de "collision" entre deux UMIs de séquence similaire est faible. Il est alors possible de corriger l'UMI par celui le plus probable.

6. Comptage des UMIs par cellule et par transcrit.
7. Filtrage des barcodes non associés à des cellules.
   - Cela permet d'éliminer les barcodes issus de billes qui n'ont pas été appareillées à des cellules.
8. Obtention d'une matrice de comptage des UMIs par cellule/barcode et par transcrit.

Le traitement de cette matrice de comptage suit pour la plupart des protocoles des approches spécialisées, car le signal est très faible. De plus, il y a un grand nombre d'observations (cellules), avec une forte dimensionnalité des données. De façon générale, voici les étapes de l'analyse des tables de comptage 10X scRNA-seq, tels qu'effectués par le package populaire Seurat [161, 167], afin d'identifier des populations cellulaires (voir Figure 3.9).

1. Filtrage des cellules à très fort pourcentage d'ARN mitochondrial et doublons.
   - Les doublons correspondent aux GEMs pouvant parfois comporter une bille et plusieurs cellules. Celles-ci portent généralement une quantité de reads anormalement haute.
   - Les cellules à fort pourcentage d'ADN mitochondrial correspondent principalement à des cellules mortes.
2. Normalisation des comptages par cellule.
   - Les approches spécifiques au RNA-seq telles que la Médiane des Ratios ne sont pas applicables à cause du trop grand nombre de zéros. Le plus souvent, le nombre d'UMIs est alors simplement normalisé par le nombre d'UMI total par cellule. Des approches telles que scran [102] ont étées développées spécifiquement pour le single cell, afin de tirer avantage du grand nombre de cellules.
3. Transformation des comptages.
   - Brièvement, l'utilisation brute des tables de comptage est difficile pour de nombreux outils qui font fréquemment l'assomption de la normalité des données. Les comptages issus de séquençage sont au contraire bien modélisées par des distributions discrètes et sur-dispersées, telles que la Binomiale Négative [116].
   - L'utilisation de transformations dites normalisantes, telles que la transformation par log(1+x) ou d'approches plus complexes telles que SCTransform [35], est donc nécessaire.
4. Sélection de variables informatives.
   - Cette étape permet de retirer les transcrits non-informatifs possédant peu de signal ou peu de variabilité (car étant probablement uniformément exprimés entre les types cellulaires).

- Cela permet de retirer des variables bruitées, et également de réduire le temps de calcul nécessaire.

5. Réduction de dimension par PCA.
   - L'expression des gènes possède une structure de corrélation complexe. L'utilisation de la PCA permet d'identifier ces axes de variations corrélés, de réduire fortement la dimensionalité des données, et réduire le bruit.
   - L'identification du nombre optimal du nombre de composantes est complexe et plusieurs approches existent. Souvent, un nombre de composantes arbitraire faible mais suffisant est choisi; il peut également être choisi en fonction du pourcentage de variance expliquée; enfin, il peut également être obtenu par permutation des données et identification des composantes significatives, on parle d'analyse parallèle par permutation [29].
6. Identification de populations ou clustering des cellules.
   - Cette étape vise à identifier des groupes de cellules possédant des transcriptomes similaires.
   - Il existe de nombreuses méthodes de clustering. Cependant, l'approche la plus populaire est une approche d'identification de clusters dans un graphe de plus proches voisins (graphe KNN des cellules). Des méthodes d'identification approximative des plus proches voisins telles que Annoy ou NNDescent [51] permettent de construire le graphe KNN très rapidement avec une haute fidélité. Les algorithmes de Louvain [22] ou Leiden [181] peuvent alors identifier les groupes de cellules les plus fréquemment reliées entre elles dans le graphe.
   - Le clustering reste toutefois assez subjectif et dépend des objectifs : il peut être plus ou moins fin sans pour autant être incorrect. Par exemple, dans le cadre de la phylogénie, l'Humain fait partie du groupe des primates, et également à plus grande échelle des mammifères.
7. Visualisation des résultats par UMAP [126] ou t-SNE [120].
   - UMAP et t-SNE sont des méthodes de réduction de dimensionnalité non-linaires. Très brièvement, ces méthodes essaient de placer les points sur un espace à dimension réduite (ici, des points en deux dimensions), tout en conservant le plus possible une structure proche de l'espace initial. Ainsi, les cellules possédant des transcriptomes similaires sont placés de façon proche sur le graphique en deux dimensions.
8. Identification de marqueurs par cluster et annotation cellulaire des clusters.
   - L'identification des marqueurs permet d'identifier des gènes dont la surexpression est spécifique à un cluster. Le plus souvent, un test d'expression différentielle d'un cluster contre tous les autres est effectué. Les approches traditionnelles telles que DESeq2 ou limma sont toutefois très coûteuses en temps de calcul pour de larges jeux de données. Le plus souvent, un test de wilcoxon ou un t-test est alors utilisé; d'autres travaux [139, 194] suggèrent

d’utiliser les coefficients d’une régression logistique multivariée, permettant d’obtenir des marqueurs plus spécifiques.

9. Annotation cellulaire des clusters.
   - A l’aide des gènes marqueurs, une annotation de type cellulaire est associée à chaque cluster (par exemple : Lymphocyte T CD4 positif).

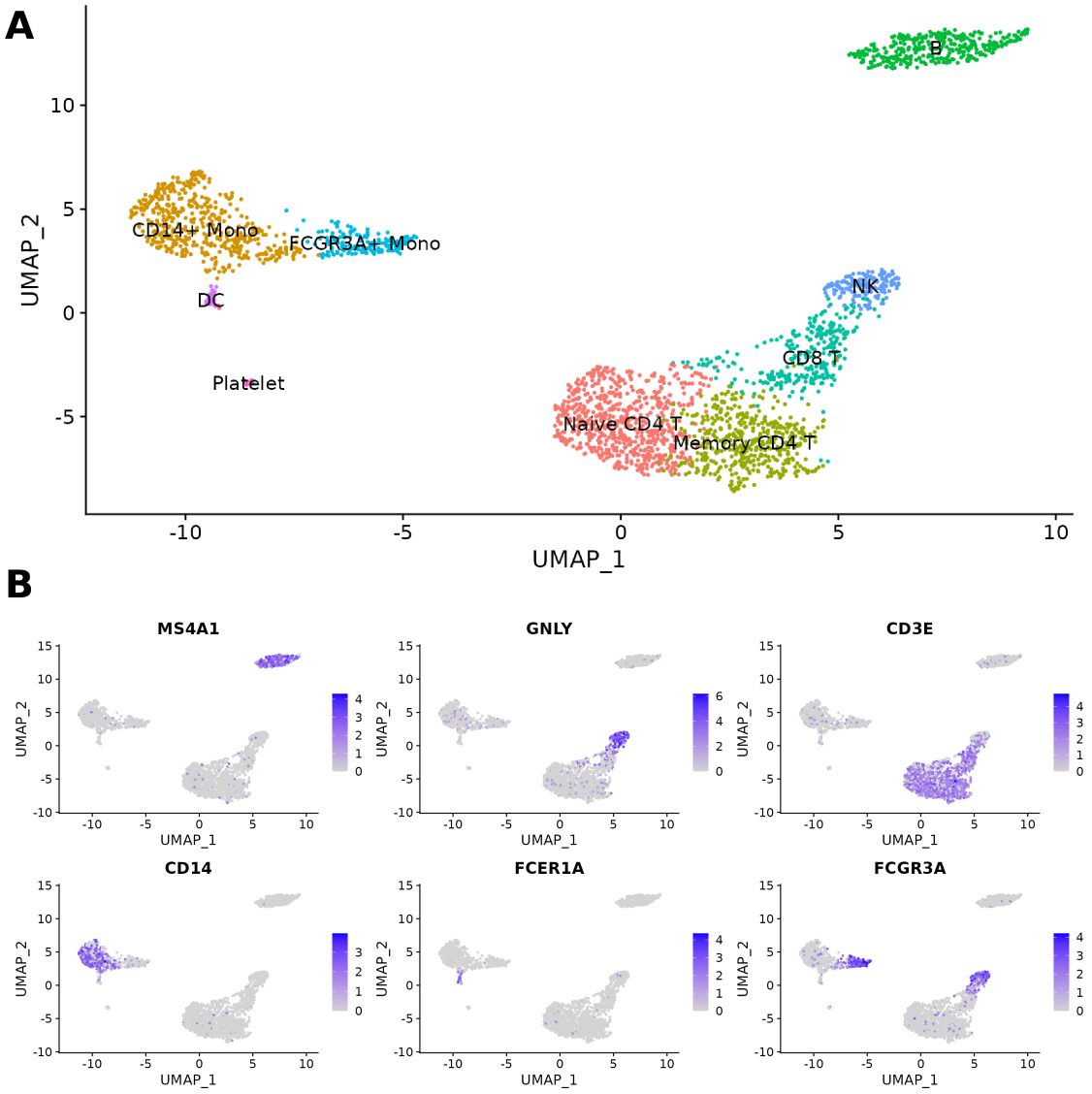


FIGURE 3.9 – Projection UMAP des transcriptomes de 2700 cellules PBMCs, analysé par l’outil Seurat. **A.** Type cellulaire inféré par clustering et analyse des marqueurs. **B.** Niveau d’expression des gènes marqueurs.

3.8.3 Limitations et évolutions

La principale limite de cette technologie est la quantité de signal faible, limitant l’interprétabilité des résultats. De plus, l’identification des types de cellules

est fortement dépendante des outils et paramètres utilisés, pouvant aboutir à des clusterings différents, et ainsi à des interprétations différentes. Enfin, tous les protocoles de séquençage ne peuvent pas être adaptés pour un séquençage single-cell.

Des techniques dites de transcriptomique spatiale permettent maintenant de quantifier le transcriptome sur un groupe de cellules ou même sur une seule cellule, tout en localisant ce groupe de cellules sur une coupe tissulaire.

4 Open Data et banques de données génomiques

Sommaire

4.1 Données Ouvertes ou Open Data, reproductibilité et réutilisabilité des données . . . 83
4.1.1 Méthodes ouvertes . . . 84
4.1.2 Données ouvertes . . . 84
4.2 Formats de données génomiques . . . 86
4.2.1 Le format FASTQ . . . 86
4.2.2 Le format FASTA . . . 86
4.2.3 Les formats SAM et BAM . . . 87
4.2.4 Le format BED . . . 87
4.3 Banques de données génomiques . . . 89
4.3.1 La base de données GEO . . . 89
4.3.2 Le consortium ENCODE (ENCyclopedia Of Dna Elements) . . . 89
4.3.3 Le consortium Genotype-Tissue Expression (GTEx) . . . 90
4.3.4 Le projet The Cancer Genome Atlas (TCGA) . . . 90
4.3.5 Le projet ReMap . . . 90
4.3.6 Le projet UK BioBank . . . 91

4.1 Données Ouvertes ou Open Data, reproductibilité et réutilisabilité des données

L'Open Science ou science ouverte a pour objectif de faciliter la transmission des résultats scientifiques, leur reproductibilité et la ré-utilisation des données générées lors du travail de recherche. Le mouvement de la science ouverte "moderne" a débuté autour des années 2000, avec l'émergence d'internet, d'une quantité en croissance exponentielle de données de recherche, et d'une "crise de la reproductibilité" avec un pourcentage croissant de travaux ne fournissant pas les éléments nécessaires à la reproductibilité des résultats de l'article ou étant simplement impossible à reproduire. La validité et la réutilisabilité des résultats obtenus est cruciale pour la science, car elle se base par essence sur les résultats des travaux précédents pour avancer. Outre le libre accès aux publications, au peer-reviewing, et à la libre diffusion des matériaux

éducatifs, la science ouverte se base également sur les concepts de méthodes ouvertes et de données ouvertes.

4.1.1 Méthodes ouvertes

Le concept de méthode ouverte vise la mise à disposition de tous les outils nécessaires pour partir d'un matériel de départ aux résultats publiés. Dans le cadre de la génomique, celle-ci comprend donc une description suffisante du matériel biologique de départ, c'est à dire,une description des tissus ou lignées cellulaires utilisées. Cela comprend également les protocoles, outils, traitements, mesures et outils informatiques (dont le code source) appliqués à ce matériel biologique. Outre la reproductibilité des résultats, afin de permettre la réutilisation des protocoles, et la comparaison de résultats, il est nécessaire de standardiser au plus possible chaque étape. Ainsi, il est nécessaire d'obtenir des identifiants, des nomenclatures communes pour les différents matériels utilisés. De même, les protocoles utilisés doivent facilement pouvoir être identifiés et suivre au plus possible un protocole standardisé afin de limiter la variabilité des résultats entre études.

Ceci s'applique également à l'analyse informatique des données, où la mise à disposition du code source, et de l'environnement logiciel utilisé permet la reproduction à l'identique des résultats, et le transfert potentiel des outils à de nouvelles données. Afin de garantir la conservation de l'environnement logiciel, le suivi des versions est essentiel : git permet de garder une trace des modifications et versions du code, et le projet conda conserve et permet une installation aisée des différentes versions de millions de logiciels et packages informatiques. L'utilisation de conteneurs tels que Docker ou Singularity permettent également de stocker l'environnement système et logiciel utilisé, garantissant la reproductibilité des résultats. Des outils tels que Snakemake [133] ou Nextflow[47] permettent également de mettre en place des pipelines reproductibles d'analyse de données.

4.1.2 Données ouvertes

Le concept de donnée est très vaste et difficilement définissable, mais peut être toutefois être résumée comme toute unité d'information qui peut être enregistrée, traitée ou transmise. Dans le contexte informatique, une donnée est typiquement une représentation numérique d'informations qui peuvent être manipulées par un ordinateur. Les données peuvent prendre diverses formes, telles que des textes, des chiffres, des images ou des sons, et sont la base de la plupart des processus informatiques. Elles peuvent être stockées dans des bases de données, traitées par des logiciels, ou transmises.

Les données FAIR sont un concept central dans le domaine de la gestion des données scientifiques et de la recherche, se référant à des données qui sont Faciles à trouver (Findable), Accessibles, Interopérables, et Réutilisables. Ce principe vise

à optimiser et à simplifier le partage et l'utilisation des données dans le milieu scientifique. Pour être considérées comme "Faciles à trouver", les données doivent être indexées dans un dispositif permettant de les rechercher, posséder un identifiant unique et pérenne et accompagnées de métadonnées détaillées. L'aspect "Accessible" insiste sur la récupérabilité des données par l'utilisateur via un protocole de communication standardisé. Cela n'empêche pas toutefois les données de n'être ouvertes qu'à certaines entités autorisées pour des raisons légales. L'Interopérabilité fait référence à la capacité des données à être intégrées avec d'autres ensembles de données, outils ou applications pour diverses analyses, nécessitant des formats standards et des langages de programmation communs. Enfin, la "Réutilisabilité" implique que les données soient suffisamment bien documentées et dans des formats standardisés pour permettre leur réutilisation dans de futures recherches. Ce cadre FAIR est devenu un standard de plus en plus adopté, favorisant une science ouverte et collaborative, et contribuant à l'avancement des connaissances dans divers domaines.

La mise à disposition des données issues de la recherche scientifique est essentielle. Dans un premier lieu, elle permet de s'assurer de la reproductibilité des résultats, et de déceler a posteriori de potentielles anomalies dans les données de départ. Leur mise à disposition permet également leur réutilisation par d'autres groupes de recherche dans d'autres projets dont des méta-analyses, ce qui permet de réduire les coûts de la recherche, et facilite la coopération entre plusieurs institutions. Le projet de séquençage du génome humain est justement un exemple concret de la réussite de la mise à disposition complète des données de la cherche dans un projet collaboratif.

Devant la quantité grandissante de données scientifiques, des entrepôts tels que Zenodo [201], ou GEO [38], dans le cadre de la génomique, ont vu le jour afin de stocker et partager de façon pérenne les données issues de la recherche. Afin de faciliter la réutilisation de ces données, l'utilisation de formats de données standardisés et facilement lisibles par des logiciels gratuits ou open source est préconisée.

4.2 Formats de données génomiques

Le séquençage produit de très grandes quantités de données à analyser par informatique. Ainsi, des formats de fichier standardisés et spécialisés ont vu le jour, afin de faciliter la mise en place d'outils d'analyse ainsi que la réutilisabilité des données en elles-même. Cette section présente les principaux formats de fichier utilisés en génomique fonctionnelle.

4.2.1 Le format FASTQ

Le format FASTQ est un fichier texte principalement utilisé pour stocker les reads de séquençage. Chaque read possède quatre attributs : son identifiant, sa séquence, son orientation (+ pour 5'->3' et - pour 3'->5'), ainsi qu'un score de qualité pour chaque base, encodé sous forme d'un caractère (voir Figure 4.1).

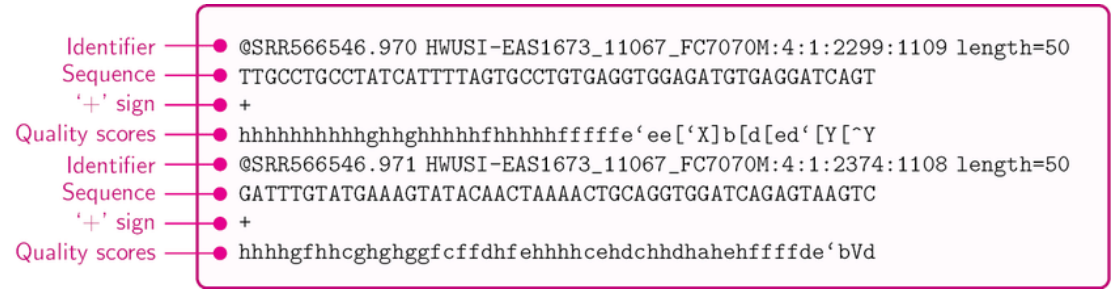


FIGURE 4.1 – Exemple d'un fichier au format FASTQ.

4.2.2 Le format FASTA

Le format FASTA est un fichier texte principalement utilisé pour stocker des séquences nucléotidiques ou d'acides aminés. Ici, chaque séquence possède son identifiant sur la première ligne, suivi de sa séquence sur les lignes suivantes. En général, 50 acides aminés ou nucléotides sont présents par ligne.

```
>gi|528476558|ref|NC_018928.2| Homo sapiens chromosome 17
TCCACTCACTGAGACAATAGACCCAAGCACATCAGCTCTGAGGCCTC
AGCACATCAGCTCTGAGGCCTCCACTCACTGAGACAATAGACCCAAG
TCACTGAGACAATAGACCCAAGCACATCAGCTCTGAGGCCTCCACTC
AACAGCTCTGAGGGCTCCACTCACTGAGACAACAGACCCAAGAACAA
>gi|528476670|ref|NC_018912.2| Homo sapiens chromosome 1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
CCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAA
ACCCTAACCCTAACCCTAACCCTAACCCAACCCTAACCCTAACCCTAACCCTAACCCTA
TAACCCTAACCCTAAACCCTAAACCCTAACCCTAACCCTAACCCTAACCCTAACCCCAA
```

FIGURE 4.2 – Exemple d'un fichier au format FASTA.

4.2.3 Les formats SAM et BAM

Le format Sequence Alignment Map (SAM) est un fichier texte utilisé pour stocker des séquences alignées sur un génome de référence. En plus d'un header, chaque ligne du fichier correspond à un read et contient :

1. L'identifiant du read.
2. Un FLAG permettant de noter certaines particularités du read, par exemple : s'il a été aligné, si c'est une paire de read de séquençage paired-end...
3. Le contig (souvent équivalent au chromosome) auquel le read a été aligné.
4. La position d'alignement du read sur ce contig.
5. La qualité de l'alignement
6. Un code CIGAR permettant de noter les mismatch, insertions ou délétions de l'alignement du read sur le génome de référence.
7. Chromosome du read pair en cas de séquençage paired-end.
8. Position du read pair en cas de séquençage paired-end.
9. Longueur de la séquence du génome de référence sur laquelle le read est aligné.
10. Séquence du read.
11. Qualité du read.
12. Tags additionnels.

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGAT *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGG   *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCT     * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCA     *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC         * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGC       * NM:i:1
```

FIGURE 4.3 – Exemple d'un fichier au format SAM.

Le format BAM est l'équivalent du format SAM, mais au format binaire et compressé, permettant un gain d'espace de stockage et de temps de lecture, et est donc préféré au format SAM.

4.2.4 Le format BED

Le format BED (Browser Extensible Data) est un format texte servant à stocker des intervalles génomiques (voir Figure 4.4). Chaque ligne correspond à un intervalle et contient les champs suivants, séparés par des tabulations :

1. Le contig (souvent équivalent au chromosome).
2. La position de départ.

3. La position de fin.
4. Le nom de l'intervalle.
5. Un score lié à l'intervalle.
6. L'orientation : + pour 5'->3', - pour 3'->5' ; . si l'intervalle n'est pas dépendant de l'orientation.
7. Position surlignée de départ.
8. Position surlignée de fin.
9. Couleur affichée sur un navigateur de génome au format R;G;B
10. Le nombre de fragments contenus dans l'intervalle.
11. Taille des blocs.
12. Positions de départ des blocs.

Il existe de nombreuses variations de ce format de fichier, qui conservent en général les 3 (BED3) ou 6 (BED6) premiers champs, et modifient les suivants, par exemple le format narrowPeak d'ENCODE utilisé pour stocker des données de peak calling, ou le format bedgraph. Le format BigBed est l'équivalent du format BED, mais sous forme binaire.

```
chr1    11873   14409   uc001aaa.3  0   +   11873   11873   (
chr1    11873   14409   uc010nxr.1  0   +   11873   11873   (
chr1    11873   14409   uc010nxq.1  0   +   12189   13639   (
chr1    14361   16765   uc009vis.3  0   -   14361   14361   (
chr1    14361   19759   uc009vit.3  0   -   14361   14361   (
chr1    14361   19759   uc009viu.3  0   -   14361   14361   (
chr1    14361   19759   uc001aae.4  0   -   14361   14361   (
chr1    14361   29370   uc001aah.4  0   -   14361   14361   (
chr1    14361   29370   uc009vir.3  0   -   14361   14361   (
chr1    14361   29370   uc009viq.3  0   -   14361   14361   (
chr1    14361   29370   uc001aac.4  0   -   14361   14361   (
chr1    14406   29370   uc009viv.2  0   -   14406   14406   (
chr1    14406   29370   uc009viw.2  0   -   14406   14406   (
chr1    15602   29370   uc009vix.2  0   -   15602   15602   (
chr1    15795   18061   uc009vjd.2  0   -   15795   15795   (
chr1    16606   29370   uc009viy.2  0   -   16606   16606   (
chr1    16606   29370   uc009viz.2  0   -   16606   16606   (
chr1    16857   17751   uc009vjc.1  0   -   16857   16857   (
chr1    16857   19759   uc001aai.1  0   -   16857   16857   (
```

FIGURE 4.4 – Exemple d'un fichier au format BED.

4.3 Banques de données génomiques

4.3.1 La base de données GEO

GEO [38] est un dépôt public qui archive et distribue librement des données de génomiques fonctionnelles à haut débit, issues de microarrays, de séquençage de nouvelle génération et autres, le tout soumis par la communauté scientifique. En plus du stockage des données, plusieurs interfaces et applications basées sur le web sont disponible pour aider les utilisateurs à interroger et télécharger les données stockés dans GEO. En septembre 2023, les données de près de sept millions d'échantillons sont accessibles sur GEO.

Cependant, si le renseignement de certaines métadonnées sont obligatoires pour procéder au stockage, il n'y a pas ou peu de standardisation de celles-ci. Ainsi, contrairement aux données d'ENCODE, il n'y a pas de standardisation des noms de réactifs utilisés, des lignées ou tissus, ni même sur les outils ou formats de fichiers utilisés. Sans inspection manuelle des métadonnées voire de l'article scientifique correspondant, il est ainsi souvent difficile d'annoter correctement une expérience. La réutilisation à grande échelle des données provenant de GEO est donc assez laborieuse de par un grand manque de standardisation des métadonnées, des données en elles-même et des méthodes.

4.3.2 Le consortium ENCODE (ENCyclopedia Of Dna Elements)

Le Consortium ENCODE [119] (Encyclopedia of DNA Elements) est une collaboration internationale en cours entre plusieurs groupes de recherche financés par le National Human Genome Research Institute (NHGRI) et initiée en 2003, à l'issue du projet de séquençage du génome humain. L'objectif d'ENCODE est de constituer une liste exhaustive des éléments fonctionnels du génome humain, y compris les éléments qui agissent au niveau des protéines et de l'ARN, ainsi que les éléments régulateurs qui contrôlent les cellules et les circonstances dans lesquelles un gène est actif. Les éléments régulateurs sont étudiés par des protocoles de séquençage fonctionnels, principalement par ChIP-seq, DNase-seq, ATAC-seq et RNA-seq, permettant un crible du génome entier pour une condition biologique donnée.

Le consortium ENCODE a réalisé de plus de 17 000 expériences de séquençage sur une vaste variété de conditions biologiques, et de protocoles de séquençage fonctionnel. Les données sont en accès libre, comprenant le séquençage brut, les fichiers intermédiaires et de résultat. Les protocoles expérimentaux, nomenclatures, et pipelines d'analyse sont standardisés, versionnés et open source.

4.3.3 Le consortium Genotype-Tissue Expression (GTEx)

Le projet GTEx [115] a étudié l'association de la variabilité génétique à l'expression des gènes, suite à l'association massive de variants non-codants à des traits phénotypiques humains [187]. Ce projet a permis la construction d'une grande base de données transcriptomique, établie par RNA-seq, associée au génotype de l'individu, obtenu par séquençage complet du génome. Au total, 17 382 échantillons de RNA-seq sont disponibles, dont 15 253 possédant un génotype appareillé, à travers 54 tissus différents et 948 donneurs sains post-mortem. Ces données ont permis d'une part le développement des méthodes de détection des expression Quantitative Trait Loci (eQTL), qui sont des variations génétiques associées à une variation de l'expression du gène, et d'autre part la détection à grande échelle de ces eQTLs dans un grand nombre de tissus.

Une des réalisations majeures du programme GTEx a ainsi été l'établissement d'un catalogue de variants génétiques qui affectent l'expression des gènes à travers de multiples tissus. Cette base de données a été une ressource inestimable pour la communauté de recherche, permettant l'évaluation de l'expression des gènes tissu-spécifique et la régulation dans de nombreux tissus différents. Le projet a également contribué à une meilleure interprétation fonctionnelle des résultats des études d'association pangénomique (GWAS) et à l'identification de gènes pertinents pour les maladies.

4.3.4 Le projet The Cancer Genome Atlas (TCGA)

Le projet The Cancer Genome Atlas (TCGA) [190], lancé en 2006, est un programme de génomique du cancer d'envergure qui a permis de caractériser moléculairement plus de 20 000 échantillons de cancer primaire et de tissus normaux appariés, couvrant 33 types de cancer. Ce projet conjoint entre le National Cancer Institute (NCI) et le National Human Genome Research Institute a réuni des chercheurs de diverses disciplines et de multiples institutions.

Le TCGA a généré plus de 2,5 pétaoctets de données génomiques, épigénomiques, transcriptomiques et protéomiques. Ces données sont désormais publiques et accessibles à la communauté de recherche. Le TCGA a notamment identifié des mutations et des dysfonctionnements dans la régulation du génome qui sont communs parmi les patients atteints du même cancer, ainsi qu'à travers différents types de cancers dans des analyses "pan-cancer".

4.3.5 Le projet ReMap

ReMap [74] est une base de données publique de sites de fixation de régulateurs transcriptionnels à l'ADN, principalement identifiés par ChIP-seq. ReMap a ré-analysé près de 20 000 expériences provenant de ENCODE, GEO et European Nucleotide

Archive (ENA), effectuées sur des tissus ou cellules couvrant les espèces Homo sapiens, Mus musculus, Arabidopsis thaliana et Drosophila melanogaster. Chacune de ces expériences a été annotée manuellement et de façon standardisée, puis a été ré-analysée de façon uniforme par le même pipeline d'analyse à partir des reads de séquençage bruts au format FASTQ.

Chez l'humain, c'est ainsi près de 1210 TFs qui ont été analysés à travers 737 cellules et tissus différents, résultant en plus de 182 millions de sites de fixation de TFs identifiés dans 8 103 expériences ChIP-seq ré-analysées uniformément et passant divers contrôles qualité.

4.3.6 Le projet UK BioBank

Le projet UK BioBank [31] a généré une cohorte unique de plus de 500 000 individus britanniques qui ont été génotypés, associés à une grande quantité d'information phénotypique, de santé et de mode de vie sur chaque participant anonymisé. Cela comprend notamment des mesures de marqueurs biologiques sanguins et urinaires, des mesures biologiques (taille, rythme cardiaque...), des données d'imagerie, et les antécédents médicaux. Cela a permis de tester l'association entre génotype et phénotype pour plus de 7000 phénotypes.

Résultats

5 Caractérisation de la transcription intergénique aux sites de liaison de l'ARN polymérase II dans les tissus normaux et cancéreux

5.1 Article

Les regions intergéniques de notre génome sont connues pour être largement transcrites dans les tissus normaux [49] et dans les cancers [34]. Cependant, l'annotation et la compréhension de ce processus reste incomplète, et ce particulièrement en comparaison des régions codantes géniques du génome. En exploitant l'énorme quantité de données de séquençage fonctionnel générée par les laboratoires de recherche et disponibles sur les entrepôts de données de GEO [38] et ENCODE [119], nous avons réanalysé dans un premier temps près de 1100 expériences ChIP-seq ciblant la Pol II, afin d'identifier plus de 180 000 régions du génome humain étant potentiellement transcrites de par la présence de la Pol II, et situées dans les régions intergéniques. Ces expériences ChIP-seq couvrent une large diversité de tissus et lignées cellulaires humaines, ce qui d'une part permet d'obtenir un catalogue assez complet, et d'autre part de relier le recrutement de la Pol II dans des catégories cellulaires données à une région génomique précise.

Par la suite, nous avons utilisé des données RNA-seq provenant d'ENCODE, TCGA et GTEx, afin de quantifier le niveau de transcription dans ces 180 000 régions occupées par la Pol II dans l'intergénique. Malgré le signal très faible, cela nous a permis de confirmer que nos régions étaient préférentiellement transcrites par rapport aux reste du génome intergénique. En étudiant de façon comparative

les signaux entre conditions biologiques, en compensant en signal faible par un grand nombre d'échantillons, nous avons pu discriminer les différentes conditions biologiques et identifier des marqueurs d'expression différentielle étant cohérents entre les trois jeux de données.

Dans les cancers, cela a permis d'identifier des régions dans l'intergénique étant différentiellement transcrites entre les tissus tumoraux et les tissus normaux, mais également entre sous types de cancers. En étudiant ces résultats à un niveau pan-cancer, cela a permis d'identifier des marqueurs "pan-cancer" étant anormalement exprimés dans de nombreux cancers. De façon similaire, nous avons associé l'expression de ces régions à la survie des patients, et identifié de nouveaux marqueurs pronostiques.

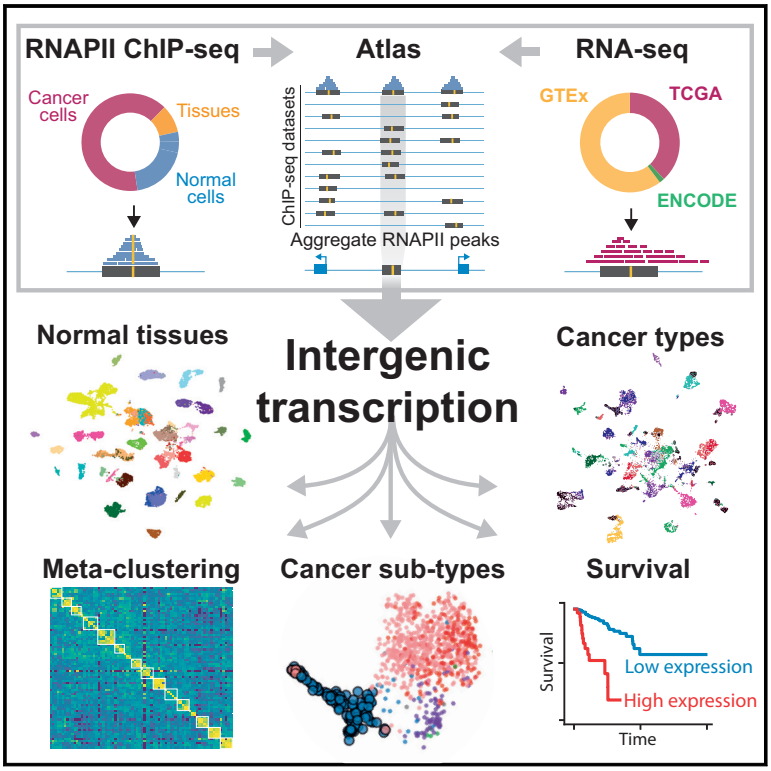
Dans un but de reproductibilité et de partage des données, les données générées par ce travail et celles ayant servi à les générer sont disponibles sur Zenodo[83, 84], tout comme le code source, qui est disponible sur Github[149].

Cell Genomics

Article

Characterizing intergenic transcription at RNA polymerase II binding sites in normal and cancer tissues

Graphical abstract



Highlights

- An RNAPII atlas of intergenic transcription in normal tissues and cancer samples
- Intergenic atlas shows enhancer-like characteristics and transcriptional signals
- Meta-clustering reveals shared transcriptional patterns among tissues and cancer types
- Identified intergenic markers associated with cancer genes and survival

Authors

Pierre de Langen, Fayrouz Hammal, Elise Guéret, Jean-Christophe Mouren, Lionel Spinelli, Benoit Ballester

Correspondence

benoit.ballester@inserm.fr

In brief

We developed an atlas of intergenic transcription using RNAPII binding sites to connect genomic and transcriptomic data in normal tissues and cancer samples. The atlas enables investigation of tissue specificity and core regulatory elements. Meta-clustering reveals shared transcription patterns among tissues and cancer types. We identified intergenic markers that are associated with known cancer genes and predictive of overall survival. Our study demonstrates the effectiveness of integrating diverse public datasets to characterize intergenic transcription in normal and cancer tissues, addressing limitations of previous techniques.

de Langen et al., 2023, Cell Genomics *3*, 100411
October 11, 2023 © 2023 The Authors.
https://doi.org/10.1016/j.xgen.2023.100411

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

Characterizing intergenic transcription at RNA polymerase II binding sites in normal and cancer tissues

Pierre de Langen,[1] Fayrouz Hammal,[1] Elise Guéret,[1] Jean-Christophe Mouren,[1] Lionel Spinelli,[1] and Benoit Ballester[1,2,*]
[1]Aix Marseille Univ, INSERM, TAGC, Marseille, France
[2]Lead contact
*Correspondence: benoit.ballester@inserm.fr
https://doi.org/10.1016/j.xgen.2023.100411

SUMMARY

Intergenic transcription in normal and cancerous tissues is pervasive but incompletely understood. To investigate this, we constructed an atlas of over 180,000 consensus RNA polymerase II (RNAPII)-bound intergenic regions from 900 RNAPII chromatin immunoprecipitation sequencing (ChIP-seq) experiments in normal and cancer samples. Through unsupervised analysis, we identified 51 RNAPII consensus clusters, many of which mapped to specific biotypes and revealed tissue-specific regulatory signatures. We developed a meta-clustering methodology to integrate our RNAPII atlas with active transcription across 28,797 RNA sequencing (RNA-seq) samples from The Cancer Genome Atlas (TCGA), Genotype-Tissue Expression (GTEx), and Encyclopedia of DNA Elements (ENCODE). This analysis revealed strong tissue- and disease-specific interconnections between RNAPII occupancy and transcriptional activity. We demonstrate that intergenic transcription at RNAPII-bound regions is a novel per-cancer and pan-cancer biomarker. This biomarker displays genomic and clinically relevant characteristics, distinguishing cancer subtypes and linking to overall survival. Our results demonstrate the effectiveness of coherent data integration to uncover intergenic transcriptional activity in normal and cancer tissues.

INTRODUCTION

Transcription is a fundamental process in biology that transcribes DNA into biologically active and cell-type-specific RNA molecules. The majority of transcription is carried out by RNA polymerase II (RNAPII), which generates mRNAs that are subsequently translated into proteins. However, intergenic active regions have been shown to cover a much larger fraction of the genome than expected.[1] Indeed, RNAPII transcribes a wide variety of intergenic active regions, such as different types of non-coding RNAs (ncRNAs)[2] or enhancer RNAs (eRNAs) that have been found to be major sites of intergenic transcription.[3]

While genes and their protein products have been the main interest in basic and cancer research, an increasing amount of genomic data support the biological and clinical relevance of intergenic transcription. Aberrant expression of ncRNAs has been found in cancer[4] and non-cancer disease,[5] and a vast majority of trait or disease-associated variants lie in non-coding regions of the genome.[6] Despite significant progress in describing enhancer transcription,[3,7–10] efforts to fully identify intergenic transcription remain a challenge. This is primarily due to a limited amount of sequencing assays like global run on sequencing (GRO-seq)[11] or its derivatives,[12,13] impacting the discovery of a broader intergenic transcription landscape.

In this study, we compiled each available RNAPII chromatin immunoprecipitation sequencing (ChIP-seq) dataset from the GEO[14] and Encyclopedia of DNA Elements (ENCODE)[1] to construct an atlas of RNAPII-bound intergenic regions in the human genome. Our approach, which targets RNAPII binding rather than the resulting ncRNA, aims to minimize the limitations of RNA abundance and stability. This approach enables exploration of active intergenic regions in a broad range of cell types and tissues, which have not been extensively studied before.

We hypothesize that intergenic RNAPII-bound regions of significance exhibit a biotype-specific signature, reflected in biotype-specific RNA sequencing (RNA-seq) expression across resources such as the Genotype-Tissue Expression (GTEx[15]), The Cancer Genome Atlas (TCGA[16]) and The Encyclopedia of DNA Elements (ENCODE[1]). In this study, we describe tissue-specific bindings by creating an atlas of intergenic RNAPII-bound regions. By analyzing the expression patterns of 28,797 RNA sequencing samples, we identify intergenic transcription on RNAPII-bound regions as a powerful indicator for characterizing tissue types. We show that using intergenic transcription on RNAPII-bound regions results in robust classification of cancer types and subtypes.

Taken together, our study indicates that intergenic transcription at RNAPII binding sites is a powerful indicator for characterizing normal and cancer tissues at the subtype level. While the functional significance of intergenic regions remains an open

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

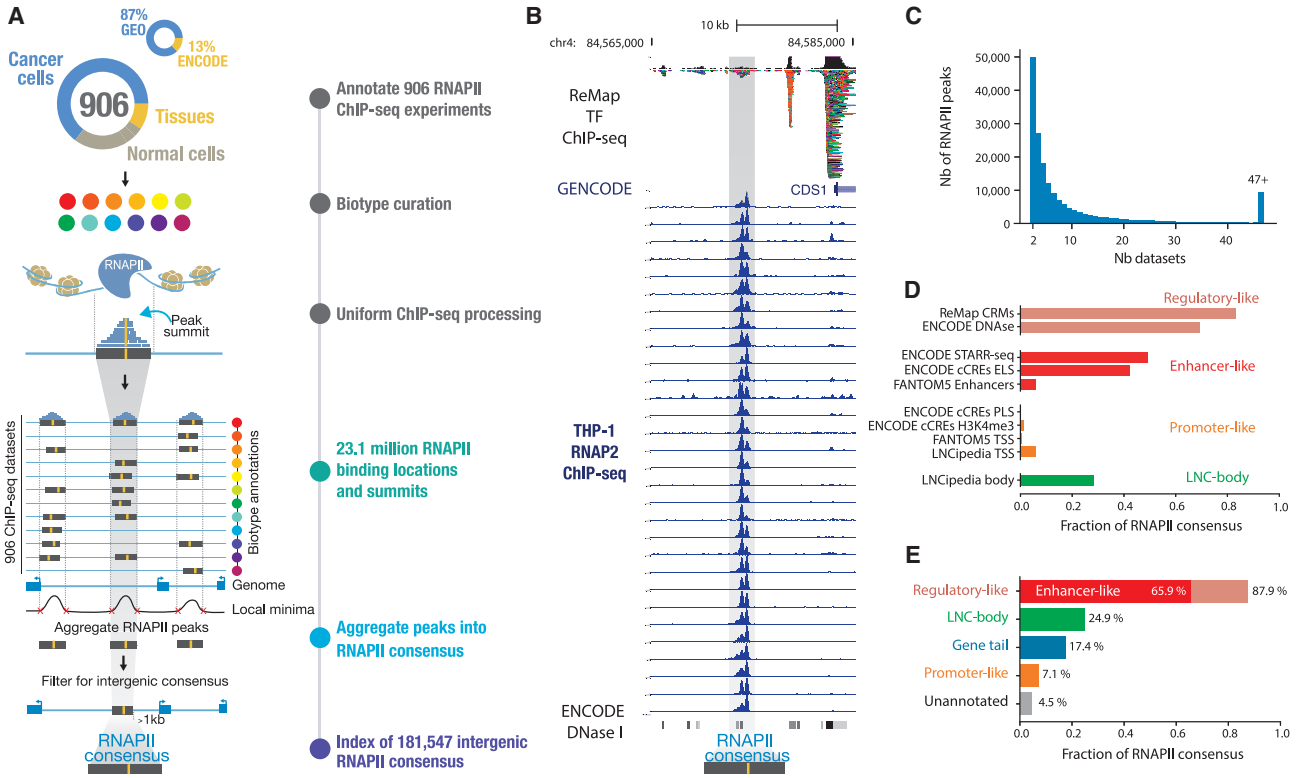


Figure 1. An atlas of intergenic RNAPII occupancy

(A) Overview of the RNAPII atlas pipeline; 23.1 million RNAPII-bound regions aggregated across 906 individual datasets jointly identify 181,547 intergenic RNAPII consensus.

(B) Genomic example on chromosome 4, showing RNAPII raw ChIP-seq signals across THP-1 cell lines (leukemia, in blue) at the location of a RNAPII consensus (gray bar), with ReMap TP ChIP-seq and ENCODE DNase I tracks.

(C) Distribution of the number of datasets across which RNAPII peaks are shared.

(D) Comparison of RNAPII consensus location with genomic resources of regulatory and non-coding elements; resources are grouped and colored by genomic characteristics.

(E) Annotation of the RNAPII atlas according to genomic characteristics: regulatory like, enhancer like, long non-coding body, gene tail, promoter like, and unannotated.

question, our findings could significantly enhance our understanding of the regulatory programs and clinical relevance of non-coding transcription in various cancers.

RESULTS

An atlas of intergenic RNAPII occupancy

To create an atlas of intergenic RNAPII binding in the human genome, we collected all available ChIP-seq data targeting RNAPII on a wide variety of cells and tissue biosamples from public biological data warehouses[1,14] (Figure 1A). The created atlas aggregates 87% of non-ENCODE datasets and 13% of ENCODE datasets (Figure 1A). This was accomplished through standardized manual curation of sample metadata, uniform biosample annotation, and consistent data processing and quality screening, initiated from the raw sequencing files using the ReMap pipeline (STAR Methods). We conservatively retained 906 RNAPII datasets from diverse cell or tissue types, utilizing various antibodies targeting the POLR2A subunit (Figure S1A). These datasets encompassed a wide range of samples, including cancer cell lines (64%) and "normal" cell lines/tissues (36%) (Figure 1A; STAR Methods). In this study, we focused specifically on intergenic RNAPII-bound regions, preventing us from detecting alternative promoters or any transcriptional events occurring within gene bodies (STAR Methods). We defined intergenic regions as all regions of the genome, excluding all GENCODE transcripts (as well as known long ncRNAs [lncRNAs]) extended by 1 kb at the transcription start site (TSS) and transcription end site (TES) and excluding ENCODE blacklisted regions.[17] We identified a total of 23,101,589 RNAPII binding events across all 906 datasets, of which 2,525,886 (11.1%) are localized within intergenic regions (averaging 2,787 intergenic binding events per dataset; Figure S1B). A large fraction of RNAPII intergenic binding events (91.7%) is shared across at least two ChIP-seq datasets, suggesting similar occupancy patterns across experiments (Figures 1B and S2). These binding events are also found to be located on clusters of transcription factor ChIP-seq peaks. We developed an aggregative approach to identify across experiments what we refer to as "consensus peaks" (Figures 1A and S3; STAR Methods). By applying this

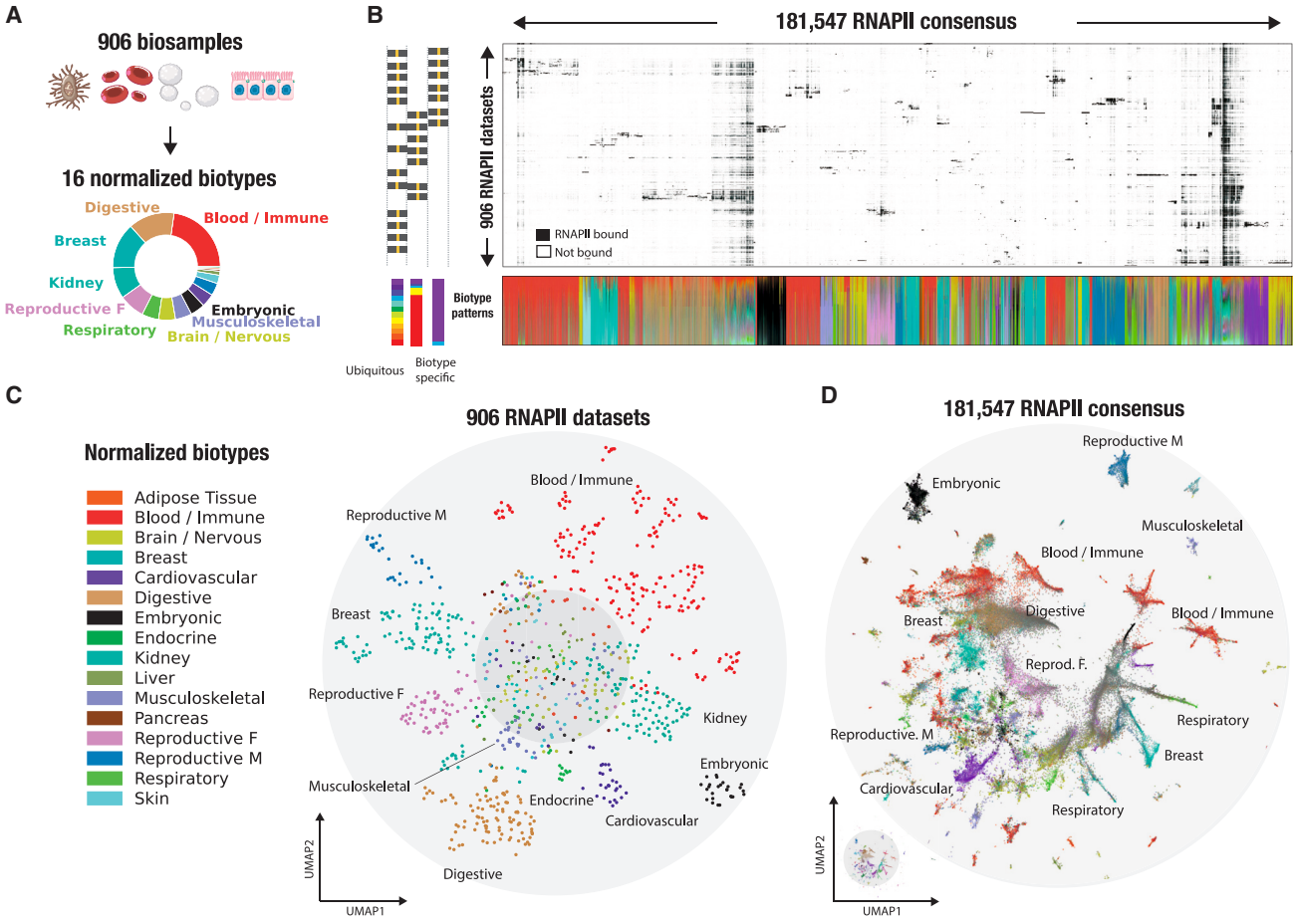


Figure 2. A normalized vocabulary captures biotype-specific intergenic RNAPII binding

(A) Distinct tissues and cell lines across 906 biosamples normalized into 16 biotypes.

(B) Intergenic RNAPII occupancy in 181,547 consensus regions across 906 biosamples displayed in a visually compressed matrix. The color code used for each RNAPII consensus region corresponds to the biosample tissue of origin, with examples representing either biotype-specific or ubiquitous signatures. This color scheme is consistently applied across all RNAPII consensus regions. Bottom: normalized contribution of a biotype, in terms of peaks, to each RNAPII consensus (STAR Methods).

(C) Two-dimensional uniform manifold approximation and projection (UMAP) of all 906 RNAPII ChIP-seq datasets across intergenic RNAPII space, colored by normalized biotype.

(D) UMAP representation of all intergenic RNAPII consensus organized by their binding patterns, colored by dominant biotype (STAR Methods; full UMAP available in Zenodo).

approach, we created an atlas of 181,547 intergenic RNAPII consensus peaks, describing distinct genomic elements bound by RNAPII across multiple biosamples. Our atlas of intergenic RNAPII-bound regions, available on Zenodo,[18] is based on consensus peaks derived from an average of 13 datasets (Figure 1C), with each consensus having an average width of 410 bp (Figure S1C). Each peak and dataset in the ChIP-seq data contributing to a representative RNAPII consensus can be traced back to its corresponding biosample or cell type category (Figures S1D and S3). We evaluated our created atlas against reference databases of regulatory and non-coding genomic elements[19–23] (Figures 1D, 1E, S4–, and S6). We found that the majority of RNAPII consensus peaks (87.9%) were categorized as regulatory regions, with 65.9% showing an enhancer signature (Figure S4A). Furthermore, we observed a concentration of RNAPII consensus downstream of genes (17.4%), within the +1- to +9-kb range (Figures S5A and S5B). Interestingly, these regions exhibit characteristics of regulatory elements and show a strong enrichment of CTCF and CTCFL (BORIS) motifs (Figure S5C). Our findings indicate that the atlas of intergenic RNAPII consensus peaks is predominantly located over regulatory elements and potentially transcribed enhancer regions.

A normalized vocabulary captures biotype-specific intergenic RNAPII binding

The RNAPII atlas covers a significant fraction of the human biological spectrum, including over 203 distinct tissues and cell lines (Figure 2A; Table S1). To facilitate biological interpretation, we grouped biosample annotations based on their tissue of origin or similarity. We then further categorized similar tissues

into 16 distinct biotypes to obtain a concise but meaningful high-level annotation of our samples (Table S1; STAR Methods). To simplify genomic interoperability across large resources, the compendium of tissues and cell lines was harmonized using Genotype-Tissue Expression (GTEx), The Cancer Genome Atlas (TCGA), ENCODE biosample nomenclature, as well as cell ontologies.[24] This results in the RNAPII consensus exhibiting a biological context ranging from biotype-specific to ubiquitous signatures (Figure 2B). Because intergenic RNAPII binding appears to be shared extensively across biosamples (Figures 1B and S2), we aimed to visualize RNAPII occupancy patterns across biosamples and consensus by employing a hierarchical clustering approach (Figure 2B; STAR Methods). Patterns of RNAPII binding were structured into mostly biotype-specific and a few ubiquitous occupancy clusters. We observed what seemed to be a sparse distribution in the intergenic RNAPII atlas, but upon further investigation, we identified diverse and intricate binding patterns. To analyze these patterns, we utilized an unsupervised dimensionality reduction technique (uniform manifold approximation and projection [UMAP][25]) on the 906 biosamples (Figure 2C) and more than 180,000 RNAPII consensus peaks (Figure 2D). The UMAP visualization across 906 ChIP-seq datasets revealed organized intergenic occupancy patterns across similar biotypes (Figure 2C). Based on their intergenic occupancy patterns, ChIP-seq datasets having similar biotypes of origin were clustered together, while the center of the plot contained datasets with ubiquitous biotype signatures. For example, ChIP-seq datasets for digestive biosamples (represented by brown dots, n = 126 samples) were predominantly clustered together, suggesting that intergenic RNAPII occupancy is representative of the sample biology but also that the biosample curation is coherent. Next, we visualized the 181,547 intergenic RNAPII consensus peaks according to their binding patterns and biotype labels (Figure 2D; STAR Methods). To facilitate biological interpretation of an RNAPII consensus, each consensus was labeled with its most frequent biotype or labeled in gray when ubiquitous. By visualizing the intergenic RNAPII atlas, we were able to identify distinct occupancy patterns that are specific to certain biotypes. This framework was also applied to 890 H3K27ac datasets, successfully demonstrating its ability to identify biotype-specific clusters of histone modifications (Figure S7). The RNAPII atlas, generated by leveraging 906 ChIP-seq datasets, provides a valuable biotype-specific summary of intergenic RNAPII binding. Its potential to uncover intergenic transcriptional activities makes this atlas an innovative tool.

Revealing tissue-specific regulatory signatures

We next aimed to retrieve and annotate each consensus group to capture its biological identity. Using an unsupervised graph clustering approach, we identified 51 RNAPII consensus clusters (Figure 3A), each harboring its own biotype specificity (Figures 3B and S8; Table S1). To independently validate their biological signatures, we compared the clusters against the biological classification of the human index of DNase I hypersensitive sites[26] (DHSs) (Figure 3C). The defined RNAPII clusters showed a coherent enrichment with the DHS regulatory vocabulary (Figure S9). For instance, "brain/nervous" RNAPII cluster 31 (light green) was enriched in neural DHSs. To capture the genomic signatures of these groups, we examined the epigenetic state for each RNAPII cluster, particularly focusing on its chromatin state specificity. As an example, we selected RNAPII cluster 4, which exhibited a distinct "embryonic" signature, and analyzed the Roadmap ChromHMM (software for learning and characterizing chromatin states) epigenetic states of embryonic stem cells (Figure 3D). We observed a strong enrichment of "active" epigenetic states, including enhancers, TSSs, and transcribed regions, within the RNAPII embryonic cluster compared with the other RNAPII clusters (Figure S10A; STAR Methods). Conversely, we observed a depletion of "inactive" epigenetic states, such as quiescent or Polycomb-repressed states. This finding suggests that RNAPII occupies intergenic space at key regulatory elements, as demonstrated previously[27] (Figures 1D and 1E). To explore the tissue specificity of RNAPII clusters, we analyzed enhancer-like histone marks and open chromatin profiles (H3K27ac, ATAC-seq; Table S2). The results revealed that RNAPII cluster-tissue pairs with matching tissues (e.g., heart-cardiovascular) exhibited the strongest activity, while non-matching pairs (e.g., lymphoid-liver) displayed a weaker signal (Figure S10B).

To further confirm the biological identity of defined clusters, we investigated the enrichment of SNP-based trait heritability from a UK Biobank genome-wide association study[24] (GWAS), transcription factor binding regions (TFBRs) from ReMap,[19] Gene Ontology (GO) terms, and HOMER[28] DNA motifs (Figures 3E and S9). "brain/nervous" cluster 31 exhibited enrichment of TFBRs for transcription factors known to be involved in neural development or diseases, such as TCF12, PITX3, and TWIST1. Similarly, at the sequence level, the embryonic RNAPII cluster exhibits enrichments of meaningful transcription factor motifs, specifically OCT4-Sox-NANOG motifs. "Cardiovascular" cluster 10 showed enrichments in multiple heart-related traits, such as intra-corneal pressure, pulse rate, and coronary heart disease. Similarly, blood/immune cluster 5 included an RNAPII consensus located near genes linked to immune response GO terms, consistent with their assigned biotypes. Our study accurately distinguishes intergenic RNAPII occupancy based on its biotype specificity, revealing tissue-specific regulatory signatures across multiple independent genomic resources. These resources range from open chromatin occupancy maps to transcription factor binding, providing comprehensive insights into the regulatory landscape.

Systematic transcription captured in the intergenic RNAPII atlas

We developed the RNAPII atlas as an innovative tool for indirectly identifying intergenic regulatory regions that are active or poised for transcription. To quantify intergenic transcription and gain a better understanding of transcriptional patterns, we utilized the RNAPII atlas to analyze transcriptional signals in three major expression resources. These resources include samples from normal and cancer cell lines: GTEx, TCGA, and the ENCODE consortium. By combining these, we conducted an extensive analysis of intergenic expression across the RNAPII atlas, leveraging data from 28,767 RNA-seq samples (Figure 4A). To quantify intergenic transcription, we first standardized each RNAPII consensus sequence to a 1-kb RNAPII-bound region. We then counted the number of reads that overlapped with these RNAPII -bound regions, generating a count table similar to conventional

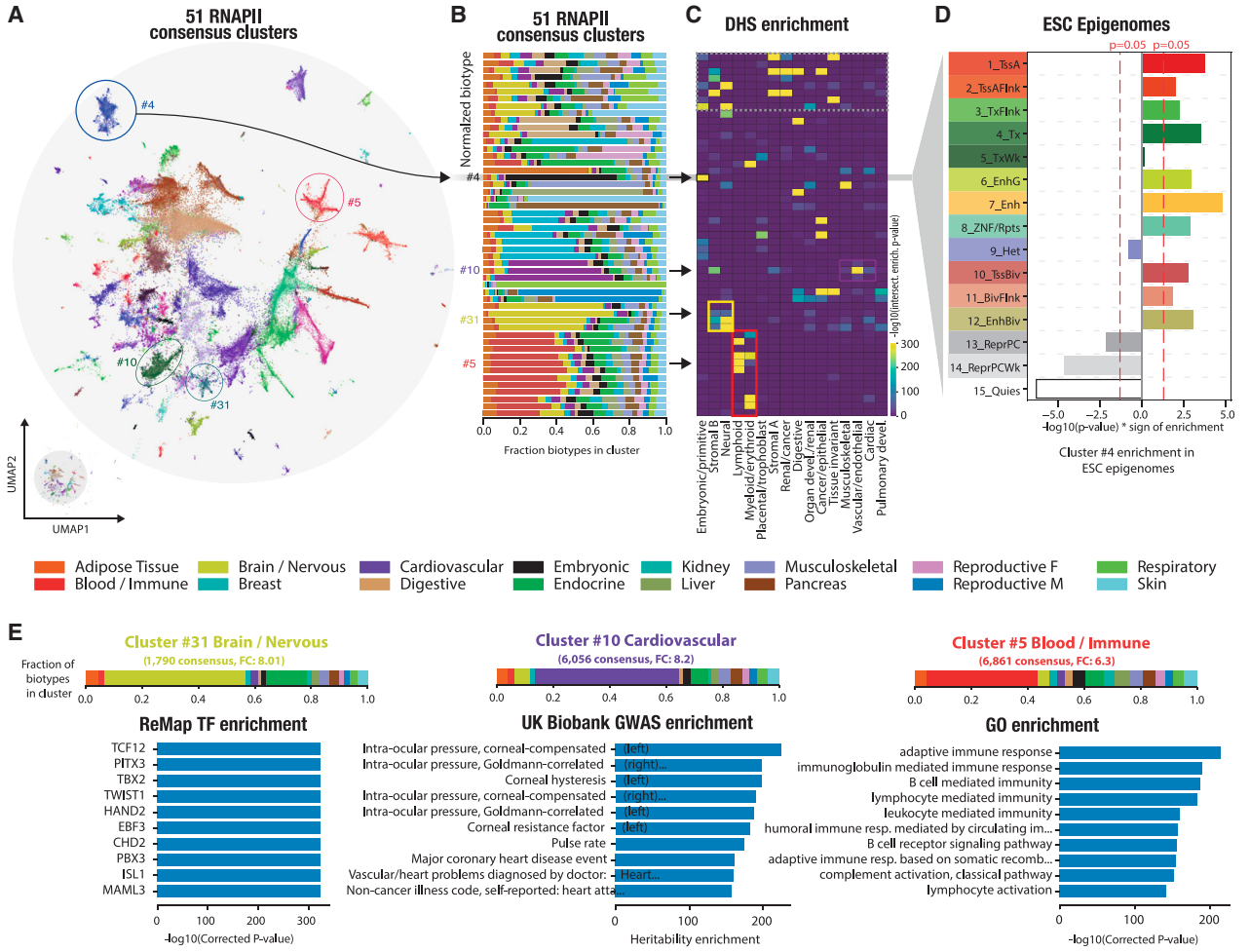


Figure 3. Revealing tissue-specific regulatory signatures

(A) Unsupervised graph clustering identifies 51 RNAPII consensus clusters. Four clusters (4, 5, 10, and 31) are highlighted across panels (A)–(C) to illustrate the analysis.

(B) The fraction of biotypes within each cluster is shown, indicating tissue-specific or ubiquitous signatures.

(C) Enrichment of DNase I hypersensitive site (DHS) biological classification in each cluster. Arrows and colored rectangles highlight correspondence between clusters and DHS categories.

(D) Enrichment of ChromHMM epigenetic states of "embryonic stem sell," sampled at the RNAPII genomic location of cluster 4, against the non-cluster 4 RNAPII consensus. Active states: active TSS transcription states (TssA and TssAFlnk), transcribed promoter and enhancer signatures (TxFlnk), actively transcribed states (Tx and TxWk), enhancer states (Enh and EnhG), zinc-finger protein gene state (ZNF/Rpts). Inactive states: heterochromatin (Het), bivalent regulatory states (TssBiv, BivFlnk, and EnhBiv), repressed Polycomb states (ReprPC and ReprPCWk), and quiescent state (Quies).

(E) Top 10 transcription factor enrichments from the ReMap database in cluster 31, top 10 UK Biobank GWAS trait heritability enrichment in cluster 10, and top 10 GO enrichment of nearby genes in cluster 5. All results shown are statistically significant. Each cluster's biotype distribution is shown as a stacked bar plot.

gene-centric RNA-seq count tables (Figure 4A). Our analysis revealed that the intergenic RNAPII atlas captured approximately 60% of intergenic reads (Figures S11A and S11B). Moreover, these RNAPII-bound intergenic regions captured significantly higher read counts compared with the rest of the intergenic genome. On average, RNAPII -bound regions had 7.13 times more transcriptional signal compared with the remaining intergenic genome (Figures 4B and S11C). By visualizing the spatial distribution of transcriptional signals within RNAPII-bound regions, we discovered two distinct types of transcriptional patterns: one displaying a mono-modal signal with a short peak and the second showcasing a broader peak spanning the entire 1-kb probe (Figure S12). Altogether, the RNAPII atlas is strongly enriched in transcriptional activity, and thus it could serve as a powerful tool for investigating intergenic transcription in normal and cancer tissues.

Intergenic transcription on the RNAPII atlas is a powerful indicator for characterizing tissues

To determine whether intergenic transcription at RNAPII-bound regions could characterize tissue specificity, we analyzed expression data from 54 non-diseased tissues, comprising a total of 17,345 samples from the GTEx project. Intergenic

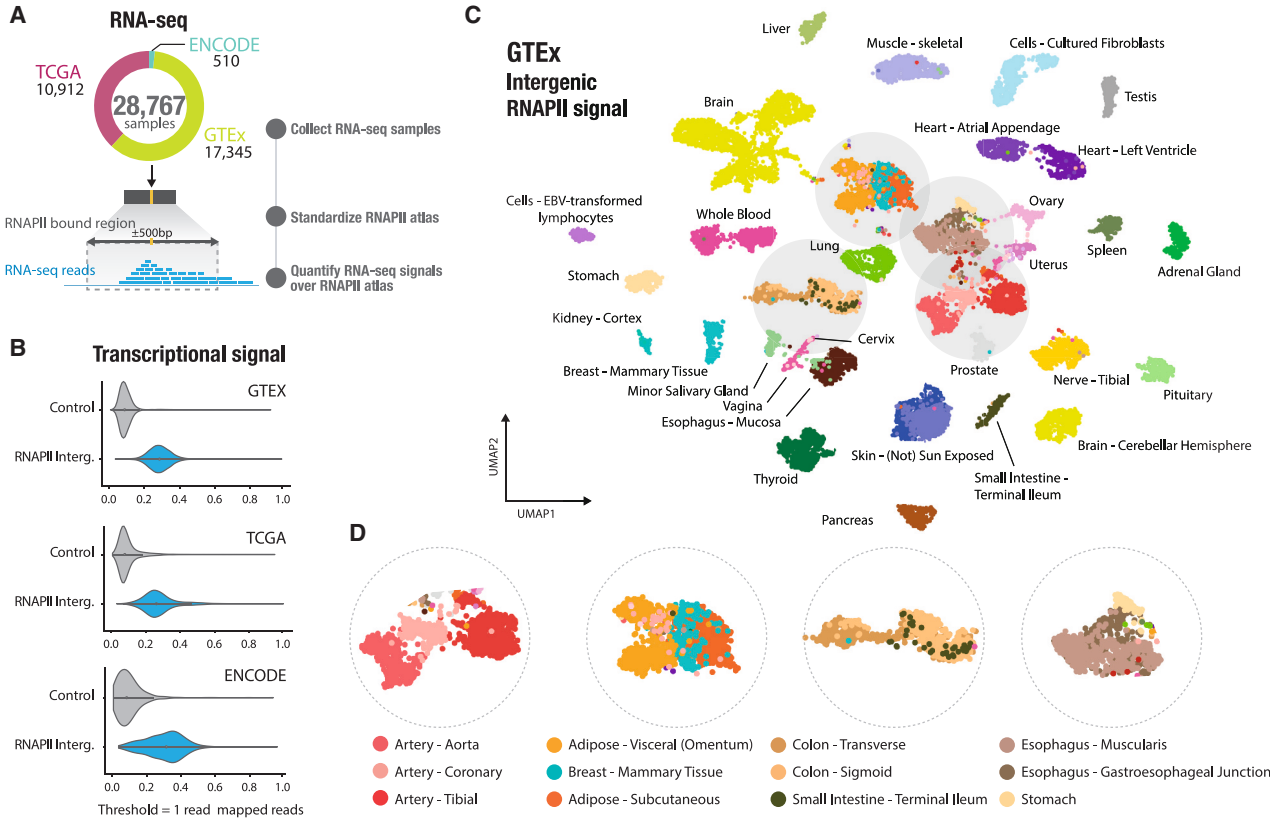


Figure 4. Intergenic transcription on the RNAPII atlas is a powerful indicator for characterizing tissues

(A) Number of RNA-seq samples from three expression resources (GTEx, TCGA, and ENCODE) and schematic depicting the standardization of RNAPII consensus to 1-kb RNAPII-bound regions to obtain read counts.

(B) Violin plots comparing transcriptional signals at intergenic RNAPII-bound regions versus non-RNAPII random intergenic regions across the three expression resources.

(C) Two-dimensional UMAP projection of 17,345 GTEx RNA-seq signals across the intergenic RNAPII atlas, with colors representing 54 tissue types, including 11 distinct brain regions (yellow) and two cell lines (light blue).

(D) Magnified view of tissue-specific expression patterns observed in similar tissues, such as different types of artery (e.g., aorta, coronary, and tibial).

transcription has been utilized previously as a marker of enhancer activity, as demonstrated in the Functional Annotation of the Mammalian Genome (FANTOM) project,[7] and across various experimental assays focused on capped and nascent RNAs.[3,10,12,13] In this study, we developed a pipeline based on single-cell RNA-seq (scRNA-seq) methods, which are commonly employed for analyzing weak signals in datasets with large sample sizes. By considering signals only within RNAPII -bound regions, we were able to extract valuable biological information from read count tables (Figure S13; STAR Methods). We used UMAP to analyze and visualize similarity between the expression levels sampled at RNAPII-bound regions of each GTEx biosample. This analysis revealed a clear distinction between tissues because biosamples originating from the same sampling site are clustered together (Figure 4C). These tissue-specific expression patterns are observed not only between similar tissues, such as artery (aorta, coronary, and tibial) but also between tissues with similar histological features, like adipose tissue (visceral, subcutaneous, and mammary tissue). Additionally, these patterns are observed between tissues located in different body regions, such as the digestive tract (colon and small intestine) (Figure 4D). To test whether intergenic transcription could accurately discriminate the 54 GTEx tissues accurately, we employed a k-nearest neighbor algorithm (KNN classifier) to classify the tissues based on expression of RNAPII-bound regions. This showed that RNAPII-bound regions could predict tissue types with a high level of accuracy, with only a slight decrease in accuracy compared with gene-centric RNA-seq counts processed using the same methods (87.1% against 90.0% balanced accuracy across 54 tissues; Figure S14). Next, we identified overexpressed intergenic RNAPII-bound regions in the GTEx tissues with an average of 4,236 regions per tissue (Figure S15A; STAR Methods). Our analysis revealed a significant association between RNAPII-bound regions with tissue-specific overexpression and tissue-specific GTEx eQTLs (Figures S15B and S15C; STAR Methods), providing evidence that these regions can serve as indicators of transcribed intergenic enhancers. Interestingly, we also detected transcriptional signals at RNAPII-bound regions located downstream of genes (>1 kb), suggesting the presence of transient RNA downstream of the polyadenylation site

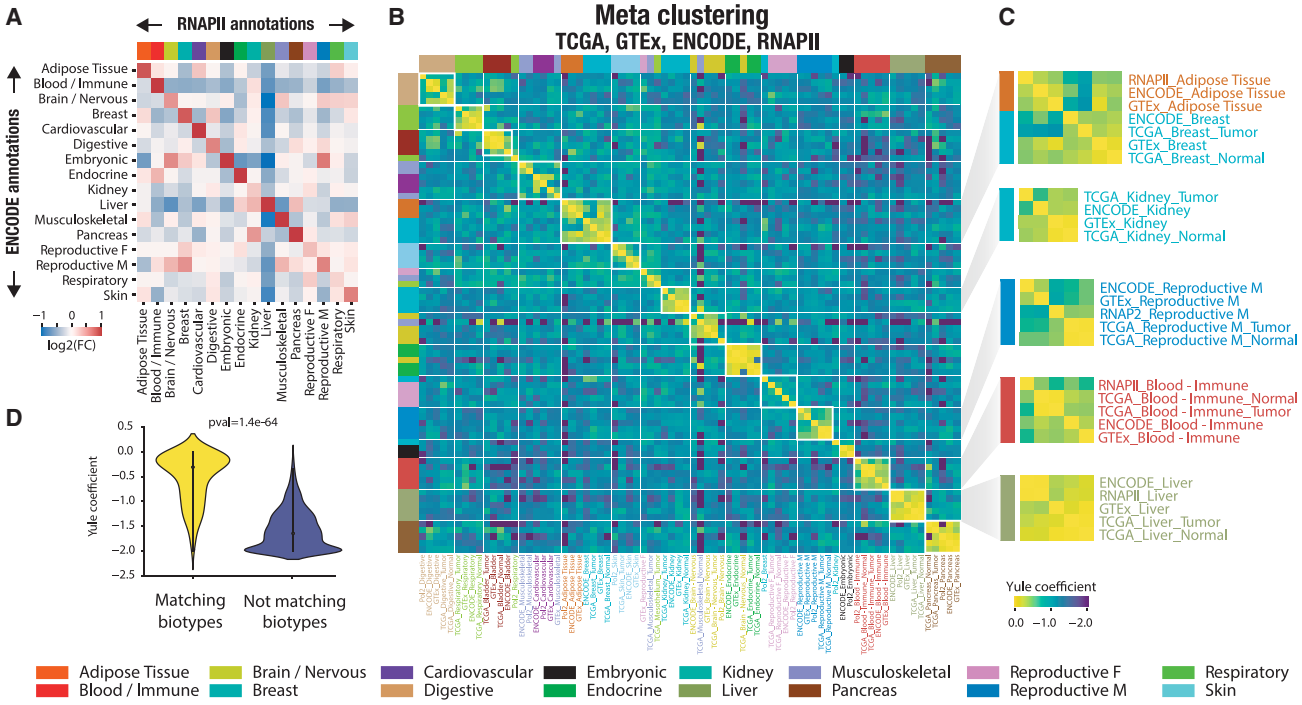


Figure 5. Meta-analysis reveals tissue- and disease-specific connections between RNAPII occupancy and transcription

(A) Association between RNAPII occupancy biotype and transcription biotype from ENCODE. The heatmap depicts log2 of ENCODE RNA-seq dataset expression fold change in each biotype (rows) between RNAPII-bound regions with biotype-specific RNAPII ChIP-seq occupancy (columns) against non-specific RNAPII-bound regions.

(B) Heatmap showing the association between biotype-specific intergenic RNAPII occupancy and biotype-specific RNAPII overexpression across four resources. A hierarchically clustered heatmap reveals the correct grouping by tissue of origin rather than data source, with each possible biotype-dataset pair represented. Yule distance between a pair of dataset-biotype lists of overexpressed RNAPII markers is indicated.

(C) Magnified view revealing meta-clusters of tissue-specific correlation between intergenic RNAPII regions and their transcription in different resources.

(D) Distributions of tissue-matching (i.e., RNAPII-liver vs. TCGA-liver) and non-matching (i.e., RNAPII-liver vs. GTEx-heart). Yule distance between two intergenic RNAPII marker sets ($p = 1.4e{-}64$) is indicated.

(Figures S5A and S5B), which is consistent with previous studies.[29] To further investigate the impact of these downstream signals, we conducted additional analyses excluding RNAPII-bound regions located up to 9 kb downstream of genes. Our findings demonstrate that RNAPII consensus peaks located within the 1- to 9-kb region downstream of genes do not drive classification of GTEx tissues (Figure S16). Furthermore, we show that our approach is applicable to smaller RNA-seq datasets (Figure S17). By comparing the expression levels in three samples of two types of heart tissues from GTEx biosamples, we identified 195 RNAPII-bound regions located near genes related to heart function, despite limited statistical power (Figure S17; STAR Methods). Here, we provide evidence that intergenic transcription detected at RNAPII-bound regions is a strong indicator of tissue specificity and can be used effectively for accurate tissue type prediction. These findings may have implications for understanding tissue-specific gene regulation.

Meta-analysis reveals tissue- and disease-specific connections between RNAPII occupancy and transcription

We examined the relationship between biotype-specific RNAPII occupancy and biotype-specific transcription by comparing the observed intergenic signal across all expression datasets, which combined 28,787 RNA-seq samples despite use of different sequencing samples and protocols. We first conducted an analysis to investigate the association between biotype-specific RNAPII occupancy in ChIP-seq and transcription in ENCODE RNA-seq biotypes by comparing biotypes pairwise (Figure 5A). This analysis revealed a significant enrichment of biotype-specific transcription in the ENCODE dataset at RNAPII probes with ChIP-seq occupancy specific to the corresponding biotype, even when considering different samples and protocols. Conversely, non-matching biotype pairs did not exhibit transcriptional signal enrichments. These findings underscore a strong link between RNAPII occupancy and effective transcription as well as the effectiveness of our biosample annotation for comparing varied data sources. Furthermore, we conducted a meta-analysis that integrated every dataset and biotype to obtain a comprehensive and interconnected view of intergenic transcription across nearly 30,000 biosamples from diverse data sources (Figure 5B). In brief, we extracted intergenic RNAPII markers (only considering up-regulated RNAPII -bound regions; STAR Methods) for each possible biotype-dataset pair (i.e., RNAPII-liver, GTEx-heart, and ENCODE-liver) and quantified pairwise similarity between marker lists for every

biotype-dataset combination, assuming that a marker list is characteristic of a specific biotype. We then applied hierarchical clustering to generate a meta-clustering that revealed similarities between tissues across all resources (Figure 5B; STAR Methods). This meta-analysis highlighted that the association between intergenic RNAPII occupancy and intergenic transcription is biotype specific, consistently observed across biotypes and independent of dataset origins or protocols used. Our approach effectively grouped similar biotypes together, independent of the data source (Figure 5B). For instance, ''adipose tissue'' and ''Breast'' tissues clustered together across resources, reflecting the presence of adipose cells in breast tissue (Figure 5C). Moreover, identical biotypes exhibited much greater similarity in markers across data sources than non-identical biotypes (Figure 5D). To ensure robustness, we extracted markers that were supported by at least half of the data sources for each biotype. These markers demonstrated a strong enrichment of heritability in biotype-related traits, confirming their biological relevance (Figure S18). For instance, markers associated with the ''reproductive female'' biotype showed a strong correlation with heritability of the ''birth weight of first child'' trait, while markers associated with the ''liver'' exhibited enrichment in heritability for ''high cholesterol.'' In summary, our meta-analysis revealed a tissue-specific correlation between intergenic transcription and RNAPII occupancy, which carries biological significance. Furthermore, we observed a remarkable consistency across diverse data sources and protocols.

Cancer type and subtype classification by intergenic transcription at RNAPII binding sites

We have shown that intergenic transcription can reliably differentiate between various tissues and biological conditions. Expanding on this understanding, we explored direct applications of our RNAPII atlas and its potential implications in human cancers. We analyzed expression data from 32 cancer types, encompassing 10,912 RNA-seq samples obtained from the TCGA cohort, to identify clinically relevant intergenic transcription patterns and potential therapeutic targets (Figure 6A). Using UMAP, we analyzed and visualized the similarity in expression profiles among biosamples, revealing an initial separation between cancer types and subsequently between normal or tumoral tissue states, suggesting that certain RNAPII-bound regions are differentially expressed in these contexts (Figure 6B). For example, in the case of brain cancers, lower-grade glioma (LGG) and Glioblastoma multiforme (GBM) display close clustering, whereas kidney tumor samples (kidney renal clear cell carcinoma [KIRC], kidney chromophobe carcinoma [KICH], and KIRP) exhibit distinct expression profiles despite the similarity observed in normal kidney samples. Interestingly, breast cancer (BRCA) samples form two distinct clusters based on expression of intergenic RNAPII-bound regions. These clusters correspond to distinct BRCA subtypes, with the basal-like subtype (triple-negative BRCA [TNBC]) being the most distinct and the luminal A, luminal B, and HER2-positive subtypes forming a separate, larger group (Figure 6C). We identified intergenic transcriptional markers specific to the basal-like/TNBC subtype, which are associated with 10 dual-specificity phosphatase genes (e.g., DUSP1, DUSP5, and DUSP7), involved in mitogen-activated protein kinase (MAPK) phosphatase activity. MAPK cascades play a central role in cell proliferation and apoptosis, and DUSP1 may contribute to development of chemoresistance in TNBC.[30,31] TNBC accounts for approximately 15%–20% of all BRCA cases, is most prevalent in women under 40,[32] and presents aggressive behavior.[33] Similar to BRCA, intergenic transcription in thyroid carcinomas (THCA) facilitated the identification of different subtypes of THCAs (Figure 6D). By using a heatmap representation of the differentially expressed RNAPII-bound regions in KICH samples, we observed distinct clusters of up-regulated and down-regulated RNAPII -bound regions. These clusters indicate potential tumor subtypes with unique intergenic expression patterns (Figure S19A). Identification of subtype-specific intergenic transcription sheds light on cancer biology by revealing active regulatory elements and potentially actionable nearby genes with clinical significance.

Identification of per-cancer and pan-cancer intergenic transcriptional markers

We identified tumor-specific RNAPII-bound regions differentially expressed in tumors compared with normal tissues for 16 cancer types, ranging from 65,050 regions for KIRC to 6,458 regions for ESCA (esophageal carcinoma) (Figure 6E). These numbers align with previously identified active enhancers in TCGA cancers.[34] The predictive power of these regions was confirmed as we accurately separated tumors from normal tissues in most cancers using a machine learning classifier (Figure 6F; STAR Methods). To uncover pan-cancer intergenic transcriptional markers that could contribute to tumorigenesis across multiple cancer types, we identified RNAPII-bound regions differentially expressed in a substantial number of cancers (7 or more of 16; Figure S19B; STAR Methods). We observed a large number of RNAPII-bound regions that did not appear to be differentially expressed in any specific type of cancer. However, on the other end of the spectrum, we observed a significant number of RNAPII-bound regions that exhibited differential expression in a greater number of cancers than expected. Specifically, we found 10,940 RNAPII-bound regions to be differentially expressed in more than seven cancers, with some expressed in each of the 16 cancers that had corresponding normal tissue samples available. Within this set of 10,940 pan-cancer differentially expressed RNAPII-bound regions, we identified previously known regions implicated in cancer as well as new loci (Figures 6G and 6H). For example, we identified two pan-cancer differentially expressed RNAPII-bound regions on enhancers located 10 kb upstream of the MDM4 gene (Figure 6G). This protein is involved in repression of the tumor suppressor TP53 and represents a potential therapeutic target in liver cancer[35] and lymphomas[36] and overall in anticancer therapy.[37] Additionally, we highlighted a group of pan-cancer RNAPII-bound regions that overlapped a known, frequently mutated[38] FOXA1 enhancer region involved in proliferation of prostate cancer cells (Figure 6H). This region has been identified as one of six *cis*-regulatory elements in the FOXA1 regulatory plexus harboring somatic single-nucleotide variants in primary prostate tumors.[38] FOXA1 acts as a pioneer factor in prostate cancer and governs expression of cell cycle regulatory genes in prostate cancer. Overall, these 10,940 regions appear to be located near cancer hallmark genes

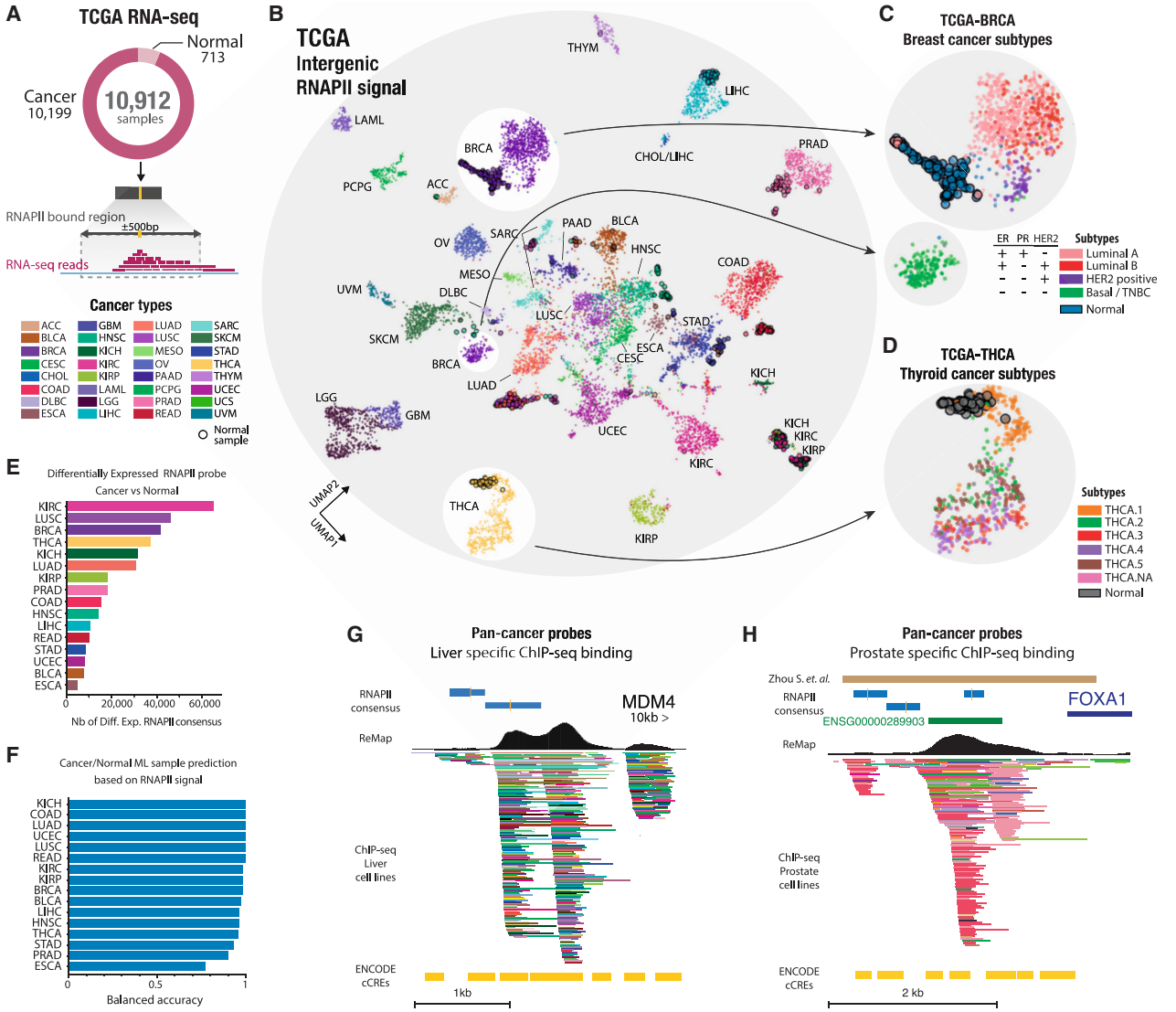


Figure 6. Cancer type and subtype classification by intergenic transcription at RNAPII binding sites

(A) A total of 10,912 TCGA RNA-seq samples were leveraged to capture intergenic signals at standardized RNAPII 1-kb bound regions.

(B) A two-dimensional UMAP of 10,912 TCGA patients based on intergenic RNAPII transcriptional signals. Each dot represents a TCGA cancer patient or normal sample, with the colors representing the cancer type. White circles highlight breast cancer (BRCA) and thyroid carcinoma (THCA) samples.

(C and D) Magnified projections of distinct BRCA and THCA patients (dots) colored by subtype categories based on intergenic transcriptional signals. Normal samples have larger solid black outlines.

(E) Number of tumor-specific intergenic RNAPII-bound regions differentially expressed in tumors compared with normal samples.

(F) Machine learning classification performance (balanced accuracy) between normal and tumor samples for each cancer type.

(G) Genomic view of a pan-cancer intergenic RNAPII-bound region differentially expressed in seven or more cancers. Two pan-cancer markers are located on enhancers (enhancer distal, cCREs) near the MDM4 gene with ChIP-seq bindings.

(H) The brown bar represents a published *cis*-regulatory element of FOXA1 harboring somatic variants in primary prostate tumors.[38] ChIP-seq ReMap tracks are filtered to show transcription factor (TF) binding specifically in liver or prostate cell lines.

(Figure S19C). Our analysis revealed differentially expressed intergenic markers in tumors or tumors subtypes compared with normal tissues, which may directly or indirectly contribute to tumorigenesis. By identifying potential intergenic transcriptional markers, our findings could pave the way for novel therapeutic strategies targeting clinically actionable genes.

Intergenic transcriptional markers showing clinical relevance in cancer

To examine the clinical relevance of intergenic transcriptional markers, we investigated the association between expression of RNAPII-bound regions and overall survival per cancer and pan cancer using a Cox proportional hazard model (per- and

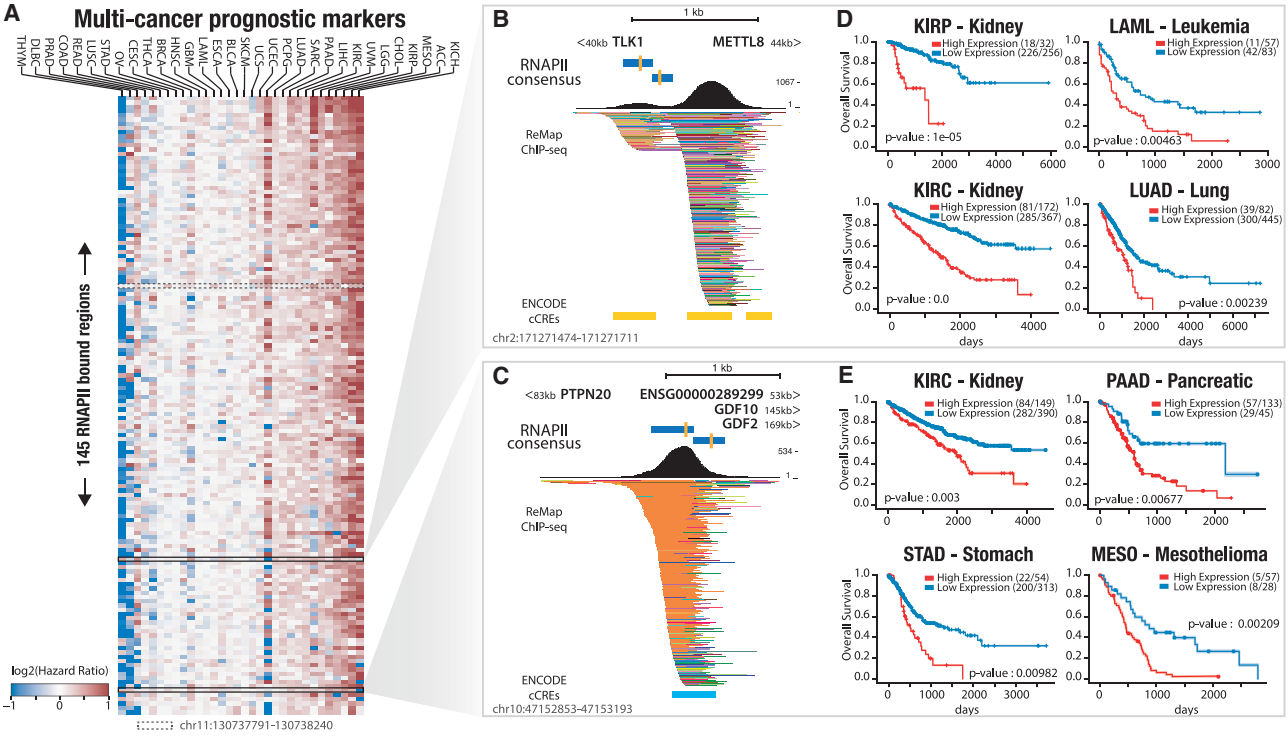


Figure 7. Intergenic transcriptional markers showing clinical relevance in cancer

(A) Heatmap of 145 transcribed RNAPII-bound regions identified as prognostic markers in multiple cancers. A color scale depicts log2(hazard ratios) of strong expression associated with a good (blue) or bad (red) prognosis. Black rectangles highlight two intergenic prognostic markers (RNAPII-bound regions in B and C). A dashed-line rectangle highlights a prognostic marker shown in the supplementary.

(B) Genomic landscape of identified multi-cancer prognostic markers (blue bars) at chr2:171,271,474–171,271,711 located 40 kb downstream of the TLK1 gene and 44 kb upstream of the METTL8 gene. Yellow bars indicate candidate *cis*-regulatory elements (cCREs, enhancer distal) and ChIP-seq binding from ReMap.

(C) Genomic view of the multicancer prognostic markers (chr2:171,271,474–171,271,711) located 83 kb downstream of the PTPN20 gene and 145/196 kb downstream of the GDF10 and GDF2 genes as well as 53 kb downstream of a, lncRNA gene, ENSG00000289299. A light blue bar indicates a cCRE CTCF region.

(D) Kaplan-Meier survival analysis of kidney cancer (papillary and clear), leukemia, and lung cancer patients with high (red) and low (blue) expression from the intergenic RNAPII-bound region in (B).

(E) Kaplan-Meier survival analysis of kidney, pancreatic, stomach, and mesothelioma TCGA cancer patients with high (red) and low (blue) expression from the intergenic RNAPII-bound region in (C).

pan-cancer marker lists and count tables are available at Zenodo[18]). At the per-cancer level, our results showed a smaller number of RNAPII-bound regions associated with overall survival compared with previous analyses of differentially expressed RNAPII regions. The largest number of associated regions was observed in LGG (n = 18,380), with an average of 2,002 regions per cancer (Figure S20A). At the pan-cancer level, we identified a set of 145 RNAPII-bound regions associated with overall survival in five or more cancer types. Most of these regions showed a positive association between overexpression and poor survival (hazard ratio > 1; Figure 7A). The 145 RNAPII-bound regions identified were found to be in close proximity to genes involved in the cell cycle, DNA metabolism and repair, and muscle development as well as hallmark genes of genome instability and mutation (Figures S20B and S20C). Perturbation and acceleration of the cell cycle are hallmarks of cancer and play a role in tumor progression and prognosis. As examples, we highlight two RNAPII-bound regions associated with overall survival (OS) and located near known cancer-associated genes and candidate regulatory elements (Figures 7B, 7C, and S21).

The first RNAPII-bound region is located between the genes TLK1 and METTL8 at 40 kb and 44 kb, respectively (Figure 7B). TLK1 has been linked to poor patient outcomes in multiple cancer types, including GBM[39] and prostate cancer metastasis,[40,41] and it is involved in DNA replication and chromatin assembly.[39] METTL8 has been identified as a potential biomarker in hepatocellular carcinoma,[42] and high levels have been associated with improved patient survival in pancreatic cancer.[42] We observed that high expression of the pan-cancer RNAPII-bound region depicted in Figure 7B is strongly linked to survival in leukemia and kidney (KIRP and KIRC) and lung cancers (Figure 7D).

The second region is located between three genes: downstream of a protein tyrosine phosphatase non-receptor (PTPN20) at 83 kb and upstream of two growth differentiation factors, GDF10 and GDF2, at 145 kb and 196 kb, respectively (Figure 7C). GDF10 and GDF2 belong to the transforming growth factor β (TGF-β) superfamily and are considered tumor

suppressors[43] in certain cancers. Studies have demonstrated that GDF10 expression is an independent prognostic factor for OS of patients with oral squamous cell carcinoma.[44] Additionally, GDF10 inhibits cell proliferation and epithelial-mesenchymal transition in nasopharyngeal carcinoma.[45] On the other hand, GDF2 exhibits pleiotropic effects in tumorigenesis, promoting ovarian cancer cell growth[46] while suppressing breast tumorigenesis[47] and increasing hepatocellular carcinoma cell growth.[48] It also plays a role in suppressing cell death in ovarian and breast epithelia.[49] Furthermore, numerous protein tyrosine phosphatases have been shown to regulate essential cellular processes, with several mutations associated with human diseases.[50] This pan-cancer RNAPII-bound region demonstrates not only a previously established correlation with mesothelioma cancers but also a new correlation between its expression and survival rates in kidney, pancreatic, and stomach cancers (Figure 7E). Taken together, these analyses suggest that these transcribed RNAPII regions, which are mostly unreferenced and undetected, may have clinically relevant roles in cancer and could serve as potential markers for OS. Additional studies are needed to fully understand the potential clinical implications of these observations.

DISCUSSION

We constructed an atlas of intergenic transcription at RNAPII binding sites to connect genomic, transcriptomic, and clinical data across normal tissues and cancer samples. Our approach utilizes a normalized vocabulary for cell lines and tissue types and integrates a compendium of 906 publicly available RNAPII ChIP-seq profiles, enabling comprehensive exploration of intergenic transcription across 28,000 expression samples. The atlas provides an efficient means to investigate tissue specificity and the activity of core regulatory elements in various tissues. Our meta-clustering approach reveals that transcription of intergenic regions is shared among similar tissues and across multiple independent resources. We identified per-cancer and pan-cancer intergenic transcriptional markers associated with known cancer genes and prognostic intergenic markers that predict overall patient survival. Additionally, we discovered that intergenic transcriptional markers can effectively discriminate between subtypes of breast and thyroid cancers.

Our mapping of intergenic transcription stands out from prior efforts to characterize enhancer activities because we directly target the RNAPII transcriptional machinery. Traditionally, studies have relied on single resources, such as histone signatures from ENCODE or Cap Analysis of Gene Expression (CAGE) transcripts from FANTOM, to identify non-coding elements. However, our study demonstrates the effectiveness of robust data integration using diverse public RNAPII ChIP-seq datasets, providing a coherent method to characterize intergenic transcriptional activity in normal and cancer tissues. To detect non-coding transcription, previous investigations[3,8,27,51] have utilized techniques such as GRO-seq or its derivatives to capture nascent transcripts. However, these techniques have limitations in terms of coverage and representation of normal tissues and cancer types. In our study, we address these limitations by integrating GTEx, TCGA, and ENCODE RNA-seq data, offering new insights into intergenic activity across cell lines, normal tissues, and cancer types.

The intergenic RNAPII consensus peaks observed in our study exhibit characteristics suggestive of enhancers, potential TSSs, or potential transcription termination sites (TTSs) that have yet to be described. The majority of identified intergenic RNAPII consensus peaks were classified as regulatory regions, with 65.9% displaying an enhancer signature and 17.4% located downstream of genes. These findings align with the current understanding of the non-coding genome, where enhancers and other regulatory elements are known to recruit RNAPII. In this study, we annotate all 181,547 intergenic RNAPII consensus with genomic characteristics and biotype signatures. Our large-scale integration approach enabled comprehensive exploration of intergenic transcription in normal tissues and cancer types.

Using signals from cancer transcriptomes (TCGA), we identified differentially expressed RNAPII intergenic regions and molecular subtypes of breast and thyroid cancers. For example, in TNBC, we showed that certain differentially expressed RNAPII regions are located near DUSP genes involved in the MAPK signaling pathway. This pathway plays a crucial role in regulating cell proliferation and apoptosis, and DUSP1 in particular may contribute to chemoresistance in TNBC.[30,31] While we demonstrated that RNAPII consensus targets intergenic enhancer elements or proximal enhancers upstream of genes, we also observed RNAPII consensus located downstream of gene TTSs. Future investigations may help identify new sites of transcription termination across our biotype panel.

Transcription of non-coding regions is a fundamental characteristic captured by our RNAPII intergenic map across cell lines, normal tissues, and cancer samples. This significantly expands the analysis horizon beyond gene-centric annotations. Our integration framework symbolizes a transition from exploratory studies centered around uncovering new regulatory elements to a map-focused phase that prioritizes identification of active transcribed elements within specific biological contexts. The significance of our study lies in its ability to enhance our understanding of the activity of non-coding regions in cancer biology and disease development, potentially guiding therapeutic approaches and ultimately improving patient outcomes.

Limitations of the study

Despite the valuable insights gained from our study, there are some limitations that should be considered. One limitation is the reliance on publicly available RNAPII ChIP-seq datasets, which may introduce potential biases and variations in experimental conditions. Additionally, the analysis focused on RNAPII-bound regions and their transcriptional activity, but the nature of the produced transcripts or other factors, such as chromatin conformation, were not directly addressed. Moreover, the use of transcriptomic data from public databases may not fully represent all tissue types and cancer subtypes, potentially limiting the generalizability of our findings. While our meta-clustering approach allowed us to integrate diverse datasets, some tissues or cell types may still have limited representation, affecting the accuracy of tissue-specific classifications. Future experiments with larger and more diverse datasets would be necessary to validate and expand the conclusions drawn from

this study. Nonetheless, our atlas of intergenic transcription at RNAPII binding sites offers a valuable resource for investigating tissue-specific regulatory elements and holds promise for advancing our understanding of non-coding transcription in normal and cancer tissues.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - RNAPII ChIP-seq data processing
  - High level biosample annotation
  - Construction of the intergenic RNAPII atlas
  - Comparison with reference databases from other large-scale efforts
  - Annotation of RNAPII consensus
  - RNAPII atlas visualisation and clustering
  - Extending the integrative approach to H3K27ac ChIP-seq
  - Epigenetic enrichments
  - Gene Ontology enrichments of nearby genes
  - TF motifs identification
  - GWAS traits and summary statistics
  - RNA-seq expression quantification of RNAPII-bound regions
  - Count normalisation and transformation
  - Reads distribution on standardised RNAPII consensus
  - Unsupervised feature selection, dimensionality reduction and predictive models
  - Identification of per tissue markers and ''meta-clustering''
  - Tissue specific eQTL enrichments
  - Differential expression
  - Using GTEx normal tissues instead of TCGA normal tissue
  - Identification of pan-cancer markers
  - Survival analysis
  - Enrichment of cancer hallmark-related genes

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100411.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Paul Flicek for providing constructive criticism of the manuscript and Dr. Sabrina Baaklini for engaging discussions on meta-clustering. We thank Dr. Jeremy J. Day (UAB Heersink School of Medicine) for granting us permission to reproduce the DNA schematic in Figure 1A. This work was supported by a PhD fellowship to P.d.L. and J.-C.M. from the French Ministry of Higher Education and Research (MESR), a 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), the Core Cluster of the Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013), and the Centre de Calcul Intensif d'Aix-Marseille by granting access to its high performance computing resources. The results shown here are based on data generated by the TCGA Research Network, the GTEx project, the ENCODE Consortium, the ENCODE production laboratories, and independent laboratories that adhered to the Open Science principles and submitted raw ChIP-seq data to repositories.

AUTHOR CONTRIBUTIONS

B.B. initiated, coordinated, and supervised the project. P.d.L. manually curated the RNAPII datasets, and E.G. assisted with primary data curation. F.H. performed ChIP-seq reprocessing. J.-C.M. collected and compiled the STARR-seq catalog. P.d.L. performed computational method development and analyzed the data under supervision from L.S. and B.B. P.d.L. and B.B. generated figures. P.d.L. and B.B. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 14, 2023
Revised: June 29, 2023
Accepted: September 4, 2023
Published: September 29, 2023

REFERENCES

1. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. https://doi.org/10.1038/nature11247.
2. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R., and Johnson, R. (2018). Towards a complete map of the human long non-coding RNA transcriptome. Nat. Rev. Genet. *19*, 535–548. https://doi.org/10.1038/s41576-018-0017-y.
3. Kim, T.-K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182–187. https://doi.org/10.1038/nature09033.
4. Anastasiadou, E., Jacob, L.S., and Slack, F.J. (2018). Non-coding RNA networks in cancer. Nat. Rev. Cancer *18*, 5–18. https://doi.org/10.1038/nrc.2017.99.
5. Salta, E., and De Strooper, B. (2012). Non-coding RNAs with essential roles in neurodegenerative disorders. Lancet Neurol. *11*, 189–200. https://doi.org/10.1016/S1474-4422(11)70286-1.
6. Ward, L.D., and Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. Nat. Biotechnol. *30*, 1095–1106. https://doi.org/10.1038/nbt.2422.
7. Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461. https://doi.org/10.1038/nature12787.
8. Hah, N., Murakami, S., Nagari, A., Danko, C.G., and Kraus, W.L. (2013). Enhancer transcripts mark active estrogen receptor binding sites. Genome Res. *23*, 1210–1223. https://doi.org/10.1101/gr.152306.112.
9. Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat. Genet. *46*, 1311–1320. https://doi.org/10.1038/ng.3142.
10. Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A.Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., et al. (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature *498*, 516–520. https://doi.org/10.1038/nature12210.

11. Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–1848. https://doi.org/10.1126/science.1162228.

12. Mahat, D.B., Kwak, H., Booth, G.T., Jonkers, I.H., Danko, C.G., Patel, R.K., Waters, C.T., Munson, K., Core, L.J., and Lis, J.T. (2016). Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat. Protoc. *11*, 1455–1476. https://doi.org/10.1038/nprot.2016.086.

13. Yao, L., Liang, J., Ozer, A., Leung, A.K.-Y., Lis, J.T., and Yu, H. (2022). A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers. Nat. Biotechnol. *40*, 1056–1065. https://doi.org/10.1038/s41587-022-01211-7.

14. 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. https://doi.org/10.1093/nar/gks1193.

15. GTEx Consortium; Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585. https://doi.org/10.1038/ng.2653.

16. Cancer Genome Atlas Research Network; Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120. https://doi.org/10.1038/ng.2764.

17. Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. *9*, 9354. https://doi.org/10.1038/s41598-019-45839-z.

18. de Langen, P., Hammal, F., Guéret, E., Mouren, J.-C., Spinelli, L., and Ballester, B. (2023). Intergenic RNAPII Atlas : Output Data. https://doi.org/10.5281/zenodo.8091826.

19. Hammal, F., de Langen, P., Bergon, A., Lopez, F., and 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 Res. *50*, D316–D325. https://doi.org/10.1093/nar/gkab996.

20. 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. https://doi.org/10.1038/s41586-020-2493-4.

21. FANTOM Consortium and the RIKEN PMI and CLST DGT; Forrest, A.R.R., Kawaji, H., Rehli, M., Baillie, J.K., de Hoon, M.J.L., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., et al. (2014). A promoter-level mammalian expression atlas. Nature *507*, 462–470. https://doi.org/10.1038/nature13182.

22. Noguchi, S., Arakawa, T., Fukuda, S., Furuno, M., Hasegawa, A., Hori, F., Ishikawa-Kato, S., Kaida, K., Kaiho, A., Kanamori-Katayama, M., et al. (2017). FANTOM5 CAGE profiles of human and mouse samples. Sci. Data *4*, 170112. https://doi.org/10.1038/sdata.2017.112.

23. Volders, P.-J., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdagh, P., and Vandesompele, J. (2019). LNCipedia 5: towards a reference set of human long non-coding RNAs. Nucleic Acids Res. *47*, D135–D139. https://doi.org/10.1093/nar/gky1031.

24. Bairoch, A. (2018). The Cellosaurus, a Cell-Line Knowledge Resource. J. Biomol. Tech. *29*, 25–38. https://doi.org/10.7171/jbt.18-2902-002.

25. McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://doi.org/10.48550/arXiv.1802.03426.

26. Meuleman, W., Muratov, A., Rynes, E., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Teodosiadis, A., et al. (2020). Index and biological spectrum of human DNase I hypersensitive sites. Nature *584*, 244–251. https://doi.org/10.1038/s41586-020-2559-3.

27. de Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.L., and Natoli, G. (2010). A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. PLoS Biol. *8*, e1000384. https://doi.org/10.1371/journal.pbio.1000384.

28. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004.

29. Schwalb, B., Michel, M., Zacher, B., Frühauf, K., Demel, C., Tresch, A., Gagneur, J., and Cramer, P. (2016). TT-seq maps the human transient transcriptome. Science *352*, 1225–1228. https://doi.org/10.1126/science.aad9841.

30. Li, J., Chen, Y., Yu, H., Tian, J., Yuan, F., Fan, J., Liu, Y., Zhu, L., Wang, F., Zhao, Y., and Pang, D. (2017). DUSP1 promoter methylation in peripheral blood leukocyte is associated with triple-negative breast cancer risk. Sci. Rep. *7*, 43011. https://doi.org/10.1038/srep43011.

31. Shen, J., Zhang, Y., Yu, H., Shen, B., Liang, Y., Jin, R., Liu, X., Shi, L., and Cai, X. (2016). Role of DUSP1/MKP1 in tumorigenesis, tumor progression and therapy. Cancer Med. *5*, 2061–2068. https://doi.org/10.1002/cam4.772.

32. McGuire, A., Brown, J.A.L., Malone, C., McLaughlin, R., and Kerin, M.J. (2015). Effects of age on the detection and management of breast cancer. Cancers *7*, 908–929. https://doi.org/10.3390/cancers7020815.

33. Zagami, P., and Carey, L.A. (2022). Triple negative breast cancer: Pitfalls and progress. npj Breast Cancer *8*, 95–10. https://doi.org/10.1038/s41523-022-00468-0.

34. Chen, H., Li, C., Peng, X., Zhou, Z., Weinstein, J.N., and Cancer Genome Atlas Research Network; and Liang, H. (2018). A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. Cell *173*, 386–399.e12. https://doi.org/10.1016/j.cell.2018.03.027.

35. Woodfield, S.E., Shi, Y., Patel, R.H., Chen, Z., Shah, A.P., Srivastava, R.K., Whitlock, R.S., Ibarra, A.M., Larson, S.R., Sarabia, S.F., et al. (2021). MDM4 inhibition: a novel therapeutic strategy to reactivate p53 in hepatoblastoma. Sci. Rep. *11*, 2967. https://doi.org/10.1038/s41598-021-82542-4.

36. Hüllein, J., Słabicki, M., Rosolowski, M., Jethwa, A., Habringer, S., Tomska, K., Kurilov, R., Lu, J., Scheinost, S., Wagener, R., et al. (2019). MDM4 Is Targeted by 1q Gain and Drives Disease in Burkitt Lymphoma. Cancer Res. *79*, 3125–3138. https://doi.org/10.1158/0008-5472.CAN-18-3438.

37. Toledo, F., and Wahl, G.M. (2007). MDM2 and MDM4: p53 regulators as targets in anticancer therapy. Int. J. Biochem. Cell Biol. *39*, 1476–1482. https://doi.org/10.1016/j.biocel.2007.03.022.

38. Zhou, S., Hawley, J.R., Soares, F., Grillo, G., Teng, M., Madani Tonekaboni, S.A., Hua, J.T., Kron, K.J., Mazrooei, P., Ahmed, M., et al. (2020). Noncoding mutations target cis-regulatory elements of the FOXA1 plexus in prostate cancer. Nat. Commun. *11*, 441. https://doi.org/10.1038/s41467-020-14318-9.

39. Ibrahim, K., Abdul Murad, N.A., Harun, R., and Jamal, R. (2020). Knockdown of Tousled-like kinase 1 inhibits survival of glioblastoma multiforme cells. Int. J. Mol. Med. *46*, 685–699. https://doi.org/10.3892/ijmm.2020.4619.

40. Khalil, M.I., and De Benedetti, A. (2022). The TLK1–MK5 Axis Regulates Motility, Invasion, and Metastasis of Prostate Cancer Cells. Cancers *14*, 5728. https://doi.org/10.3390/cancers14235728.

41. Khalil, M.I., Singh, V., King, J., and De Benedetti, A. (2022). TLK1-mediated MK5-S354 phosphorylation drives prostate cancer cell motility and may signify distinct pathologies. Mol. Oncol. *16*, 2537–2557. https://doi.org/10.1002/1878-0261.13183.

42. Schöller, E., Marks, J., Marchand, V., Bruckmann, A., Powell, C.A., Reichold, M., Mutti, C.D., Dettmer, K., Feederle, R., Hüttelmaier, S., et al. (2021). Balancing of mitochondrial translation through METTL8-mediated m3C modification of mitochondrial tRNAs. Mol. Cell *81*, 4810–4825.e12. https://doi.org/10.1016/j.molcel.2021.10.018.

43. Upadhyay, G., Yin, Y., Yuan, H., Li, X., Derynck, R., and Glazer, R.I. (2011). Stem cell antigen-1 enhances tumorigenicity by disruption of

growth differentiation factor-10 (GDF10)–dependent TGF-β signaling. Proc. Natl. Acad. Sci. USA *108*, 7820–7825. https://doi.org/10.1073/pnas.1103441108.

44. Cheng, C.-W., Hsiao, J.-R., Fan, C.-C., Lo, Y.-K., Tzen, C.-Y., Wu, L.-W., Fang, W.-Y., Cheng, A.-J., Chen, C.-H., Chang, I.-S., et al. (2016). Loss of GDF10/BMP3b as a prognostic marker collaborates with TGFBR3 to enhance chemotherapy resistance and epithelial-mesenchymal transition in oral squamous cell carcinoma. Mol. Carcinog. *55*, 499–513. https://doi.org/10.1002/mc.22297.
45. He, F., Feng, G., Ma, N., Midorikawa, K., Oikawa, S., Kobayashi, H., Zhang, Z., Huang, G., Takeuchi, K., and Murata, M. (2022). GDF10 inhibits cell proliferation and epithelial-mesenchymal transition in nasopharyngeal carcinoma by the transforming growth factor-β/Smad and NF-κB pathways. Carcinogenesis *43*, 94–103. https://doi.org/10.1093/carcin/bgab122.
46. Herrera, B., van Dinther, M., ten Dijke, P., and Inman, G.J. (2009). Autocrine Bone Morphogenetic Protein-9 Signals through Activin Receptor-like Kinase-2/Smad1/Smad4 to Promote Ovarian Cancer Cell Proliferation. Cancer Res. *69*, 9254–9262. https://doi.org/10.1158/0008-5472.CAN-09-2912.
47. Ren, W., Sun, X., Wang, K., Feng, H., Liu, Y., Fei, C., Wan, S., Wang, W., Luo, J., Shi, Q., et al. (2014). BMP9 inhibits the bone metastasis of breast cancer cells by downregulating CCN2 (connective tissue growth factor, CTGF) expression. Mol. Biol. Rep. *41*, 1373–1383. https://doi.org/10.1007/s11033-013-2982-8.
48. Herrera, B., García-Álvaro, M., Cruz, S., Walsh, P., Fernández, M., Roncero, C., Fabregat, I., Sánchez, A., and Inman, G.J. (2013). BMP9 is a proliferative and survival factor for human hepatocellular carcinoma cells. PLoS One *8*, e69535. https://doi.org/10.1371/journal.pone.0069535.
49. Varadaraj, A., Patel, P., Serrao, A., Bandyopadhay, T., Lee, N.Y., Jazaeri, A.A., Huang, Z., Murphy, S.K., and Mythreye, K. (2015). Epigenetic Regulation of GDF2 Suppresses Anoikis in Ovarian and Breast Epithelia. Neoplasia *17*, 826–838. https://doi.org/10.1016/j.neo.2015.11.003.
50. Tautz, L., Critton, D.A., and Grotegut, S. (2013). Protein Tyrosine Phosphatases: Structure, Function, and Implication in Human Disease. Methods Mol. Biol. *1053*, 179–221. https://doi.org/10.1007/978-1-62703-562-0_13.
51. Franco, H.L., Nagari, A., Malladi, V.S., Li, W., Xi, Y., Richardson, D., Allton, K.L., Tanaka, K., Li, J., Murakami, S., et al. (2018). Enhancer transcription reveals subtype-specific gene expression programs controlling breast cancer pathogenesis. Genome Res. *28*, 159–170. https://doi.org/10.1101/gr.226019.117.
52. de Langen, P., Hammal, F., Guéret, E., Mouren, J.-C., Spinelli, L., and Ballester, B. (2023). Intergenic RNAPII Atlas : Input Data. https://doi.org/10.5281/zenodo.7785392.
53. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.
54. 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. https://doi.org/10.1101/gr.136184.111.
55. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137. https://doi.org/10.1186/gb-2008-9-9-r137.
56. 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. https://doi.org/10.1093/nar/gkaa1087.
57. Lee, B.T., Barber, G.P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J.N., Hinrichs, A.S., Lee, C.M., et al. (2022). The UCSC Genome Browser database: 2022 update. Nucleic Acids Res. *50*, D1115–D1122. https://doi.org/10.1093/nar/gkab959.
58. Stovner, E.B., and Sætrom, P. (2020). PyRanges: efficient comparison of genomic intervals in Python. Bioinformatics *36*, 918–919. https://doi.org/10.1093/bioinformatics/btz615.
59. Tamura, Y., Obara, N., and Miyamoto, S. (2014). A Method of Two-Stage Clustering with Constraints Using Agglomerative Hierarchical Algorithm and One-Pass k-Means++. In Knowledge and Systems Engineering Advances in Intelligent Systems and Computing, V.N. Huynh, T. Denoeux, D.H. Tran, A.C. Le, and S.B. Pham, eds. (Springer International Publishing), pp. 9–19. https://doi.org/10.1007/978-3-319-02821-7_3.
60. Müllner, D. (2013). fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. J. Stat. Software *53*, 1–18. https://doi.org/10.18637/jss.v053.i09.
61. Dong, W., Moses, C., and Li, K. (2011). Efficient k-nearest neighbor graph construction for generic similarity measures. In Proceedings of the 20th international conference on World wide web WWW '11, pp. 577–586, (Association for Computing Machinery). https://doi.org/10.1145/1963405.1963487.
62. Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci. Rep. *9*, 5233. https://doi.org/10.1038/s41598-019-41695-z.
63. 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. https://doi.org/10.1038/nature14248.
64. Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187–W191. https://doi.org/10.1093/nar/gku365.
65. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501. https://doi.org/10.1038/nbt.1630.
66. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235. https://doi.org/10.1038/ng.3404.
67. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209. https://doi.org/10.1038/s41586-018-0579-z.
68. Pan-UKB team (2020). https://pan.ukbb.broadinstitute.org.
69. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.
70. Lun, A.T.L., Bach, K., Marioni, J.C., and Marioni, J.C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. *17*, 75. https://doi.org/10.1186/s13059-016-0947-7.
71. Choudhary, S., and Satija, R. (2022). Comparison and evaluation of statistical error models for scRNA-seq. Genome Biol. *23*, 27. https://doi.org/10.1186/s13059-021-02584-9.
72. Dinno, A. (2009). Exploring the Sensitivity of Horn's Parallel Analysis to the Distributional Form of Random Data. Multivariate Behav. Res. *44*, 362–388. https://doi.org/10.1080/00273170902938969.
73. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.
74. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
UK Biobank	UK Biobank[67]	https://www.ukbiobank.ac.uk/
UK Biobank GWAS summary statistics	UK Biobank[67]	http://www.nealelab.is/uk-biobank
UK Biobank GWAS summary statistics	UK Biobank[67,68]	https://pan.ukbb.broadinstitute.org
RNAPII ChIP-seq summary	This paper	Table S1
Tissue-wide RNA-seq expression	GTEx project[15]	https://gtexportal.org/home/datasets
Cell lines RNA-seq expression	ENCODE project[1,20]	https://www.encodeproject.org/
Cancer samples RNA-seq expression	TCGA project[16]	http://gdc.cancer.gov
Tissue specific eQTL v8	GTEx project[15]	https://gtexportal.org/home/datasets
Human genome annotation v38	GENCODE project[56]	https://www.gencodegenes.org/
Human ChIP-seq peaks	ReMap project[19]	http://remap.univ-amu.fr/
ENCODE backlisted regions	Amemiya et al.[17]	https://github.com/Boyle-Lab/Blacklist/
Human candidate Cis Regulatory Elements	ENCODE project[1,20]	https://www.encodeproject.org/
DNase I hypersensitive	Meuleman et al.[26]	https://doi.org/10.1038/s41586-020-2559-3
Long non-coding RNA encyclopedia v5	LNCipedia[23]	https://lncipedia.org/
Transcribed Enhancers CAGE eRNAs	FANTOM5 project[21,22]	https://fantom.gsc.riken.jp/
ENCODE STARR-seq	ENCODE project[1,20]	https://www.encodeproject.org/
Repeat elements	UCSC Genome Browser[57]	https://genome.ucsc.edu/
H3K27Ac ChIP-seq experiments	ENCODE project[1,20]	Accession IDs are in Table S2
ATAC-seq experiments	ENCODE project[1,20]	Accession IDs are in Table S2
Epigenome states (Core 15-states model)	ROADMAP project[63]	https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html
Cancer Hallmarks Genes database	CHG database	http://www.bio-bigdata.com/CHG/
Software and algorithms		
Bowtie2 (2.4.2)	Langmead and Salzberg[53]	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Samtools (1.11)	Li et al.[74]	http://samtools.sourceforge.net/
DESeq2 (1.40.2)	Love et al.[73]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
ReMap ChIP-seq pipeline (2022)	Hammal et al.[19]	https://github.com/benoitballester/
Trim Galore (0.6.5)	Babraham Bioinformatics	https://github.com/FelixKrueger/TrimGalore
MACS2 (v2.1.2)	Zhang et al.[55]	https://github.com/macs3-project/MACS
Python (3.6.12)	Python Core Team	https://www.python.org/
Python PyRanges library (0.0.129)	Stovner and Saestrom[58]	https://pypi.org/project/pyranges/
Python fastcluster library (1.2.6)	Müllner, D.[60]	https://pypi.org/project/fastcluster/
Python Pynndescent library (0.5.10)	Dong et al.[61]	https://pypi.org/project/pynndescent/
Python Statsmodels library (0.14.0)	Python Package Index (PyPI)	https://pypi.org/project/statsmodels/
Python Lifelines library (0.27.7)	Python Package Index (PyPI)	https://pypi.org/project/lifelines/
Python kaplanmeier library (0.1.9)	Python Package Index (PyPI)	https://pypi.org/project/kaplanmeier/
Pandas	https://pandas.pydata.org	RRID:SCR_018214
Numpy	http://www.numpy.org	RRID:SCR_008633
Scipy	https://www.scipy.org	RRID:SCR_008058
scikit-learn	http://scikit-learn.org	RRID:SCR_002577
R (4.0.5)	R Core Team	https://www.R-project.org
R maxstat library (0.7–25)	CRAN project	https://cran.r-project.org/package = maxstat
deepTools (3.5.2)	Ramirez et al.[64]	https://github.com/deeptools/deepTools

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
GREAT (v4.0.4)	Mc Lean et al.[65]	http://great.stanford.edu/public/html/
HOMER (v4.11)	Heinz et al.[28]	http://homer.ucsd.edu/homer/
Stratified LD-Score Regression	Finucane et al.[66]	https://doi.org/10.1038/ng.3404
featureCounts	Liao et al.[69]	https://doi.org/10.1093/bioinformatics/btt656
SCTransform	Choudhary et al.[71]	https://doi.org/10.1186/s13059-021-02584-9
Other		
Code for main figures and analysis	This paper; Github	https://github.com/benoitballester/Pol2Atlas
Intergenic RNAPII Atlas: input data	This paper; Zenodo	https://zenodo.org/record/7785393
Intergenic RNAPII Atlas: output data	This paper; Zenodo	https://zenodo.org/record/8091826
Intergenic RNAPII Atlas: annotated consensus	This paper	Data S1

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Benoit Ballester (benoit.ballester@inserm.fr).

Materials availability

This study did not generate new unique reagents.

Data and code availability

RNA polymerase II ChIP-seq data are publicly available in NCBI-GEO, and data accessions for ChIP-seq are listed in Table S1. The GTEx[15] eQTL data were obtained from GTEx v8. Human regulatory TF catalog was obtained from ReMap 2022 release.[19] ENCODE RNA-seq raw sequencing data (Accession IDs in Table S4) are available at https://www.encodeproject.org/. TCGA and GTEx RNA-seq raw sequencing data are available under controlled access to ensure appropriate data usage. Access to these protected data must be requested through the dbGaP portal. The Cancer Genome Atlas[16] (TCGA) RNA-seq BAM files are accessible through dbGaP under accession no. phs000178.v11.p8.c1 (TCGA) and at NCI's Genomic Data Commons (http://gdc.cancer.gov) under project TCGA. Genotype-Tissue Expression (GTEx) RNA-seq BAM files are accessible through dbGaP under accession no. phs000424.v8.p2.c1 (GTEx) and at the GTEx portal (https://gtexportal.org/home/). Identified and annotated RNAPII consensus are available in Data S1.

We deposited the codes and bioinformatics environments in GitHub at https://github.com/benoitballester/Pol2Atlas. The processed data matrices and files can be accessed on Zenodo.[18,52] Both data and codes are publicly available for the replication of the whole study.

METHOD DETAILS

RNAPII ChIP-seq data processing

We recovered from NCBI-GEO all existing RNA Polymerase II (RNAPII) ChIP-seq experiments targeting the POLR2A subunit (n = 1,135) in human, following the ReMap procedures and pipeline.[19] Briefly, we manually annotated and standardised the cell line and tissue of origin names (Table S1). Every experiment was downloaded and processed uniformly starting from the fastq files, to quality checks, up to the peak calling stage using the ReMap pipeline. In more detail, ChIP-seq experiments were retrieved from the NCBI Gene Expression Omnibus (GEO) and ENCODE databases. For GEO, the query 'Genome binding/occupancy profiling by high-throughput sequencing' AND 'homo sapiens'[organism] AND NOT 'ENCODE'[project]' was used to return a list of all potential studies. The selected experiments metadata are then manually curated and annotated with official nomenclatures for target names and biotypes. For incomplete metadata, the materials and methods of associated and published papers are often examined to complete the curation. We used the BRENDA Tissue Ontologies for cell lines at the EBI Ontology Lookup Service (www.ebi.ac.uk/ols/ontologies/bto) as well as the Cellosaurus database to homogenize cell and tissue names (e.g., MCF-7 not MCF7, Hep-G2, not HepG2, Hepg2 etc.). We define a dataset as a DNA-binding experiment in a given GEO/ENCODE series (e.g., GSE37345), for a given RNA Polymerase II subunit (e.g., POLR2A), and in a particular biotype (e.g., LNCaP, MCF-7) in a given biological condition. Datasets are labeled with the concatenation of these information (e.g., GSE37345.POLR2A.LNCAP_45min-DMSO). All data were re-analysed starting from raw FASTQ files. Both GEO and ENCODE datasets were manually curated, processed and analyzed in the same way. Bowtie 2[53] (version 2.2.9) with options -end-to-end -sensitive was used to align all reads on the human genome GRCh38/hg38 assembly. 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[54] we used the MACS2[55] peak-calling tool (version 2.1.1.2) to identify the RNAPII-bound regions. For all the datasets, the corresponding bed file is available for download. In order to study only the intergenic part of the genome, we filtered out peaks overlapping GENCODE[56] v38 transcripts ±1kb. We also excluded ENCODE blacklisted regions.[17] We retained peaks with a MACS2 q-value under 10^{-5}, and removed uninformative datasets with less than 100 intergenic peaks. In the end, we conserved 906 out of 1,135 datasets after all Quality Checks (Figures S1B and S1D). Finally we investigated the distribution of antibody usage across the 906 datasets (Figure S1A). The Table S1 includes standardised antibody information, which was manually curated from ENCODE, GEO or the associated paper methods.

High level biosample annotation

Due to the very large biological diversity of the experiments, it is necessary to have a high level annotation to make the interpretation of the results easier, as well as comparing results between datasets. We annotated samples according to their tissue of origin, with the simplified GTEx tissue (30 tissues) annotation as a baseline, to which we added additional tissues: bone, eye, embryo and trachea. To make some results more interpretable, we grouped similar tissues (e.g., various brain tissues into 'Brain') obtaining an annotation with 18 categories (Table S3). A full sample-annotation table is available in Table S1.

Construction of the intergenic RNAPII atlas

A naive approach to delineate groups of RNAPII peaks corresponding to a similar biological signal across experiments would be to merge overlapping peaks. However, when the number of experiments is large, the entire genome becomes covered with peaks which makes this approach impractical. To create consensus RNAPII peaks, we first computed the density function of the peak summits (the single base pair genomic location with the maximum signal of the peak) across each chromosome. Due to the inherent inaccuracy on the summit position of these sequencing techniques and the undersampling, this estimate is extremely noisy. To reduce the amount of noise, we applied a Gaussian filter to this density function across the genome (Figure 1A). Consensus peaks were defined at each local minima of the smoothed density function. A peak belongs to a consensus if its peak summit falls in between the identified flanking local minimas. The boundaries of the defined consensus peak were reduced to the ones of the farthests peaks. By default 1/8th of the average peak size was used as the standard deviation of the Gaussian kernel, and to be valid each consensus was required to contain at least 2 peaks from different experiments. Consensus peaks centroid were defined as the mean position of the peak summits. The middle of the peak was used, if a summit coordinate was not available. A binary data matrix was generated to summarise all datasets. For each consensus peak, this matrix stores if a biosample has a RNAPII peak that belongs to it, similar to the DNAse 1 binary matrix from ENCODE.[26] A schematic of the whole approach is available in Figure S1. Identified and annotated RNAPII consensus are available in Data S1.

Comparison with reference databases from other large-scale efforts

The RNAPII atlas was intersected against GENCODE[56] v38, LNCipedia[23] v5, FANTOM5,[21,22] ReMap[25] 2022, ENCODE cCREs,[20] ENCODE STARR-seq and Repeat elements downloaded from UCSC[57] (hg38). Intersections are computed using the centroid of RNAPII consensus (1bp) against the whole genomic features. The PyRanges python library[58] was used to compute intersections between genomic features. We computed overlap enrichments for the whole dataset using a binomial test where: n, the number of trials, is the number of RNAPII consensus; p, the probability of intersection, is the base pair coverage of the feature of interest divided by the coverage of the intergenic regions (+-1kb from genes, excluding ENCODE blacklisted regions); k, the observed number of successes, is the number of RNAPII consensus intersecting the feature of interest; The fold change is computed as $\frac{k/n}{p}$. We computed overlap enrichments for subsets of the whole RNAPII atlas using a hypergeometric test, which removes the RNAPII-specific intersection bias: where N, the population size, is the number of RNAPII consensus; K, the number of successes in the population, is the number of RNAPII consensus intersecting the feature of interest; n, the number of draws, is the number of RNAPII consensus of the subset of interest; k, the number of observed successes, is the number of RNAPII consensus of the subset of interest intersecting the feature of interest.

Annotation of RNAPII consensus

We performed functional annotation of the 181,547 RNAPII consensus using a simplified approach, where each RNAPII consensus can overlap multiple categories such as Promoter-like, LNC-body, Enhancer-like, Regulatory-like, Gene-tail, Unannotated (Figure S4A). In the following analyses, genomic intersections are performed at the RNAPII consensus centroid against the whole genomic feature. Promoter-like: RNAPII consensus were annotated as promoter-like if they met any of the following criteria: presence of cCREs PLS (Promoter-like Sequence) or cCREs H3K4Me3 (histone H3 lysine 4 trimethylation) or F5 TSS (FANTOM5 Transcription Start Site) or if they overlapped with a LNCipedia promoter, a comprehensive database of long non-coding RNA transcripts. Here, LNCipedia promoters are defined as ±1kb regions around their promoter. LNC-body (potential LNC RNA transcripts): RNAPII consensus were annotated as LNCipedia transcripts if they matched any of the transcripts present in the LNCipedia database, excluding those already annotated as Promoter-like. Here LNCipedia transcripts are extended by 1kb. Enhancer-like: RNAPII consensus were classified as Enhancer-like if they fulfilled the following criteria: presence of cCREs ELS (Enhancer-like Sequence)

or F5 Enhancer (FANTOM5 Enhancers) or if they showed overlap with enhancer regions identified by ENCODE STARR-seq. RNAPII consensus that were already annotated as Promoter-like were excluded from this category. Regulatory-like: RNAPII consensus were labeled as unannotated regulatory if they met the criteria of being present in ReMap CRM or ENCODE DNase (DNase I hypersensitivity sites) datasets. Additionally, they were excluded if they were already classified as Promoter-like, Enhancer-like, or LNC-like. Gene-tail: RNAPII consensus were labeled as gene-tail if they were located +1kb to +9kb downstream of a GENCODE gene. Unannotated: RNAPII consensus that did not fulfill any of the aforementioned criteria were considered unannotated. By applying these specific criteria, we were able to assign functional annotations to 90.9% RNAPII consensus, enabling us to gain insights into their putative regulatory roles and characterising their potential functional significance within the context of our study. In addition, we compared the RNAPII consensus to reference databases through a more detailed and non-exclusive analysis (Figure S6). Identified and annotated RNAPII consensus are available in Data S1.

RNAPII atlas visualisation and clustering

To visualise the similarity between datasets, we applied UMAP with the Yule similarity, with 30 neighbors and the minimum distance set to 0.5. To visualise the similarity between RNAPII consensus, we use the Sorensen-Dice similarity, 30 neighbors and the minimum distance is set to 0. Other parameters were left to default. For the consensus peaks UMAP, to highlight consensus peaks specific to a biotype annotation, each consensus peak was colored by its most frequent biotype annotation. To do so, we compute the sum of the number of peaks per dataset of each annotation at each consensus ($s_{i,j}$), which is then normalised by the total number of peaks for each annotation ($n_{i,j}$): $s_{i,j} = \sum_{k=1}^{N} M_{k \in a_i, j}$ and $n_{i,j} = \frac{s_{i,j}}{\sum s_{i,*}}$, where a_i is a set storing the index of each datasets belonging to annotation i, M is the dataset binary matrix, N the number of experiments, and j the consensus index. This prevents over-represented annotations or annotations with some datasets with a large number of peaks to annotate most of the consensus peaks. We chose as representative for consensus j the annotation i for which $n_{i,j}$ is the largest.

Finally, to identify consensus peaks that are not condition specific, each one is linearly grayed according to its Gini-Simpson index: $\lambda_j = 1 - \sum_{i=1}^{R} p_{i,j}^2$, where R is the number of annotations, and $p_{i,j} = \frac{n_{i,j}}{\sum n_{*,j}}$. The Gini-Simpson index is a measure of diversity: in this study, it tends toward one if the annotations are equidistributed, and is equal to zero if the consensus only has peaks belonging to datasets with the same annotation.

A three steps Hierarchical Clustering (HC) approach was used to order datasets and RNAPII consensus peaks. First, we performed a UMAP dimensionality reduction to 10 dimensions, and used the same metric as the 2D UMAP transform. This step allows the use of any metric, as UMAP optimises to a lower dimensional space using the euclidean distance, which is used by k-means and Ward HC, and also improves k-means and HC quality. Second, we reduced the effective number of points using k-means clustering and grouped very similar points into 50,000 clusters (step performed only when >50,000 points). This approach is documented and allows to scale Ward HC to very large datasets.[59] Third, we performed Ward HC on the k-means clusters centroids using the fastcluster library.[60] The bottom part of the heatmap displays each $p_{i,j}$ as defined in the previous section as a stacked barplot.

A Shared Nearest Neighbor (SNN) Graph Clustering approach was used to identify clusters of RNAPII consensus peaks. This approach is common in single-cell RNA sequencing (scRNA-seq) analyses to identify clusters of cells without *a priori* on the number of clusters. To scale to a large number of points to cluster, we used an Approximate Nearest Neighbor (ANN) method to build the NN graph (python library pynndescent[61]). This approach avoids the quadratic time complexity of building exact nearest neighbors, can use any metric and runs in an almost linear time complexity. The Sorensen-Dice coefficients were used to measure distances between points. In the SNN graph, vertices are weighted by the number of shared nearest neighbors between the two nodes. To identify communities in the SNN graph, we used the Leiden graph clustering algorithm implemented in the python leidenalg[62] library.

Extending the integrative approach to H3K27ac ChIP-seq

We collected all H3K27ac ChIP-seq experiments from ENCODE, retrieved processed files in bed narrowPeak format, mapped for hg38, and without audit error appearing on the sample metadata (n = 890 samples). Each sample was annotated with the same biotype methodology as our RNAPII consensus. The same integrative approach and settings as the RNAPII atlas presented above were run on this dataset.

Epigenetic enrichments

We downloaded every 15 states epigenome available from ROADMAP[63] (hg38). We intersected (consensus centroid only) each consensus peak with each epigenome to get the epigenetic state of each consensus in each epigenome. We computed the proportion of epigenetic states for the subsets/clusters of RNAPII consensus of interest for each epigenome (i.e., the sum of the epigenetic states proportions is equal to one in an epigenome). We used the "GROUP" column of the epigenome metadata to annotate and group epigenomes. We used a paired t-test to statistically assess the difference in proportions between subsets of RNAPII consensus across epigenomes. We downloaded H3K27ac ChIP-seq and ATAC-seq processed bam files from ENCODE for Heart, Liver and T cells samples (Table S2). We used deepTools[64] to compute the mean profiles at each RNAPII consensus of the studied clusters (+-5kb from centroid).

Gene Ontology enrichments of nearby genes

To assign consensus peaks to genes, we used a similar heuristic as GREAT[65] at default settings: a basal domain of 5kb upstream and 1kb downstream, extended in both directions up to 1Mb or the nearest basal domain (whichever is the closest). For each gene we obtained the number of consensus peaks in its regulatory region, for all the consensus peaks (n) and its subset of interest (k). To compute Gene Set enrichments, we used a Negative Binomial GLM: $ln(\mu) = \beta_0 + \beta_1 \times G + E$. Where G is equal to 1 if the studied gene belongs to the Gene Set of interest and 0 otherwise. The term E corrects for the intersection bias of the background regions, with E being the expected number of hits for a particular gene: $E = n * \frac{K}{N}$, with K being the number of query regions, and N the number of background regions. We tested whether β_1 is greater than zero using a Wald Test. The model is fitted using the python statsmodels[53] library. We considered GO terms with more than 3 genes and less than 1,000, and applied the Benjamini-Hochberg FDR correction. The approach is similar to Chip-Enrich and Poly-Enrich, which has shown that gene-wise modeling is required to reduce false discoveries, but these two methods do not offer a model for our case, where the query regions are a subset of a set of background regions. To improve the readability of the GO enrichments, we identified clusters of GO terms given to similar genes using a graph clustering approach to reduce term redundancy. Starting from a binary matrix with genes as columns and significant (5% FDR) GO terms as rows, we built a nearest neighbor graph of GO terms using the Yule metric. We performed graph clustering on this NN graph and chose the GO term with the smallest p-Value as the cluster representative for each cluster.

TF motifs identification

We employed HOMER[28] 4.11 to identify TF motifs within the HOMER database designed for the hg38 genome. For each of the 51 RNAPII clusters, we conducted a systematic search for known motifs (default parameters), leveraging the extensive collection of motifs available within the HOMER database. The HOMER database encompasses a diverse range of experimentally validated motifs, which have been curated and annotated to provide reliable and accurate motif predictions.

GWAS traits and summary statistics

We used Stratified LD-Score Regression[66] (S-LDSR) to compute enrichments of heritability phenotypes for subgroups of consensus peaks. We downloaded all available GWAS summary statistics files from UK Biobank[67,68] (http://www.nealelab.is/uk-biobank), and only kept traits with strong heritability (noted "z7", as recommended by the documentation for this kind of analysis). RNAPII consensus coordinates were lifted to hg19 (UCSC liftover) which caused 873 RNAPII consensus to be removed. An SNP was assigned to a RNAPII consensus if it overlaps any part of the RNAPII consensus. The LDSC pipeline was run at default settings and Bonferroni correction was applied on the obtained p values. Heritability enrichments are defined as: $\frac{Proportion\ of\ h^2}{Proportion\ of\ intersected\ SNPs}$, where h^2 is the SNP-based heritability (see LD-Score paper[66]).

RNA-seq expression quantification of RNAPII-bound regions

We quantified RNA-seq expression on RNAPII-bound regions using featureCounts[69] similar to RNA-seq gene quantification. Instead of genes as sampling points, RNAPII-bound regions were used and standardised to 1kb long, centered on the consensus centroids (±500bp). Multi-mapping reads were excluded. Given the similarities of our data with scRNA-seq datasets, we employed several methods commonly used in the scRNA-seq field. Our data has a large number of samples (equivalent to cells) with much lower read counts compared to traditional gene-centric RNA-seq experiments. Before conducting each analysis, we preprocessed the data by filtering out RNAP2-bound regions that did not have at least one read count in three samples. This filtering step helped to remove noise and increase the quality of the data. The "gene-centric" GTEx count table was retrieved from GTEx.[15]

Count normalisation and transformation

An overview of our count processing is available in Figure S14. Counts were normalised using the scran pooling and deconvolution[70] approaches, as RNAPII-bound region counts have a large fraction of zeroes causing issues on approaches such as DESeq2's median of ratios. A small modification of the method was used to compute the size factors only using the top 5% "most detectable" bound regions. We defined detectability as the number of samples that have at least one read at an RNAPII-bound region. To break ties, we computed detectability at 2 reads, 3 reads … up to 5 reads, which is sufficient to break most ties. This helped to reduce the number of non-expressed RNAPII-bound regions or RNAPII-bound regions that are expressed in a single condition only, which should not be considered for an optimal normalisation. This can be seen as something analogous to the use of the geometric mean in DESeq2's median of ratios, which considers only genes with at least 1 read in each sample. RNAPII counts were transformed using the Pearson residuals of a regularised Negative Binomial (NB) model, similarly to what is used in SCTransform.[71] This kind of transform has been shown to better discriminate between biological conditions in scRNA-seq experiments, as well as reducing batch effects caused by differences in sequencing depths at small counts. The model for a gene/RNAPII-bound region expression is: $ln(\mu) = \beta X + ln(s)$. Where μ are the predicted means for a bound region/gene for each sample, β are the fitted model coefficients, X is the design matrix, and s are the count normalisation factors for each sample. In this work we use a simple intercept but the model allows more complex experimental designs. We use the following NB(μ, α) variance formulation: $V(\mu) = \mu + \alpha\,\mu^2$. We fitted a trendline of the overdispersion parameter α as a function of the mean, so we obtain a regularised estimate of the variance that only depends on the mean. To do so, we binned genes/RNAPII-bound regions into 20 groups according to their quantile of mean expression, evaluated each gene/bound region overdispersion parameter, find the modal value of overdispersion within a group using a kernel

density estimate (with Silverman's rule to estimate bandwidth), and linearly interpolate results between each group/quantile of mean expression. We only used up to 5000 RNAPII-bound regions/genes to fit the mean/overdispersion relationship to speed up computations. The python statsmodels library was used to fit the NB models with the more robust Nelder-Mead solver instead of BFGS. The pearson residuals are then computed as following: $= \frac{x-\mu}{\sqrt{V(\mu)}}$, where x is the count value. A custom python implementation was employed as the SCTransform package failed to run on the RNAPII count matrices, possibly due to much larger counts than UMI scRNA-seq experiments, causing numerical instability when fitting the models. The original implementation clips the pearson residuals at $\pm \sqrt{n/4}$ by default, where n is the number of cells/samples, in order to reduce the influence of outliers. We found these bounds to be quite small when dealing with smaller sample sizes, which can remove biological signals. Instead, we clipped values at $\pm \sqrt{9+n/4}$, creating larger bounds for small sample sizes without changing the large sample size behavior.

Reads distribution on standardised RNAPII consensus

To visualise the read profiles, we employed the following methodology. Initially, we sampled the RNA-seq signal within 10-base pair (bp) windows, which were positioned within the standardised 1-kilobase (kb) RNAPII consensus regions. To integrate the data from multiple samples, we generated a pooled "meta-sample" by summing the number of reads of each sample. Furthermore, we normalised the read counts for each sample by dividing them by the total number of reads, ensuring accurate comparisons across samples. To standardise the windowed signal of each RNAPII consensus region, we normalised it by dividing by the maximum pooled signal across all 10bp windows within that specific RNAPII consensus region. This normalisation step allowed us to eliminate any potential bias and enabled fair comparisons between different regions and datasets. Next, we applied ward hierarchical clustering to arrange the rows of the heatmap for each dataset. This clustering approach facilitated the identification of similar transcriptional patterns. Finally, to generate an overall profile for each dataset, we calculated the mean value of the pooled, maximum normalised transcriptional signal across all RNAPII consensus regions.

Unsupervised feature selection, dimensionality reduction and predictive models

Feature selection in scRNA-seq is a common step that allows to remove a large fraction of potentially uninformative bound regions/genes (i.e., those with very low expression or those with ubiquitous expression, which are not informative of the sample/cell biology). Typically, around 2000 to 3000 genes are kept in scRNA-seq experiments, but this number is generally tuned for each experiment. To automatically select ''highly variable'' features for each dataset, we computed the sum of the squared pearson residuals, which are asymptotically following a χ^2 distribution with n - p degrees of freedom, n being the number of samples, and p the number of parameters of the model (1 in our case). We performed an upper tail test for each gene/bound region and kept bound regions at an FDR of 5%. This selects only sufficiently expressed genes above the mean-variance trendline, and due to the clipping of the pearson residuals also removes outliers with an extreme variance (Figure S22A). We performed PCA on the Pearson Residuals of these highly variable features. To automatically identify the optimal number of Principal Components, we used Horn's Permutation Parallel Analysis, which has been found to be one of the most effective approaches to identify the number of components in factor analysis[72] (cit). This approach generates row–wise permutations for each feature, computes PCA on these permuted datasets, then the selected number of components is the threshold at which the eigenvalues from the randomised dataset are larger than the real dataset. We performed 3 permutations due to the computational cost of this approach, which is acceptable as the randomised eigenvalues are very stable on large matrices (Figure S22B). We used the fast ''randomised'' solver from the python sklearn[59] library to compute PCAs.

For UMAP visualisation, we used 30 neighbors, a min_dist parameter of 0.5, Pearson correlation as the metric and use data in PCA space as input. For heatmaps, we used a similar approach as the RNAPII heatmap, except that the data was used in PCA space as input to the UMAP pass for the samples, and used Pearson correlation as the metric for both samples and RNAPII-bound regions. The predictive model uses a Catboost gradient boosted decision tree model that takes as input the data in PCA space. Default settings were used with the exception of balanced class weights (where each sample is reweighted by class proportion). We used balanced accuracy (where each sample is reweighted by class proportion) as the main metric to evaluate the model over a stratified 10–Fold Cross Validation.

Identification of per tissue markers and "meta-clustering"

For each dataset (TCGA, ENCODE, GTEx, RNAPII), we identified markers for each annotation (i.e., Pol2+Liver, GTEx+Blood). To identify markers in the three RNA-seq datasets, we performed a group-versus-rest, one sided t-test on the Pearson Residuals. We kept over-expressed bound regions with log2 Fold Change above 0.25, and detectable in at least 10% or 2+ samples (whichever is the largest). For the RNAPII dataset, we performed an hypergeometric test for each RNAPII consensus, where: N, the population size, is the number of peaks across all experiments; K, the number of successes in the population, is the number of peaks across all experiments with the annotation of interest; n, the number of draws, is the number of experiments that has a peak at the studied consensus peak; k, the number of observed successes, is the number of experiments with the annotation of interest that has a peak at the studied consensus peak. We used a BH FDR cutoff of 5% in both cases. This yields a binary vector which indicates whether a RNAPII-bound region is a marker or not for each dataset+annotation. We removed RNAPII-bound regions which are markers in more than 10% of the dataset+annotation combinations or in less than two dataset+annotation. An Average Linkage clustering using the Yule binary metric was performed, which we found to be less sensitive to the number of identified markers.

Tissue specific eQTL enrichments

We downloaded per tissue eQTL data from GTEx "GTEx_Analysis_v8_eQTL.tar" and used the list of significant eQTL-gene pairs for each tissue. For each SNP listed as an eQTL we stored whether it is listed as an eQTL or not in each tissue. SNPs listed as eQTLs in more than 10% of the tissues (6 or more) were removed to keep variants that are likely located in tissue specific regulatory regions. Using the list of per dataset, per tissue marker RNAPII-bound regions, we kept RNAPII-bound regions that are markers in less than 10% of the tissues in a dataset (6 or less) and removed non-marker RNAPII-bound regions. We removed tissues having less than 50 markers left after this step. We computed pairwise intersection enrichment p values between the tissue-specific eQTLs and the marker RNAPII-bound regions (SNP intersection against whole bound region). We computed an hypergeometric enrichment p value for each of these intersections as following: N, the population size, is the number of RNAPII-bound regions (after filtering); K, the number of successes in the population, is the number of RNAPII-bound regions intersecting an eQTL from the eQTL-wise tissue of interest; n, the number of draws, is the number of marker RNAPII-bound regions of the second tissue of interest; k, the number of observed successes, is the number of marker RNAPII-bound regions of the second tissue of interest intersecting an eQTL from the eQTL-wise tissue of interest.

Differential expression

To identify differentially expressed (DE) RNAPII-bound regions between tumor and normal tissues, we performed a t-test on the Pearson Residuals with an FDR cutoff of 5%. We constrained DE RNAPII-bound regions to have an absolute log2 Fold Change above 0.25, and to be detectable in at least 10% or 2+ samples (whichever is the largest) of either class (normal/tumor). For the detection of Tumor Subtype specific markers, we compared the expression of samples of a subtype to reference normal samples. We used the same significance cutoffs. We considered a marker to be subtype-specific only if it appeared for this subtype. Linear modeling methods such as DESeq2 ran out of memory on large datasets and required an excessive computation time. A t-test was used to accommodate large datasets (100+ samples x 180 000 bound regions in most datasets) and to keep an uniform processing for each dataset for our cross-dataset analyses. With sufficiently large sample sizes, the t-test yields robust markers, although with less statistical power (Figures S22C and S22D). To evaluate our approach on a much smaller dataset with less statistical power, we selected samples from two similar types of heart tissues from GTEx and downsampled to obtain an n = 3 comparison. Here, we used DESeq2[73] to maximise statistical power. We performed 100 random sampling iterations to obtain 3 samples for each tissue, evaluated DE in each iteration, then kept bound regions supported as DE in at least half of the downsampling iterations. To evaluate the relationship between sample size and statistical power, we performed 10 downsampling iterations for each sample size. GREAT[65] v4.0.4 analyses were performed on BRCA TNBC specific RNAPII-bound regions, identifying enrichment of the "MAP kinase tyrosine/serine/threonine phosphatase activity" GO:0017017 term (Binom FDR Q-Val 1.34e-14), with TNBC specific RNAPII-bound regions associated with 10 DUSPs genes.

Using GTEx normal tissues instead of TCGA normal tissue

We evaluated the variation of cancer-specific intergenic RNAPII markers when using GTEx normal tissues instead of TCGA normal tissues. We selected tumor samples from TCGA and utilised GTEx normal samples as the reference group (Figure S23). Differential expression analysis was performed between these two groups, employing the same methodology as in the primary TCGA cancer vs. TCGA normal analysis. The RNAPII consensus regions were then ranked based on their p values, and an equal number of markers were retained as in the main analysis. To evaluate the overlap between the TCGA normal vs. GTEx normal analyses, we computed two statistics: A) the recall, which represents the fraction of shared markers, and B) the fold change of the observed number of shared markers compared to the expected value if they were chosen randomly. It is important to note that the original TCGA normal vs. TCGA tumor samples were not paired in the analysis (there are, in fact, more cancer samples than normal ones). Additionally, GTEx and TCGA employ different protocols for tissue sampling, conservation, and sequencing, which introduces a noticeable batch effect. Our findings indicate a substantial and statistically significant overlap between the two analyses, with the exception of ESCA (Esophageal carcinoma) cancer, which also exhibited lower classification accuracy in machine learning.

Identification of pan-cancer markers

To identify RNAPII-bound regions whose expression is associated with survival or DE (separately) in tumor tissues in multiple cancers ("pan-cancer markers"), we randomly selected for each cancer the same number of marker bound regions as observed in this cancer. This process was repeated 100 times to obtain the expected distribution of the number of cancers in which a bound region is a marker. We identified the "pan-cancer threshold" as the threshold where less than 5% of the observed markers are expected to belong to the null distribution (equivalent to 5% FDR, see Figure S24). This approach allowed us to set a statistically meaningful threshold to identify bound regions that are markers in more cancers than expected at random, instead of an arbitrary threshold.

Survival analysis

For survival analysis, we fit a linear Cox Proportional Hazards regression model on the Pearson residuals using the Python Lifelines library, and use a 5% FDR threshold. Kaplan-meier survival curves were created using the kaplanmeier python library. The maxstat R library was used to obtain the optimal expression cutpoint as well as the associated maximally selected log -rank statistic p value (using the most accurate "condMC"′ method with 100 000 samples to compute the p value).

Enrichment of cancer hallmark-related genes

To investigate the enrichment of cancer hallmark-related genes, we utilised a gene set enrichment analysis (GSEA) approach. The list of candidate genes associated with specific cancer hallmarks was obtained from the CHG database (http://www.bio-bigdata.com/CHG/) which provides a comprehensive collection of genes with verified and putative links in various cancer-associated biological processes. We conducted GSEA analysis on RNAPII markers genomic regions. However, instead of utilising Gene Ontology (GO) terms, we employed the cancer hallmarks as defined in the CHG database. Cancer hallmarks encompass key biological processes and pathways that contribute to tumorigenesis and cancer progression. We assessed the enrichment of cancer hallmark-related genes nearby the pan-cancer RNAPII markers. This analysis allowed us to identify potential associations between the RNAPII marker and specific cancer hallmarks, providing insights into the functional relevance of these regions in cancer biology.

Cell Genomics, Volume *3*

Supplemental information

Characterizing intergenic transcription at RNA polymerase II binding sites in normal and cancer tissues

Pierre de Langen, Fayrouz Hammal, Elise Guéret, Jean-Christophe Mouren, Lionel Spinelli, and Benoit Ballester

SUPPLEMENTAL INFORMATION

Characterising intergenic transcription at RNA polymerase II binding sites in normal and cancer tissues

Figure S1 : Characteristics of the RNAPII atlas, related to STAR Methods, related to STAR Methods 2
Figure S2 : Examples of intergenic RNAPII occupancy in MCF-7 and HEK293 cells, related to Figure 1 3
Figure S3 : Large scale integration methodology of RNAPII ChIP-seq, related to STAR Methods 4
Figure S4 : Intergenic RNAPII bound regions are co-localizing with regulatory elements, related to Figure 1 5
Figure S5 : Intergenic RNAPII bound regions are enriched in end of gene regulatory elements, related to STAR Methods 6
Figure S6 : Comparison of intergenic RNAPII consensus with genomic resources, related to Figure 1 7
Figure S7 : Genome-wide large scale integration of 890 Human H3K27Ac Histone ChIP-seq experiments, related to STAR Methods 8
Figure S8 : RNAPII occupancy clustering and cluster annotation statistics, related to STAR Methods 9
Figure S9 : Tissue-specific biological characteristics of RNAPII consensus, related to Figure 3 10
Figure S10 : Tissue-specific epigenetic states of RNAPII consensus, related to Figure 3 11
Figure S11 : RNAPII-bound regions captures a majority of intergenic transcriptional signal, related to Figure 4 12
Figure S12 : Transcriptional profiles of intergenic RNAPII-bound regions, related to STAR Methods 13
Figure S13 : Flowchart of the RNA-seq pipeline, related to STAR Methods 14
Figure S14 : Intergenic transcription by itself is sufficient to characterise biological conditions, related to STAR Methods 15
Figure S15 : Tissue-specific regulatory variants are enriched within tissue-specific Intergenic transcripts, related to STAR Methods 16
Figure S16 : The intergenic transcriptional signal is not driven by end-of-gene transcription, related to STAR Methods 17
Figure S17 : Differentially expressed RNAPII-bound regions can be detected at smaller sample sizes, related to STAR Methods 18
Figure S19 : Non coding transcription captured at RNAPII-bound regions discriminates normal and tumour tissues, related to Figure 6 20
Figure S20: Non coding transcription captured at RNAPII-bound regions is prognostic of the patient's survival, related to Figure 7 21
Figure S21: Supplementary multi-cancer prognostic marker, related to Figure 7 22
Figure S22 : Mean-variance trendline, Feature selection and PCA Permutation Parallel analysis, related to STAR Methods 23
Figure S23 : Using GTEx normal tissues instead of TCGA normal tissue, related to STAR Methods 24
Figure S24 : Distribution of differentially expressed RNAPII-bound regions in cancers, related to STAR Methods 25

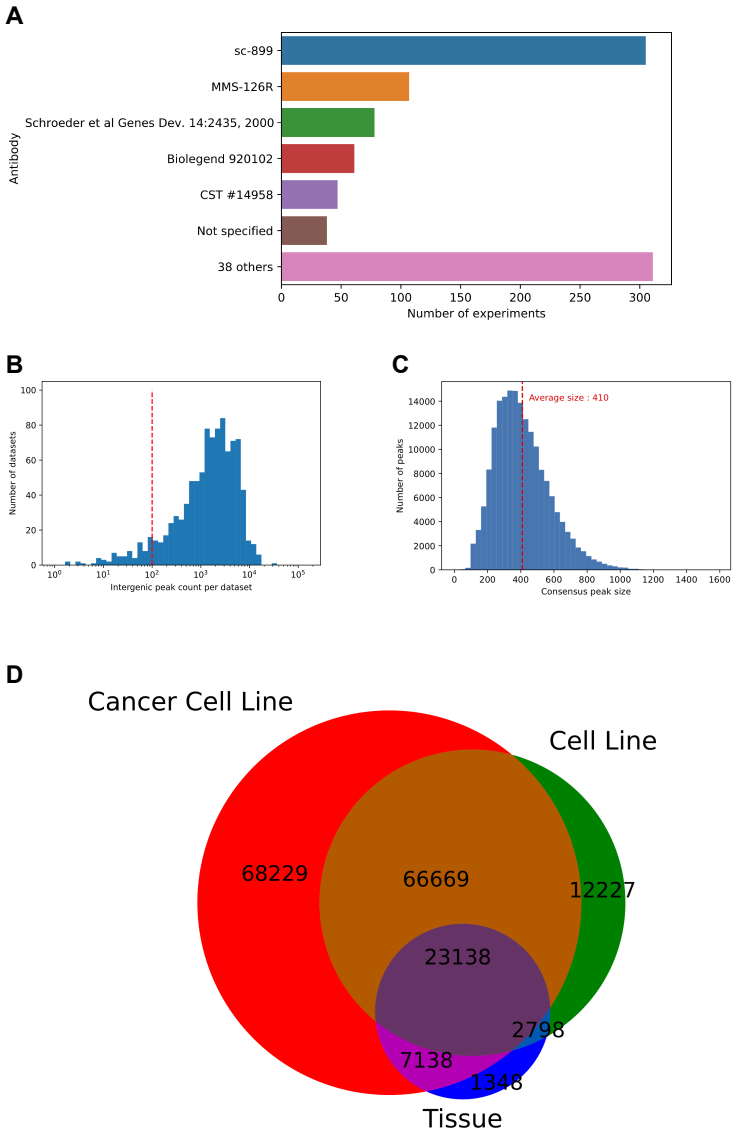


Figure S1 : Characteristics of the RNAPII atlas, related to STAR Methods, related to STAR Methods

A. Number of RNAPII ChIP-seq experiments in which each antibody is used. **B.** Histogram of the distribution of the number of Intergenic RNAPII peaks per dataset, average 2,787 intergenic binding events per dataset. Red line indicates minimal cutoff for a dataset to be retained. **C.** Histogram of the distribution of the intergenic consensus peaks sizes, with an average consensus width of 410bp (red line). **D.** Detectability of RNAPII consensus peaks per sample category ; Cancer cell lines, Cell lines and Tissues.

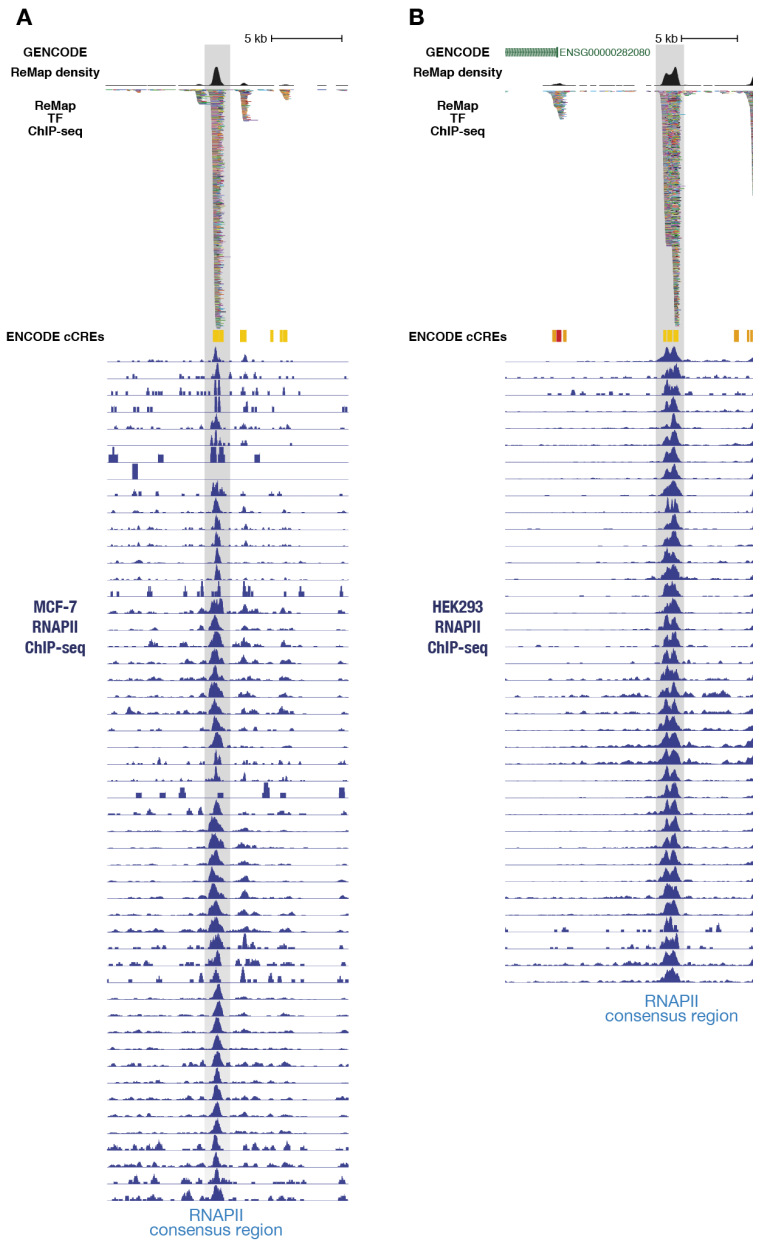


Figure S2 : Examples of intergenic RNAPII occupancy in MCF-7 and HEK293 cells, related to Figure 1

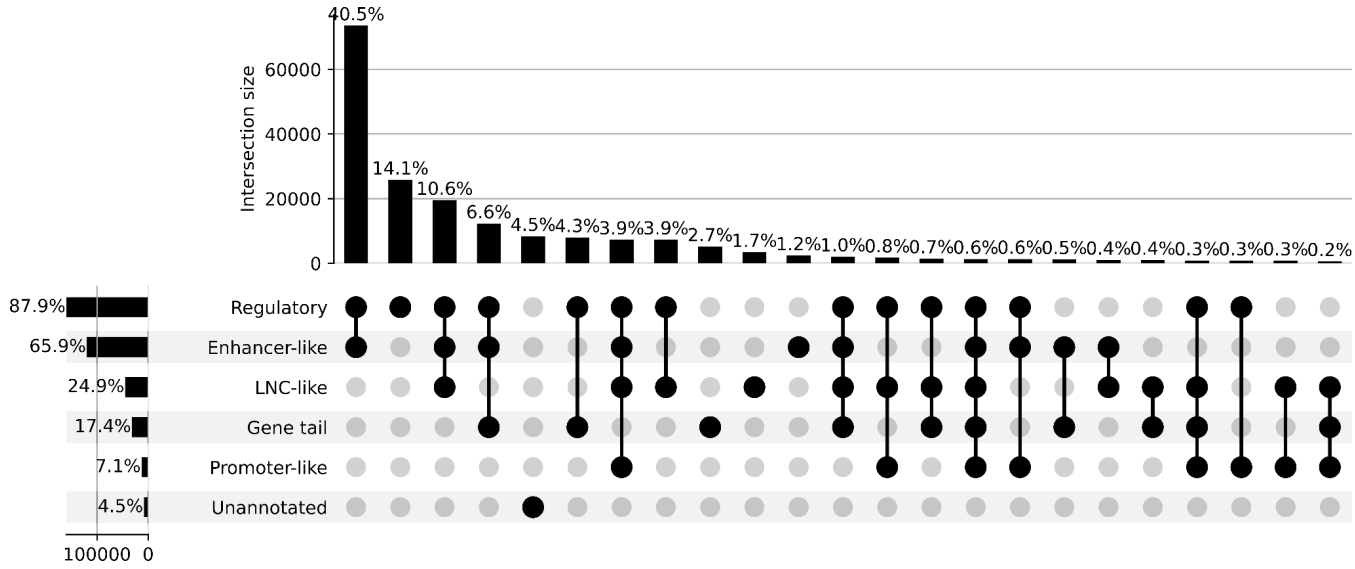
A. Genomic example on chr1:31170514-31171424, displaying RNAPII raw ChIP-seq signals across MCF-7 cell lines (Blue) at the location of a RNAPII consensus region (grey bar), with GENCODE. ReMap TF ChIP-seq and ENCODE cCREs tracks. **B.** Similarly, a second example of RNAPII ChIP-seq signals at chr12:8023634-8024204 with HEK293 cell lines.



Figure S3 : Large scale integration methodology of RNAPII ChIP-seq, related to STAR Methods

Schematic of our post peak-calling integration methodology : **A.** Identification of consensus peaks via a peak density-based approach. Naive approach refers to a simple merge on overlap. **B.** Summarization of all datasets / consensus peaks in a binary matrix storing the presence or absence of RNAPII at each consensus peak in each experiment.

A



B

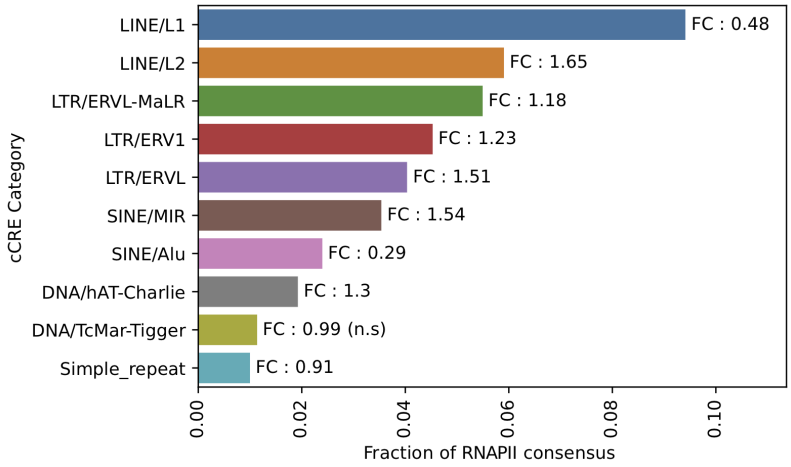


Figure S4 : Intergenic RNAPII bound regions are co-localizing with regulatory elements, related to Figure 1

A. UpSet plot of putative functional annotations of intergenic RNAPII consensus, derived from reference databases (see methods for details). Rare combinations with intersection sizes smaller than 0.1% of all intergenic RNAPII consensus have been removed for clarity. **B.** Fraction of consensus peaks intersected for the top ten most intersected repeat families. FC corresponds to fold change enrichment versus random regions. All Fold Changes are statistically significant (FDR < 0.05) unless otherwise mentioned (hypergeometric test, see methods, p < 1e-100).

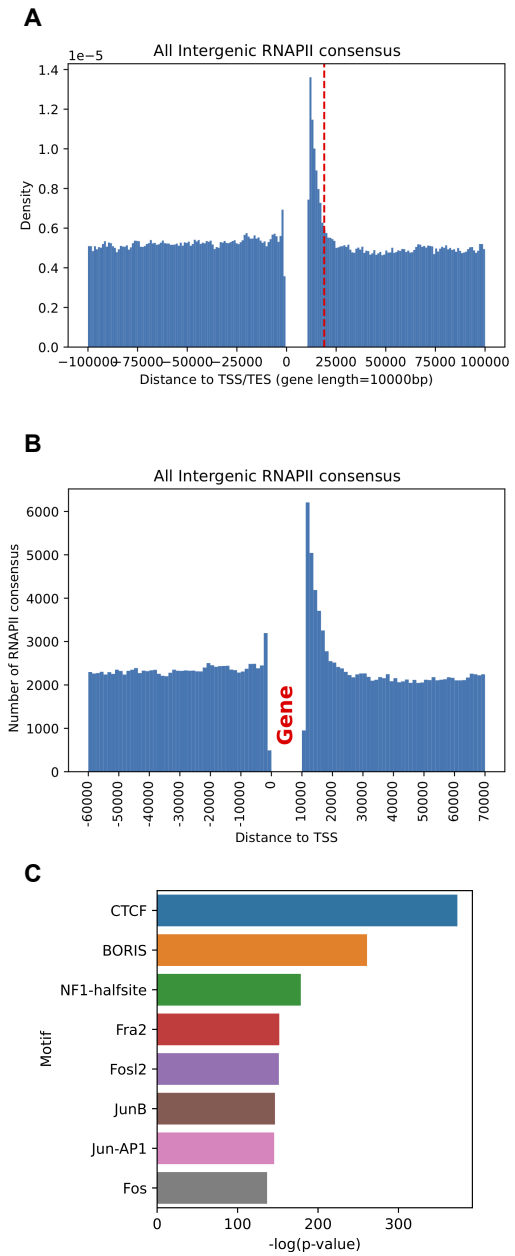


Figure S5 : Intergenic RNAPII bound regions are enriched in end of gene regulatory elements, related to STAR Methods

A. Distributions of RNAPII consensus centroids relative to protein coding genes (gene length standardised to 10kb), red line indicates 9kb after Transcription End Site (TES). **B**. Distribution of RNAPII consensus counts by 5kb windows relative to TSS of and TES of protein coding genes (red), in narrower genomic coordinates. **C.** HOMER top 10 known motif enrichment -log10(p-values) for RNAPII consensus within 9kb of a TES. The top 2 TFs are CTCF and CTCFL (BORIS).

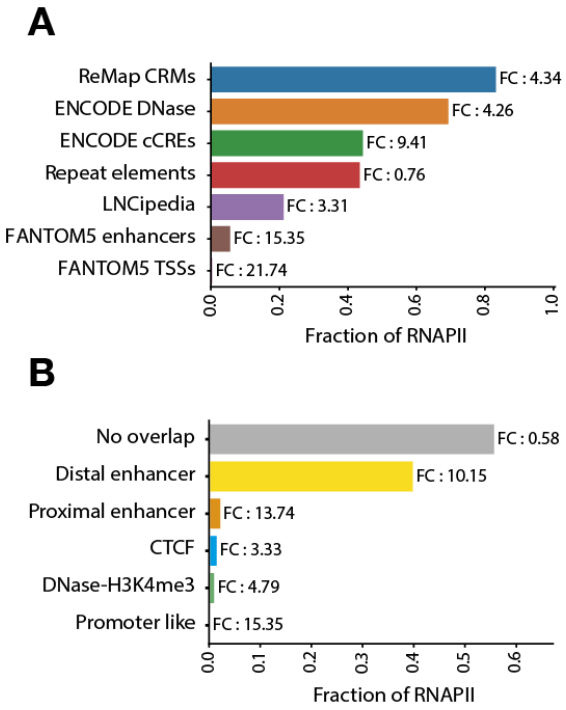


Figure S6 : Comparison of intergenic RNAPII consensus with genomic resources, related to Figure 1

A. Comparison of RNAPII consensus location with major genomic resources of regulatory and non-coding elements. **B.** Distribution of candidate cis-Regulatory Elements (cCREs) derived from ENCODE compared to the RNAPII atlas. FC denotes fold change versus random intergenic regions; all results are strongly statistically significant (hypergeometric test, see methods, $p < 1e\text{-}300$).

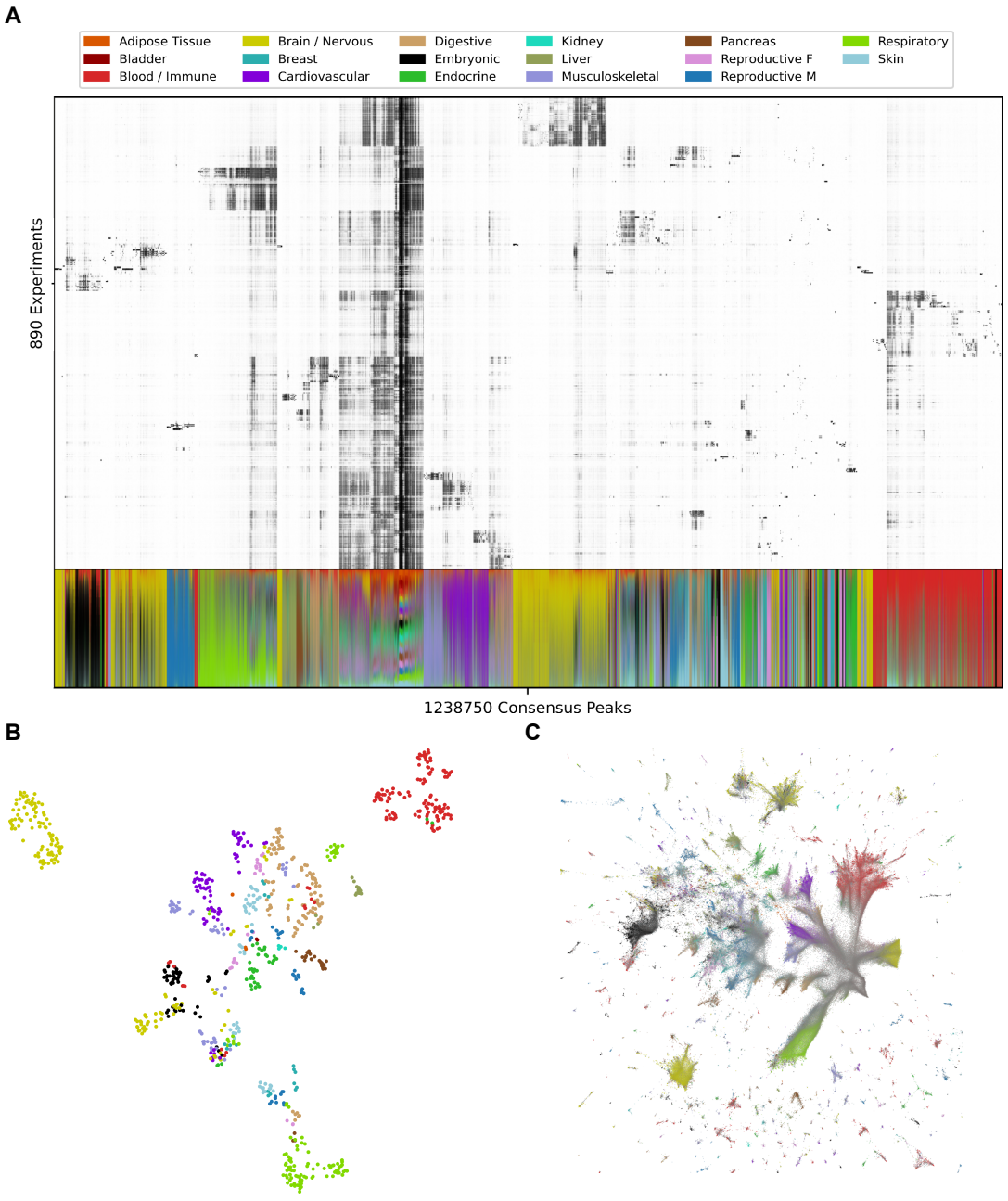


Figure S7 : Genome-wide large scale integration of 890 Human H3K27Ac Histone ChIP-seq experiments, related to STAR Methods

A. H3K27Ac occupancy in 1,238,750 consensus regions across 890 biosamples. Lower panel indicates the normalised contribution of a biotype, in terms of peaks, to each consensus. **B.** Two-dimensional Uniform Manifold Approximation and Projection (UMAP) projection of all 890 H3K27Ac ChIP-seq datasets. **C.** UMAP projection of all H3K27Ac consensus according to their binding patterns, coloured by dominant biotype.

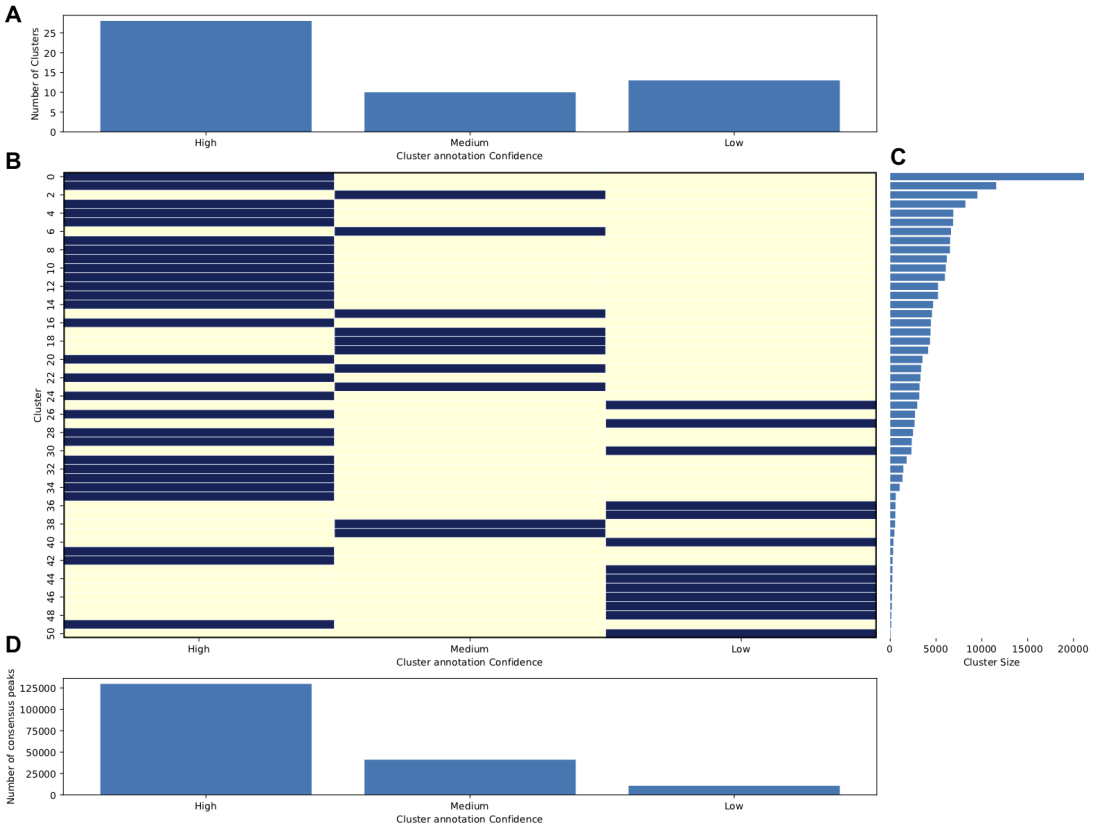


Figure S8 : RNAPII occupancy clustering and cluster annotation statistics, related to STAR Methods

A. Number of clusters by level of annotation confidence (High, Medium, Low), derived with concordance with biological enrichments. **B.** Annotation confidence by cluster (dark = selected annotation). **C.** Number of RNAPII consensus peaks within one of the 51 clusters of annotation **D.** Number of RNAPII consensus peaks across the three levels of annotation confidence.

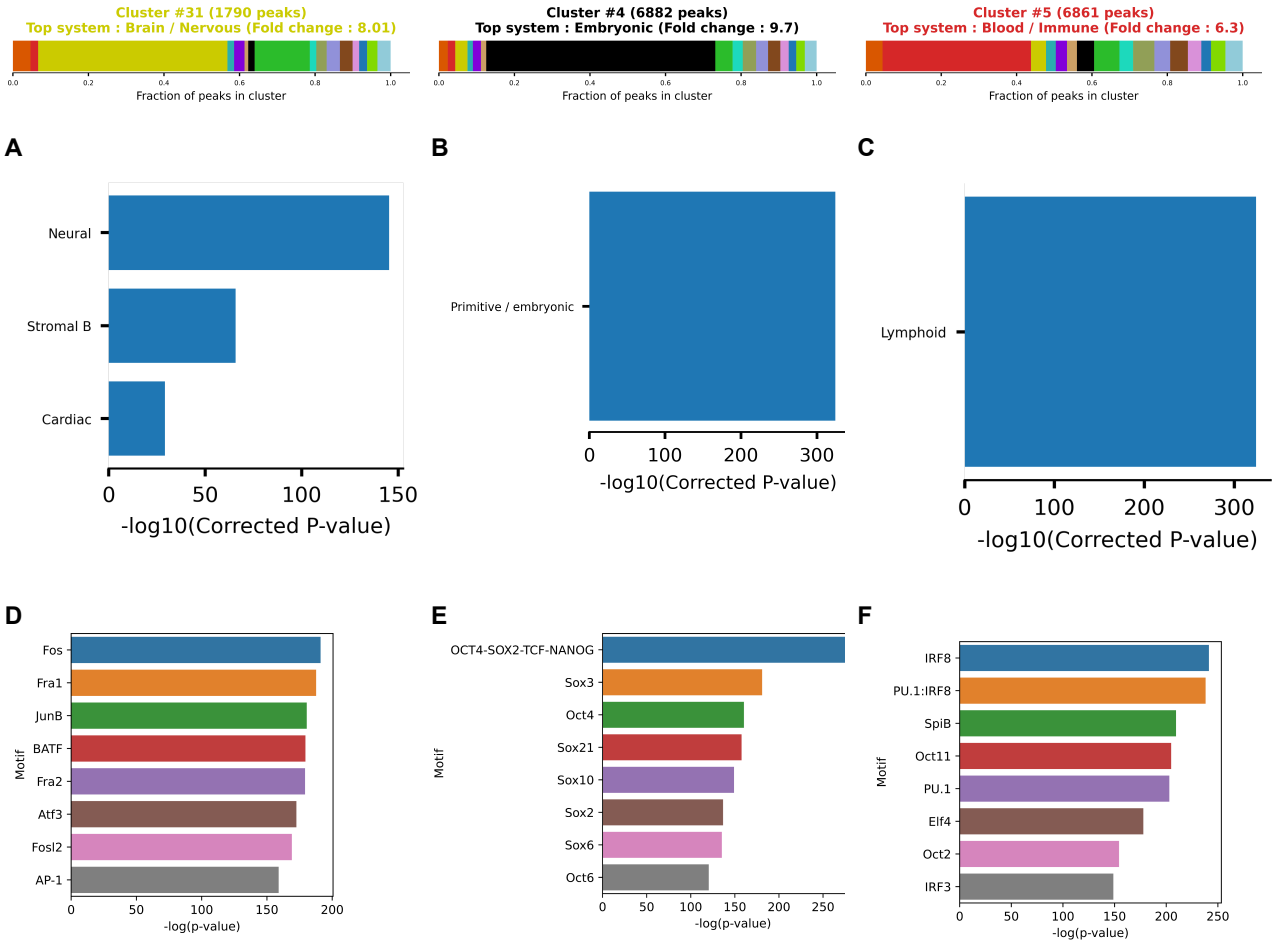


Figure S9 : Tissue-specific biological characteristics of RNAPII consensus, related to Figure 3

Detailed characteristics of three selected RNAPII clusters (#31, #4, #5) compared to the biological classification of the human index of DNase I hypersensitive sites (DHSs) (A,B,C) and HOMER top 8 known DNA motifs enrichments. **A.** RNAPII cluster #31 defined as Brain/Nervous biotype is enriched in Neural, Stromal B and Cardiac DHS. **B.** RNAPII cluster #4 defined as Embryonic biotype is enriched in Primitive/Embryonic DHSs. **C.** RNAPII cluster #5 defined as Blood/Immune biotpype is enriched in Lymphoid DHSs. **D.** The HOMER top 8 DNA motifs for the RNAPII cluster #31 defined as Brain/Nervous biotype contains Fox and Jun. **E.** The RNAPII cluster #4 defined as Embryonic biotype is enriched in Oct4, Sox2 and Nanog DNA motifs. **F.** The RNAPII cluster #5 defined as Blood/Immune biotpype is enriched in IRF8, Pu.1 DNA motifs.

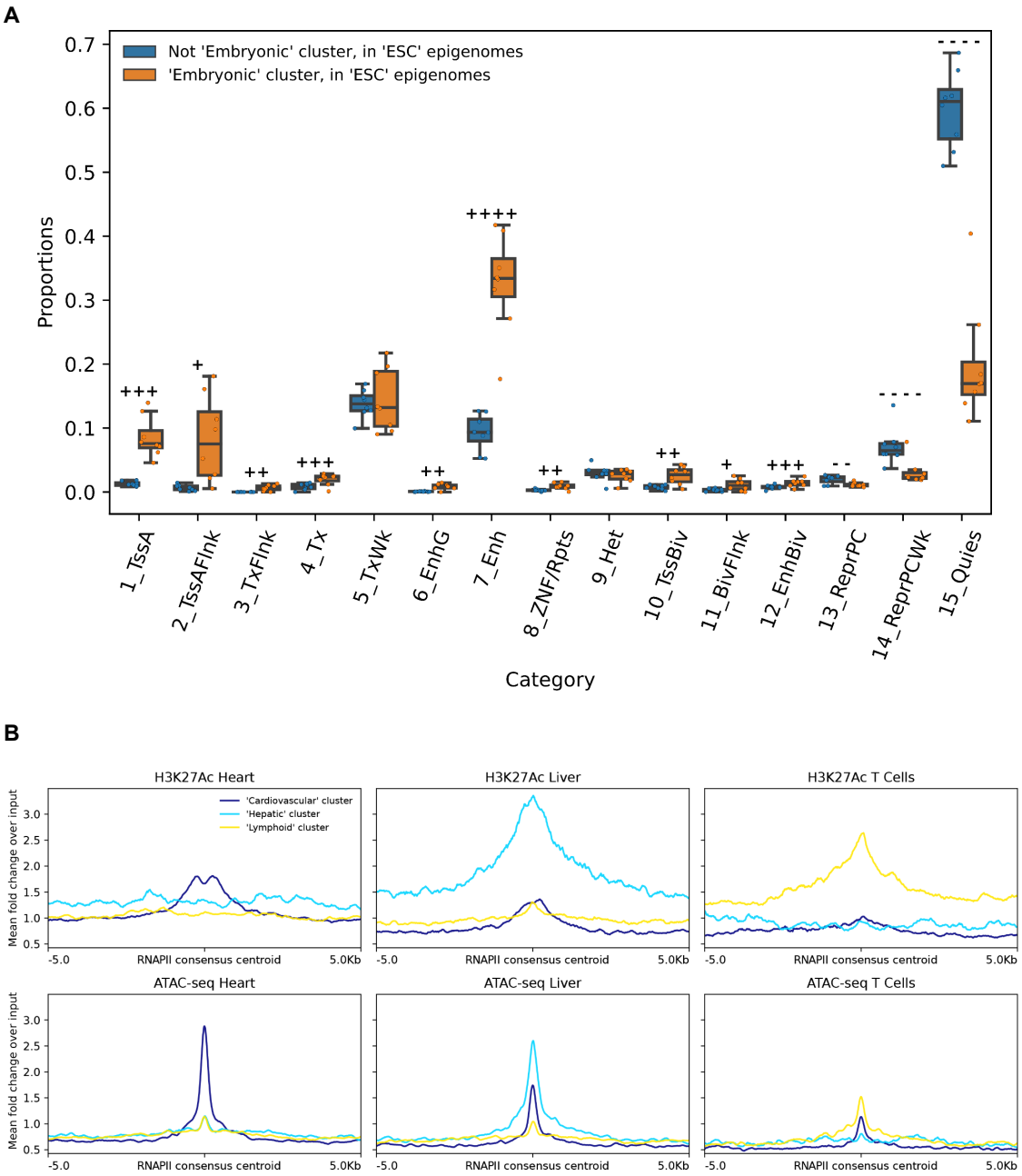


Figure S10 : Tissue-specific epigenetic states of RNAPII consensus, related to Figure 3

A. Proportions of ChromHMM epigenetic states in the “Embryonic” cluster against other Intergenic RNAPII consensus. (* : p < 0.05, ** : p < 0.01, *** : p < 0.001, **** : p < 0.0001; + or - indicate sign of mean difference; two-sided paired t-test). Blue boxes correspond to RNAPII peaks not included in the Embryonic cluster #4, but overlapping ESC epigenomes. Orange boxes correspond to RNAPII peaks forming the Embryonic cluster #4, and overlapping ESC epigenomes. **B.** H3K27Ac ChIP-seq and ATAC-seq profiles at clusters of RNAPII consensus. Selected clusters are respectively #5, #10 and #20 for “Cardiovascular”, “Hepatic” and “Lymphoid” (largest representative clusters, see supplementary data).

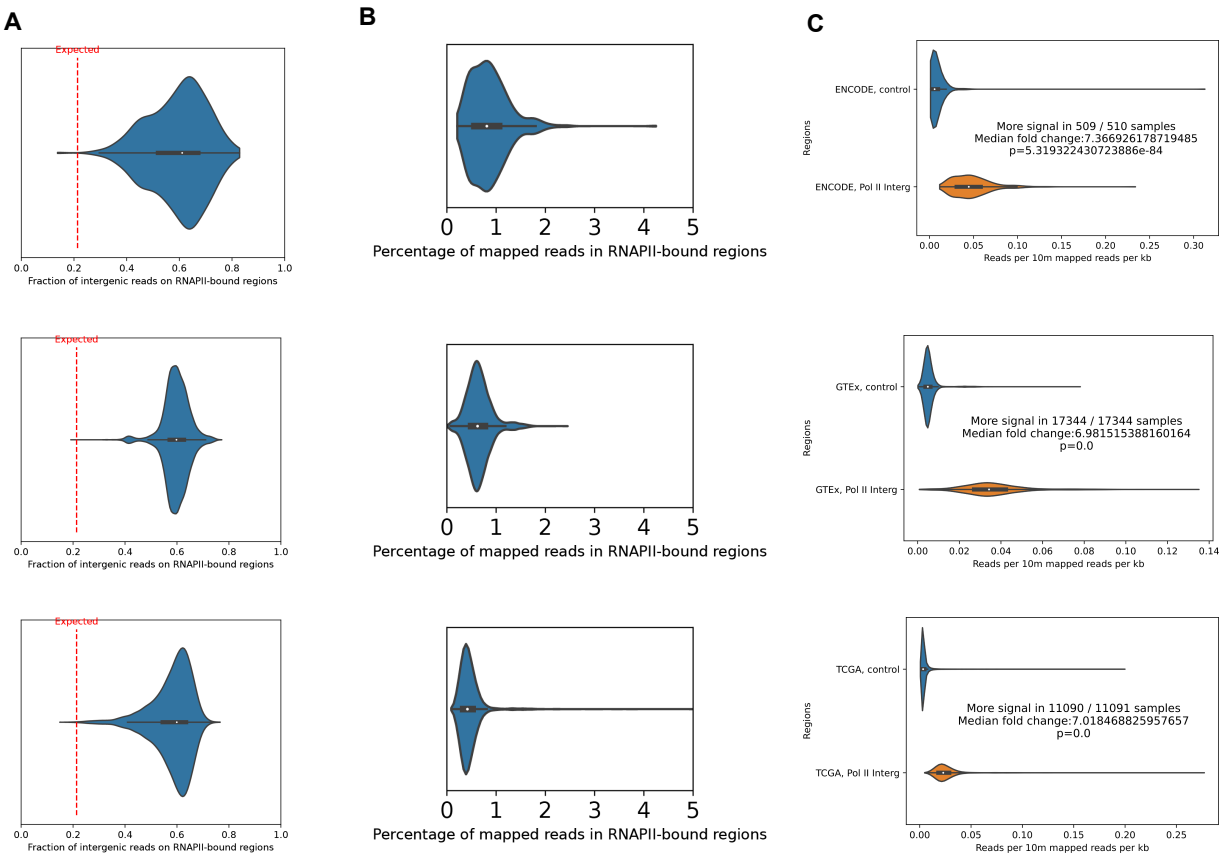


Figure S11 : RNAPII-bound regions captures a majority of intergenic transcriptional signal, related to Figure 4

A. Distribution as violin plots of the fraction of intergenic reads captured by RNAPII-bound regions across samples. Dashed line indicates expected value due to RNAPII-bound regions coverage. From top to bottom : ENCODE, GTEx, TCGA RNA-seq datasets. **B.** Distristribution of the percentage of mapped reads captured by RNAPII-bound regions. **C.** Distribution of the average number of reads per RNAPII-bound region (blue) or control regions (orange) (non RNAPII-bound region, all 1kb binned intergenic regions), across samples (p-values from Mann-Whitney U-test).

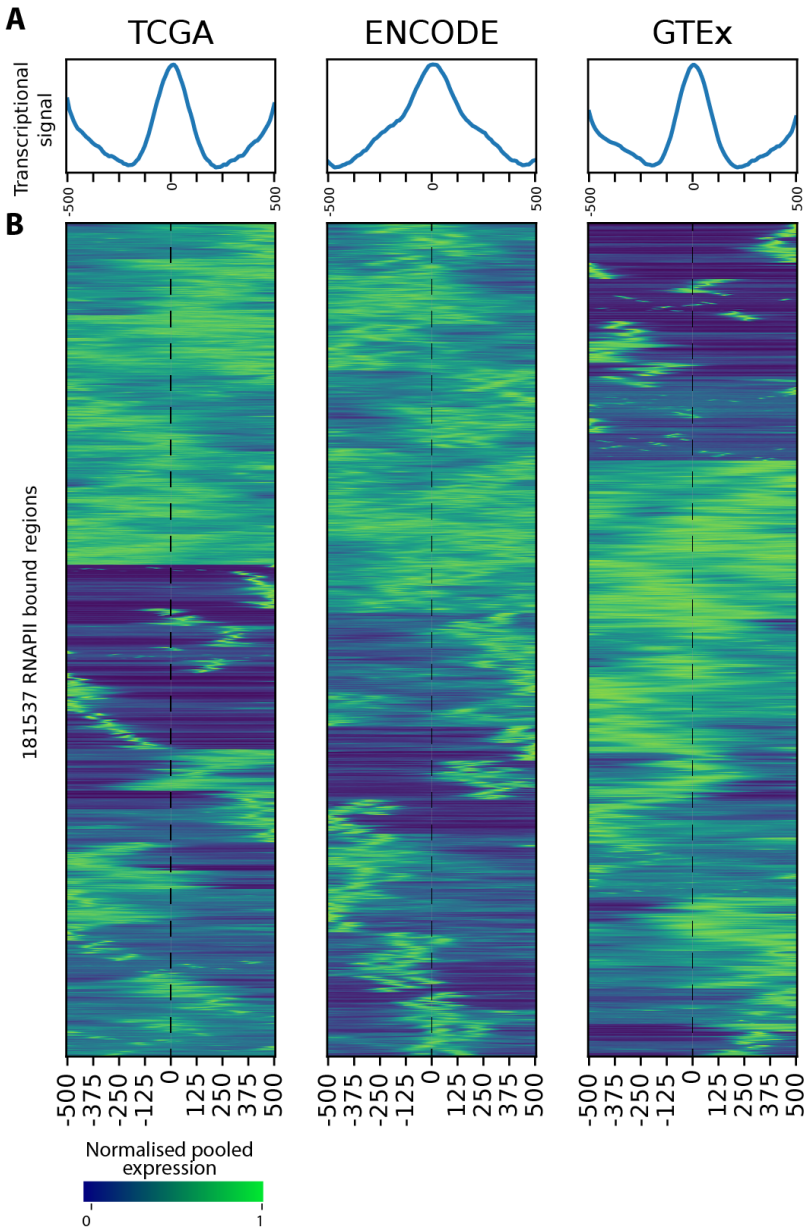


Figure S12 : Transcriptional profiles of intergenic RNAPII-bound regions, related to STAR Methods

A. Pooled, normalised, and averaged transcriptional profiles of intergenic RNAPII-bound regions (methods). **B.** Pooled, normalised, and clustered heatmaps depicting the transcriptional profiles of the 184,547 intergenic RNAPII-bound regions. Each heatmap clustering is performed independently for TCGA, ENCODE, and GTEx datasets.

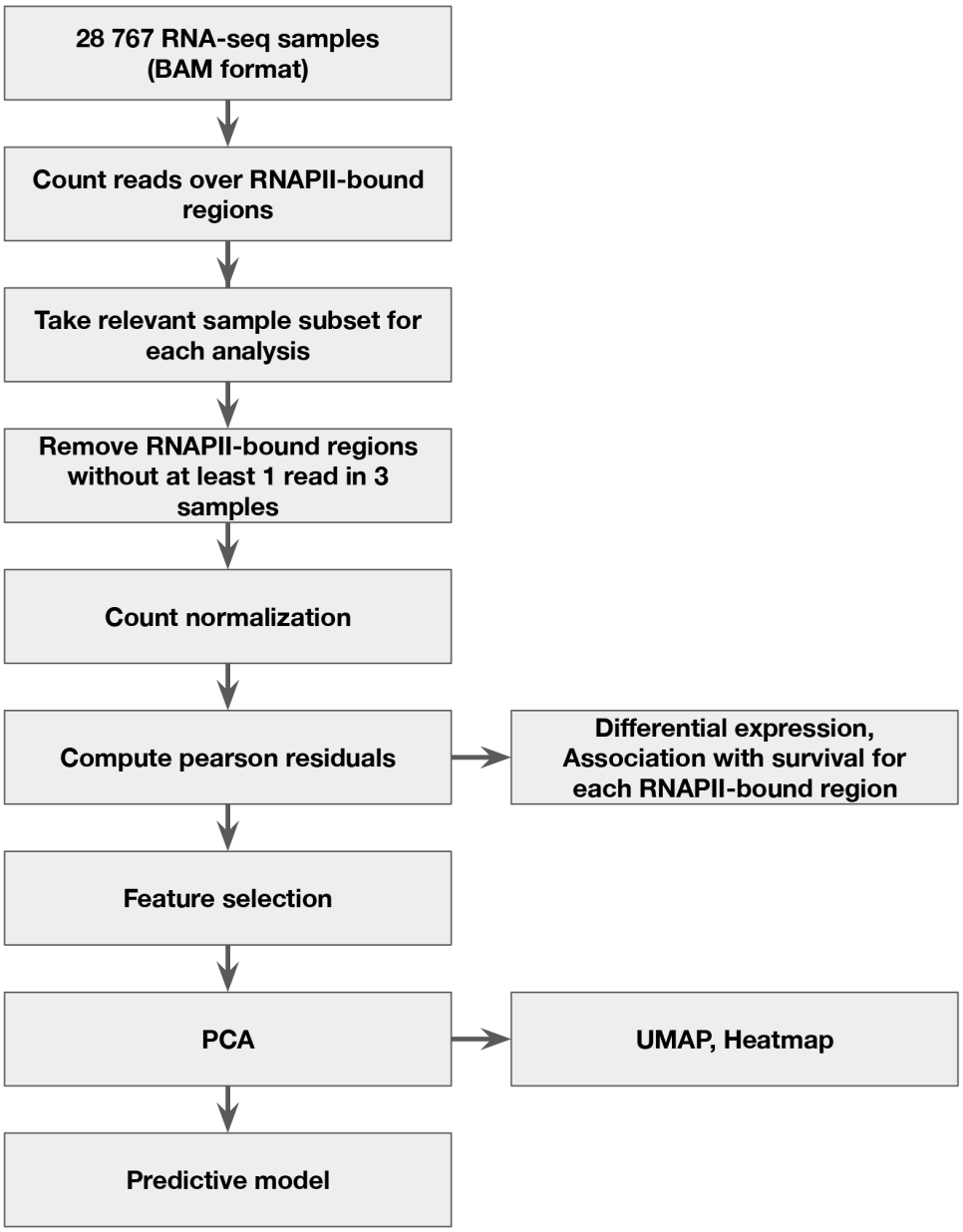


Figure S13 : Flowchart of the RNA-seq pipeline, related to STAR Methods

Simplified schematic of the RNA-seq processing pipeline. See methods for additional details.

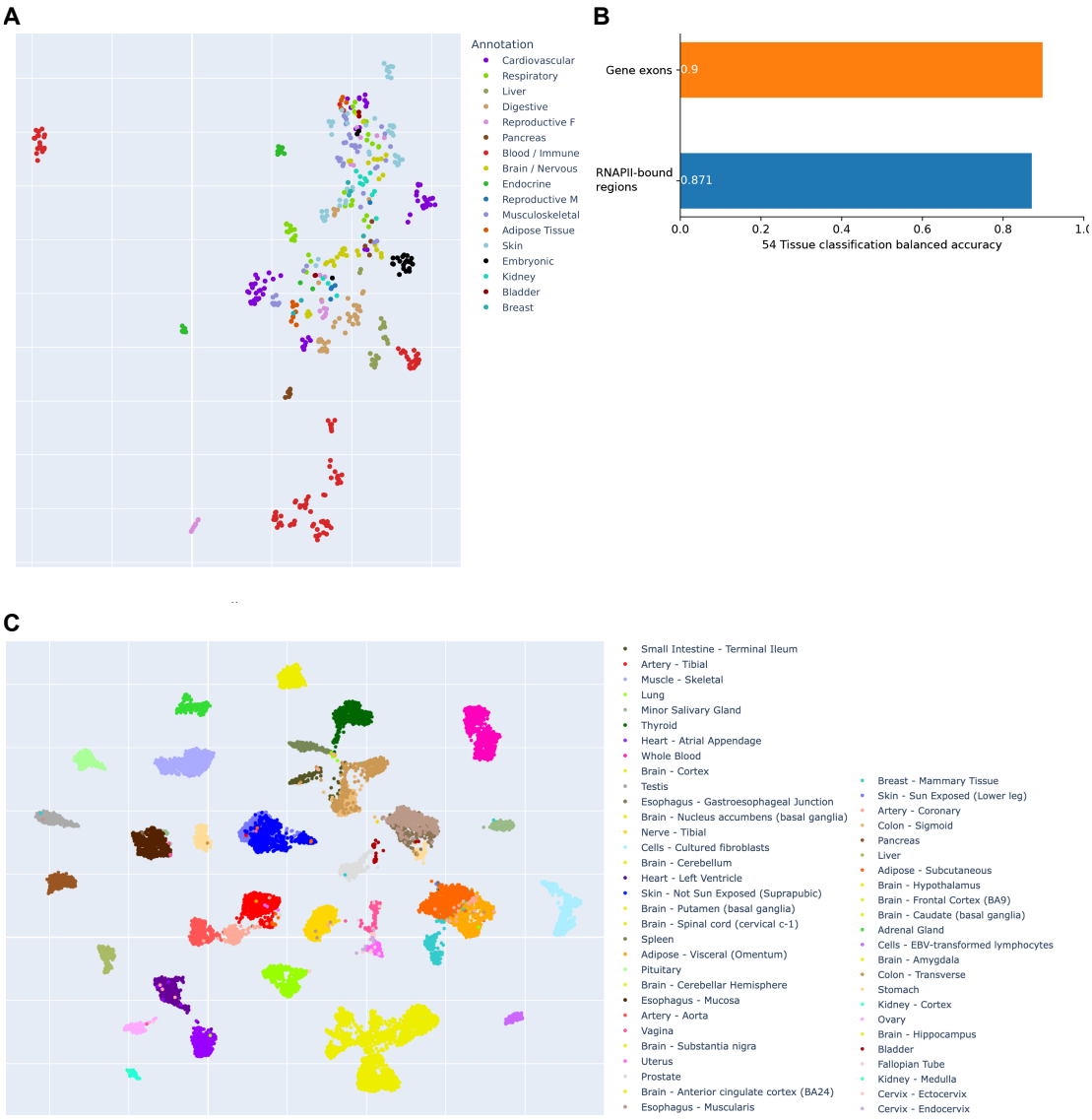


Figure S14 : Intergenic transcription by itself is sufficient to characterise biological conditions, related to STAR Methods

A. UMAP of ENCODE total RNA-seq samples using RNA-seq signal at RNAPII-bound regions. **B.** KNN (5 NN, Pearson correlation as metric) classification balanced accuracy using either Gene expression or RNAPII signal as input. **C.** UMAP of GTEx RNA-seq samples using RNA-seq signal at genes exons. Displayed colours correspond to the official GTEx Tissue colouring conventions.

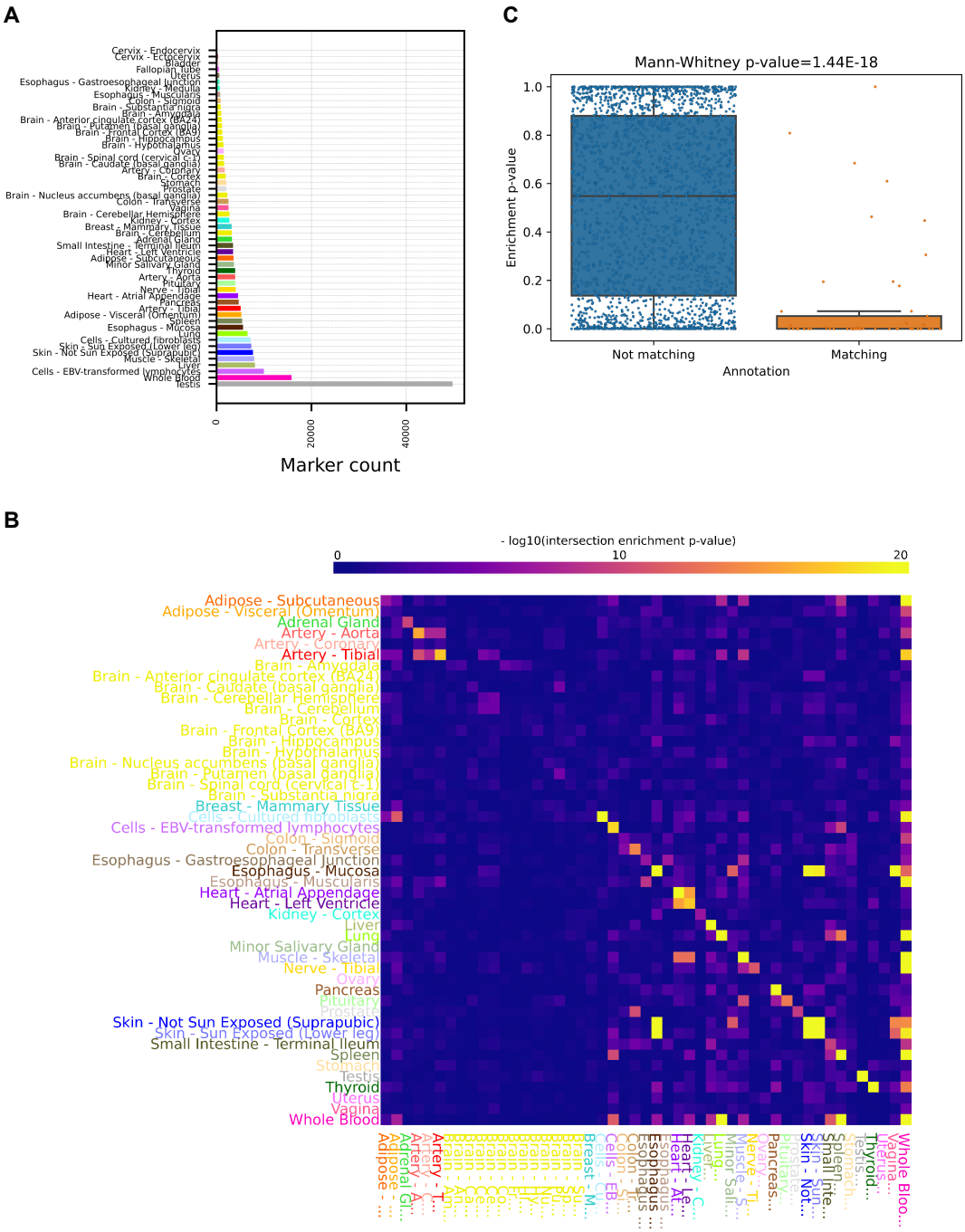
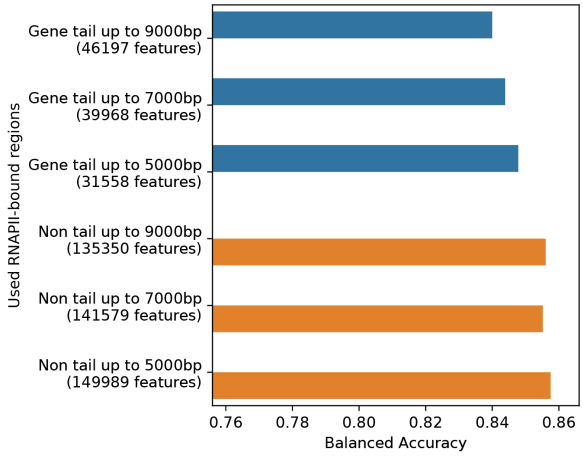


Figure S15 : Tissue-specific regulatory variants are enriched within tissue-specific Intergenic transcripts, related to STAR Methods

A. Number of overexpressed RNAPII-bound regions per GTEX tissue. **B.** Intersection enrichment heatmap between tissue-specific eQTLs (rows) and tissue-specific RNA-seq over-expressed marker RNAPII-bound regions. **C.** Distributions of tissue-matching (i.e. “Artery - Aorta” vs “Artery - Aorta”) and non-matching (i.e. “Artery - Aorta” vs “Artery - Coronary”,“Liver”...) enrichment p-values between tissue-specific eQTLs (rows) and tissue-specific RNA-seq over-expressed markers.

A



B

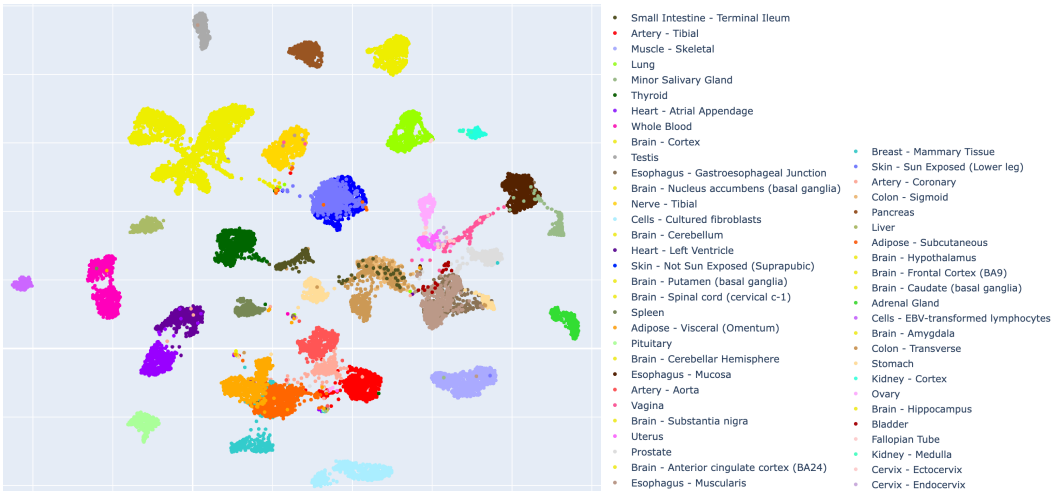


Figure S16 : The intergenic transcriptional signal is not driven by end-of-gene transcription, related to STAR Methods

A. KNN (5 NN, Pearson correlation as metric) classification balanced accuracy using different subsets of intergenic RNAPII-bound regions for classification. **B.** UMAP of GTEx RNA-seq samples using RNA-seq signal at intergenic RNAPII, excluding those located at less than 9,000bp of a Transcription End Site. Displayed colours correspond to the official GTEx Tissue colouring conventions.

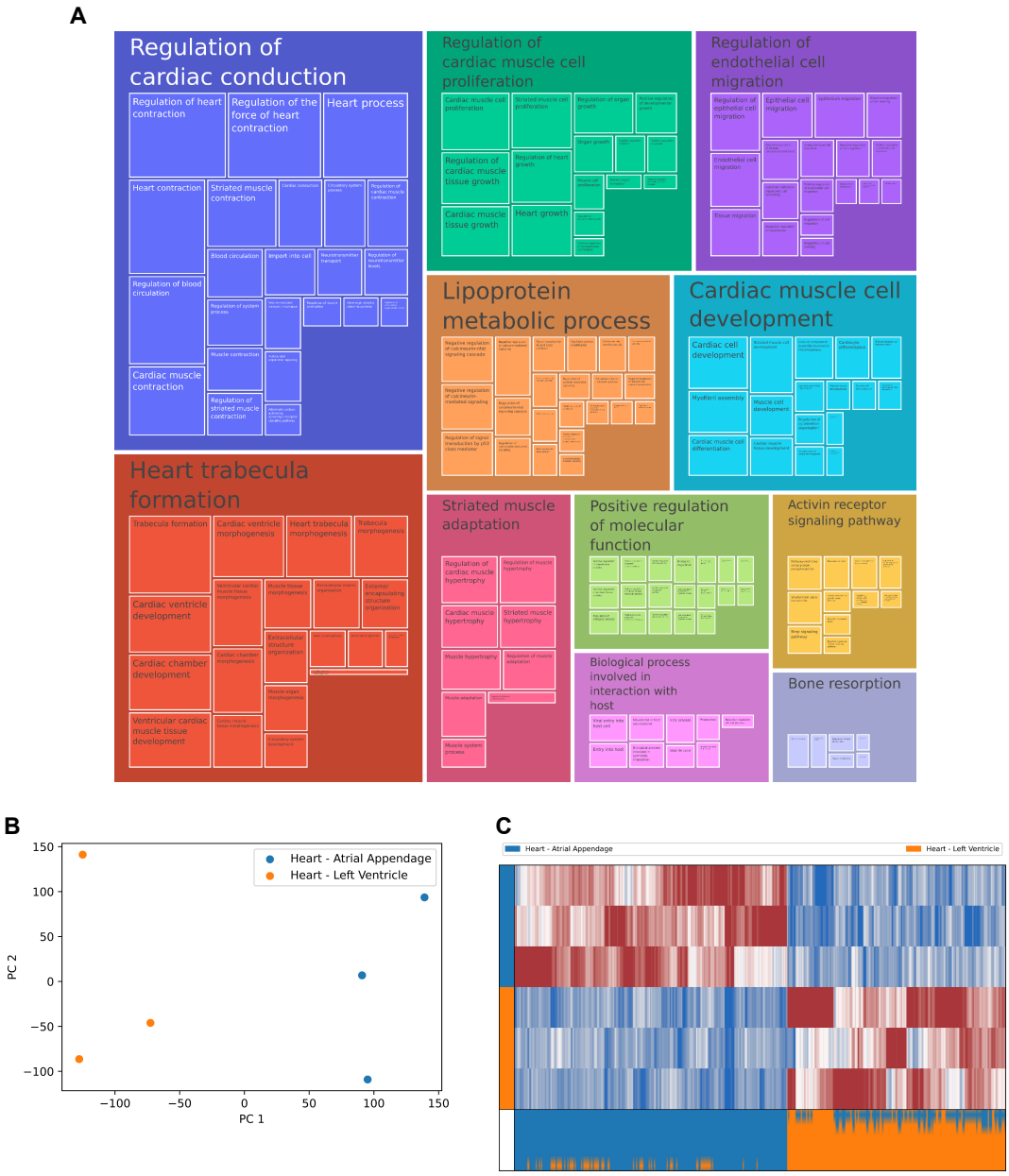


Figure S17 : Differentially expressed RNAPII-bound regions can be detected at smaller sample sizes, related to STAR Methods

A. Clustered GO terms enrichments for genes nearby RNAPII-bound regions differentially expressed between 'Heart - Atrial Appendage' and 'Heart - Left Ventricle' tissues from GTEx in a downsampled n=3 comparison (methods). **B.** First two principal components of RNAPII-bound region expression in a downsampled n=3 comparison. **C.** Heatmap of the Pearson residuals of DE RNAPII-bound regions in a n=3 comparison.

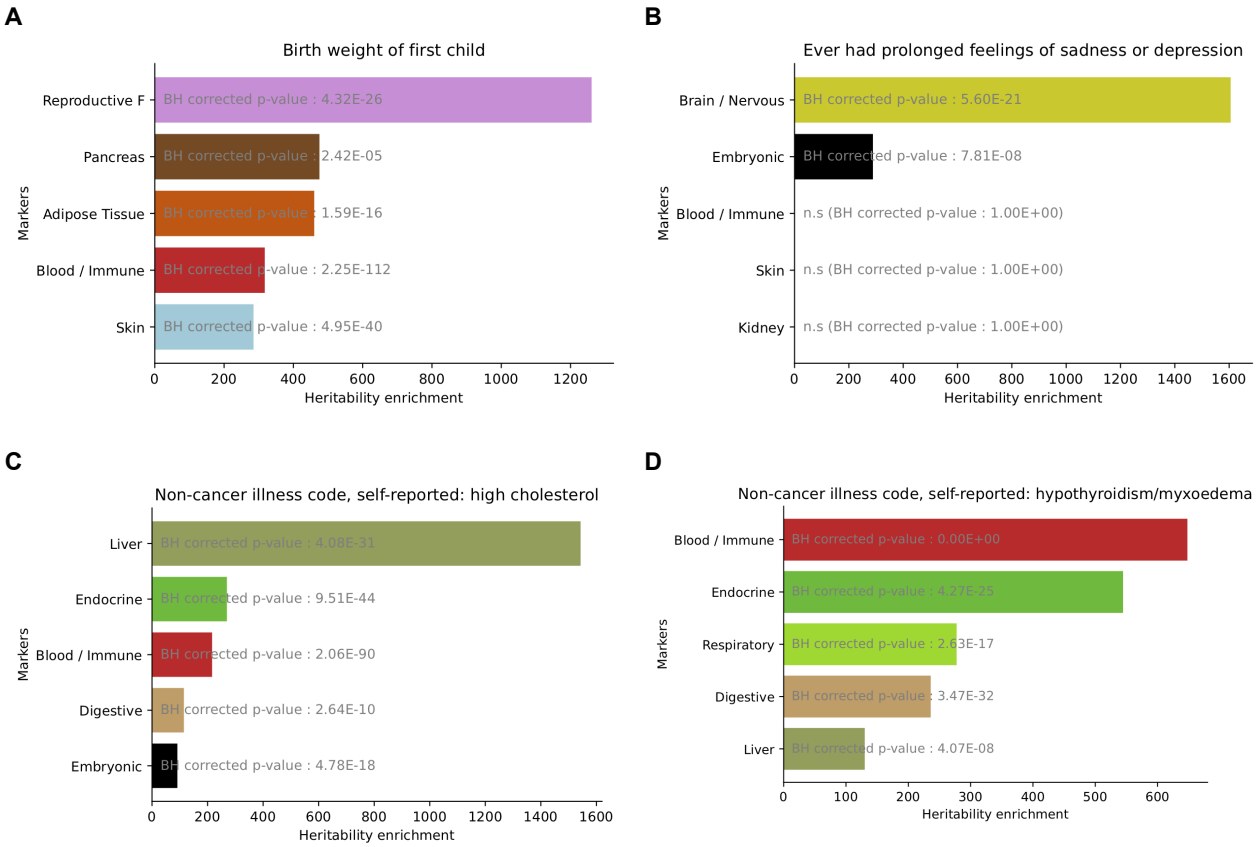


Figure S18 : Per-biotype robustly over-expressed markers display meaningful disease-associated heritability enrichments, related to Figure 5

Top 5 heritability enrichment over robust tissue-specific over-expressed markers (Methods), for 4 disease-associated UK Biobank GWAS traits. Benjamini-Hochberg corrected LD-score regression p-values are indicated and detailed for each trait - marker group pair.

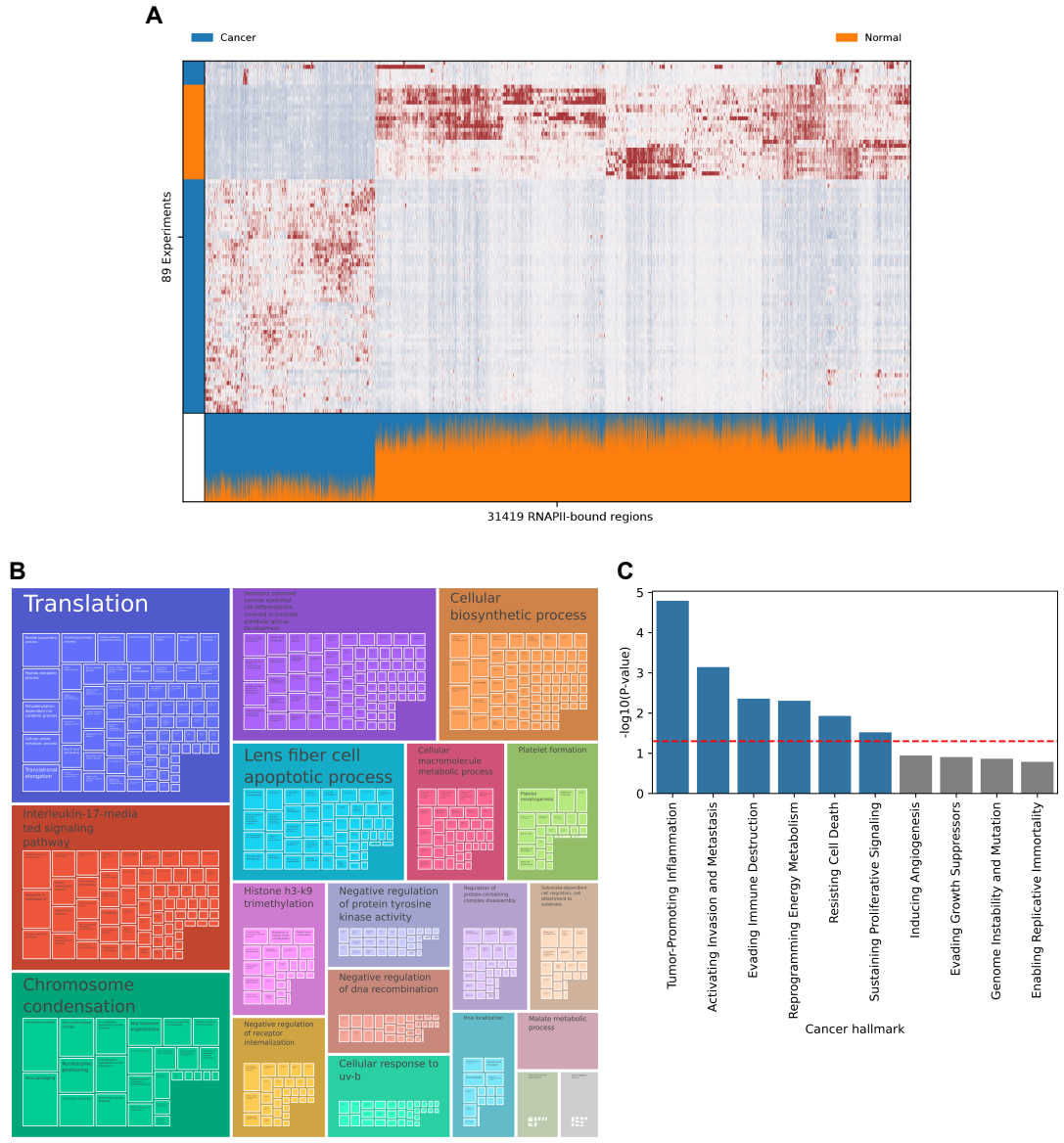


Figure S19 : Non coding transcription captured at RNAPII-bound regions discriminates normal and tumour tissues, related to Figure 6

A. Heatmap of Pearson residuals (clipped at +-3) of DE RNAPII-bound regions in the Kidney Chromophobe Carcinoma dataset (KICH). Red = strongly expressed. Bottom part of the heatmap represents the fraction of normalised reads belonging to either class in each RNAPII-bound region (weighted by class imbalance). **B.** Clustered GO terms enrichments for genes nearby RNAPII-bound regions DE in 7 or more cancers (see methods). All enriched terms are statistically significant (FDR < 0.05, Wald test, see Star Methods). **C.** Gene set enrichment of cancer hallmark genes for genes nearby RNAPII-bound regions DE in 7 or more cancers (see methods). Red dashed line indicates 0.05 p-value threshold.

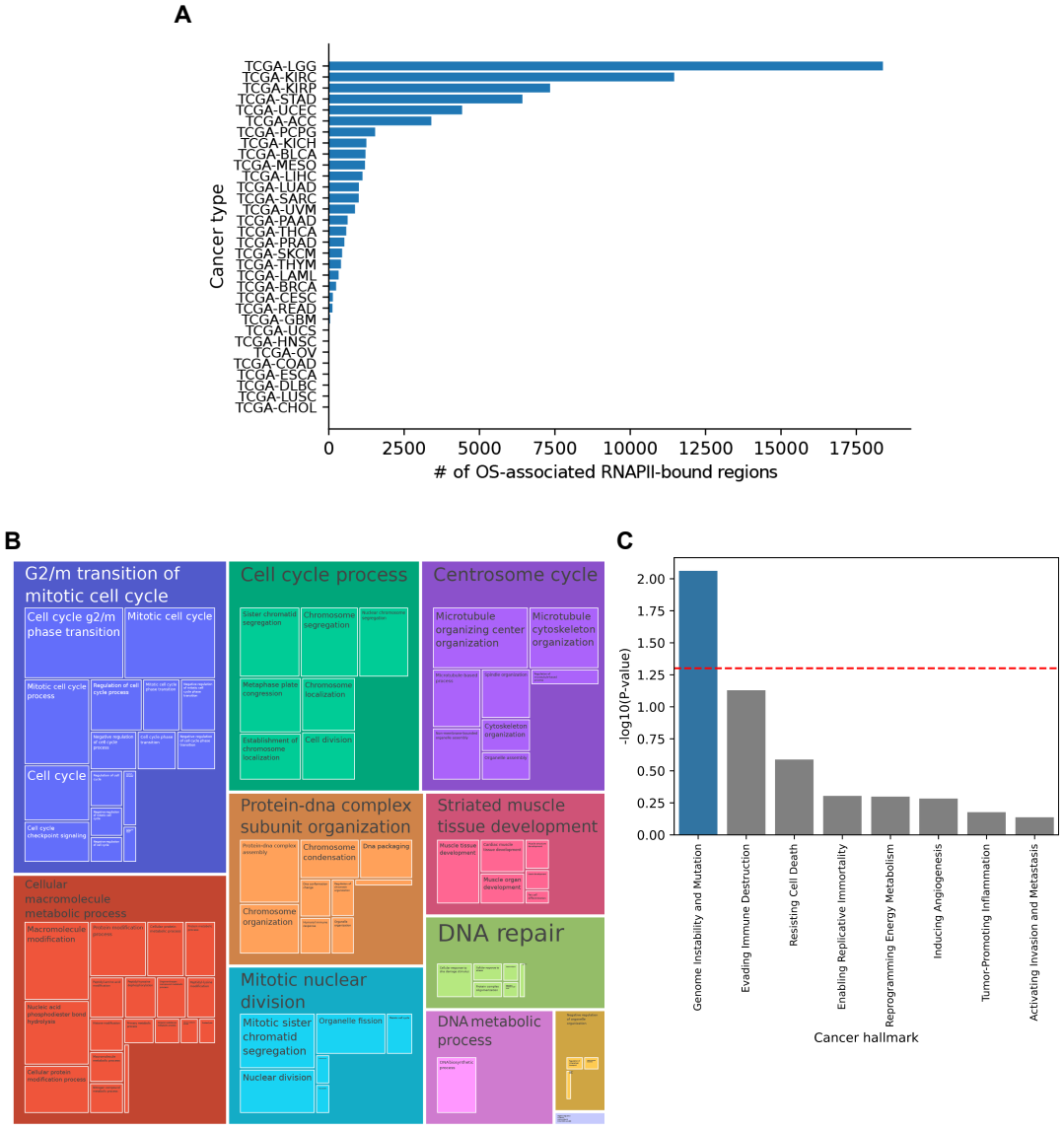


Figure S20: Non coding transcription captured at RNAPII-bound regions is prognostic of the patient's survival, related to Figure 7

A. Number of OS-associated RNAPII-bound regions for each TCGA cancer. **B.** Clustered GO terms enrichments for genes nearby RNAPII-bound regions OS-associated in 5 or more cancers (see methods). **C.** Gene set enrichment of cancer hallmark genes for genes nearby RNAPII-bound regions OS-associated in 5 or more cancers (see Methods). Red line indicates 0.05 p-value threshold.

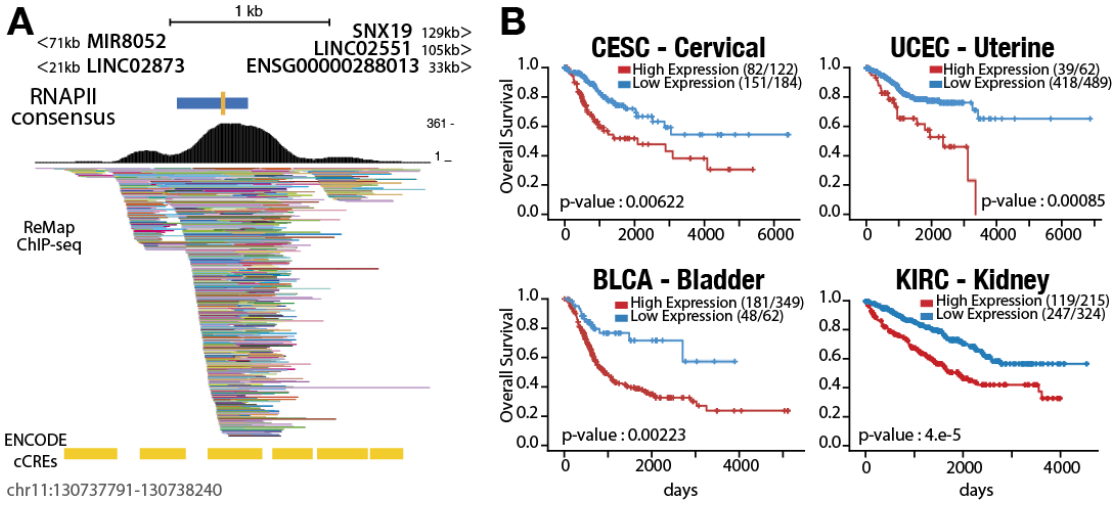


Figure S21: Supplementary multi-cancer prognostic marker, related to Figure 7

A. Genomic landscape of an identified multi-cancer prognostic marker blue bar at chr11:130737791-130738240, located upstream (21kb min) and downstream (33kb min) of non-coding and coding genes. Yellow bars indicate candidate Cis Regulatory Elements (cCREs, Enhancer distal) and ChIP-seq binding from ReMap. This region is located between two non-coding LincRNAs (LINC02873 and ENSG00000288013) at 21kb and 33kb respectively. It is closest to the coding gene SNX19, which encodes a sorting nexin located at 128kb. SNX19 has not been directly linked to cancer survival, other SNXs family members have shown potential prognostic value in various cancers. Decreased expression of SNX1 has been associated with overall survival in colorectal cancer, and down-regulation of SNX2 leads to drug resistance in lung cancer. **B.** Kaplan-Meier survival analysis of Cervical, Uterine, Bladder and Kidney TCGA cancer patients with high (red) and low (blue) expression from the intergenic RNAPII bound region in A. This RNAPII-bound region reveals a previously unknown correlation between its expression and survival in cervical, uterine, and breast cancers (**Figure 7D**).

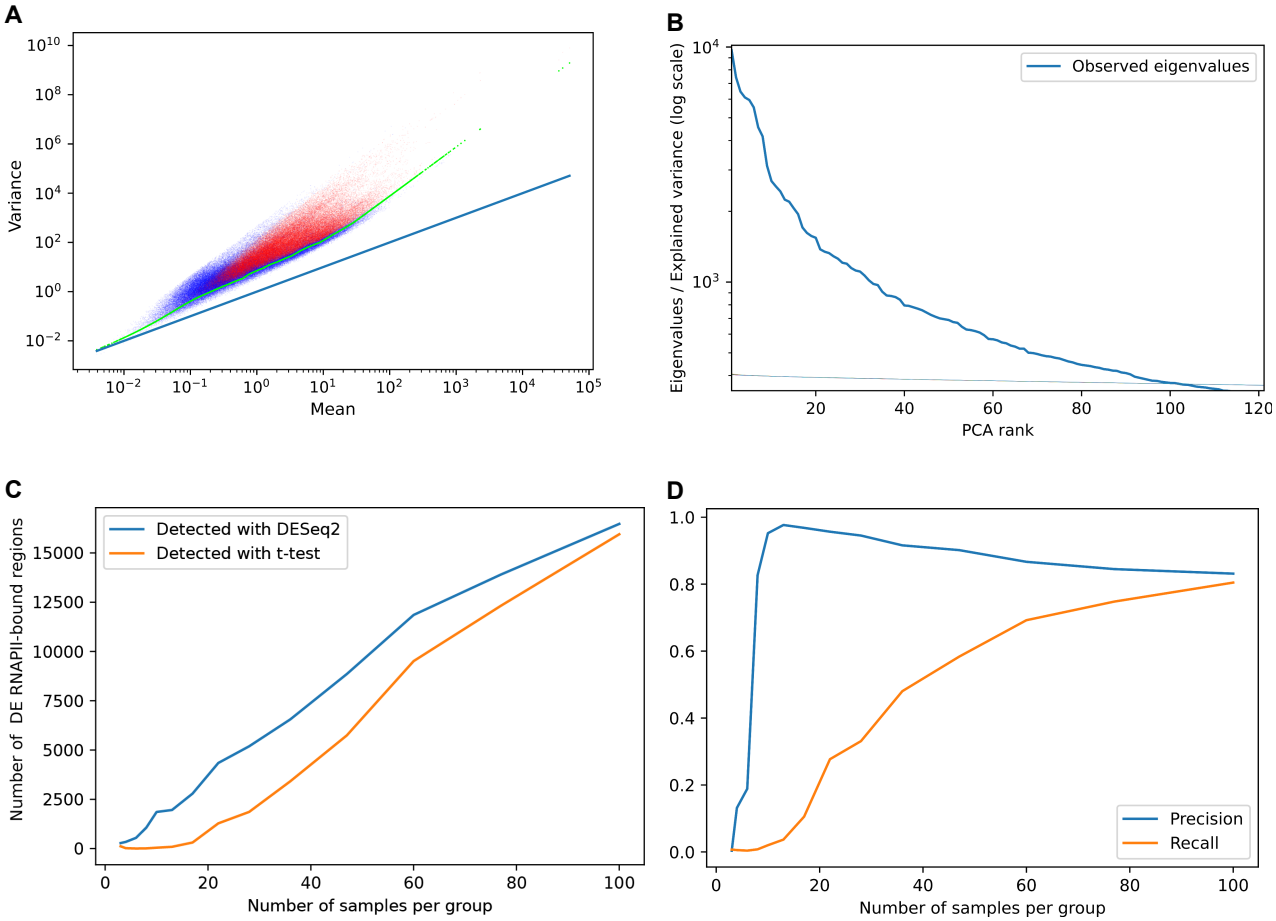
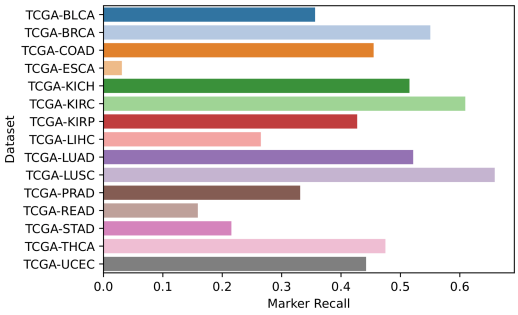


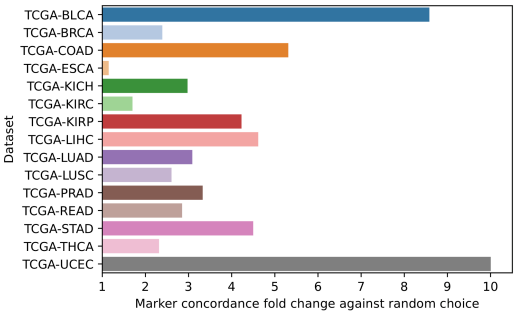
Figure S22 : Mean-variance trendline, Feature selection and PCA Permutation Parallel analysis, related to STAR Methods

A. Scatterplot of the mean and variance of the normalised counts of the 181,547 RNAPII-bound regions in the ENCODE RNA-seq dataset. Red dots are selected “highly variable” RNAPII-bound regions, green dots represent the fitted mean-variance trendline. Blue line is a Poisson mean-variance relationship. **B.** Observed eigenvalues (or explained variance) for each component of the PCA performed on the pearson residuals of selected RNAPII-bound regions in the ENCODE RNA-seq dataset. 100 thinner lines (stacked on the graph) are corresponding to the PCA eigenvalues of each of the 100 permutations of the dataset. Here, only the 102 first components are retained. **C.** Average (over 10 downsampling iterations) number of detected DE RNAPII-bound regions between the two heart tissues in function of the number of samples per group. **D.** Average (over 10 downsampling iterations) precision and recall of the t-test in function of the sample size, using DESeq2 DE-bound regions as a reference.

A



B



C

TCGA Cancer	Reference GTEx tissue
TCGA-BLCA	Bladder
TCGA-BRCA	Breast
TCGA-COAD	Colon
TCGA-ESCA	Esophagus
TCGA-KICH	Kidney
TCGA-KIRC	Kidney
TCGA-KIRP	Kidney
TCGA-LIHC	Liver
TCGA-LUAD	Lung
TCGA-LUSC	Lung
TCGA-PRAD	Prostate
TCGA-READ	Colon
TCGA-STAD	Stomach
TCGA-THCA	Thyroid
TCGA-UCEC	Uterus
TCGA-HNSC	N/A

Figure S23 : Using GTEx normal tissues instead of TCGA normal tissue, related to STAR Methods

The predictive power of per-cancers markers was confirmed as we accurately separated tumours from normal GTEx tissues (instead of TCGA normal tissues) using a machine learning classifier **A.** Recall statistics, the fraction of shared markers between markers identified from TCGA normal or GTEx normal. **B.** The Fold Change analysis corresponding to the observed number of shared markers against the expected value if they were chosen at random. All fold changes are statistically significant (hypergeometric test, p<1e-300) except for TCGA-ESCA. **C**. Table showing GTEx normal tissue to TCGA cancer matching.

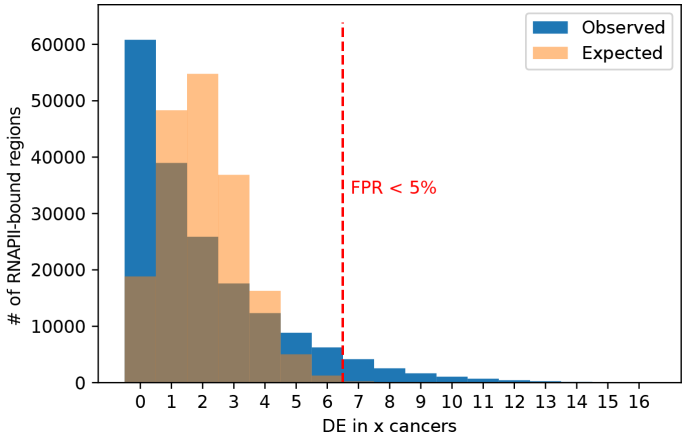


Figure S24 : Distribution of differentially expressed RNAPII-bound regions in cancers, related to STAR Methods

Distributions of the number of cancers of RNAPII-bound region is DE in, as observed in the TCGA dataset, and by cancer-wise random permutations. Red dashed line indicates the threshold at which less than 5% of observed DE RNAPII-bound regions are DE in more cancer than expected by chance.

5.2 Étude de l'occupation de la Pol II et de l'activité transcriptionelle pour les transposons

Les transposons couvrent près de la moitié de notre génome et en restent néanmoins une partie mal comprise. Il a été observé que certains transposons étaient exprimés et produisaient des lncRNAs, certains étant même essentiels à l'identité pluripotente des cellules embryonnaires[117]. Il est ainsi intéressant d'étudier quels transposons et quelles classes de transposons peuvent être potentiellement transcrites au niveau des 180 000 pics consensus Pol II que nous avons identifié.

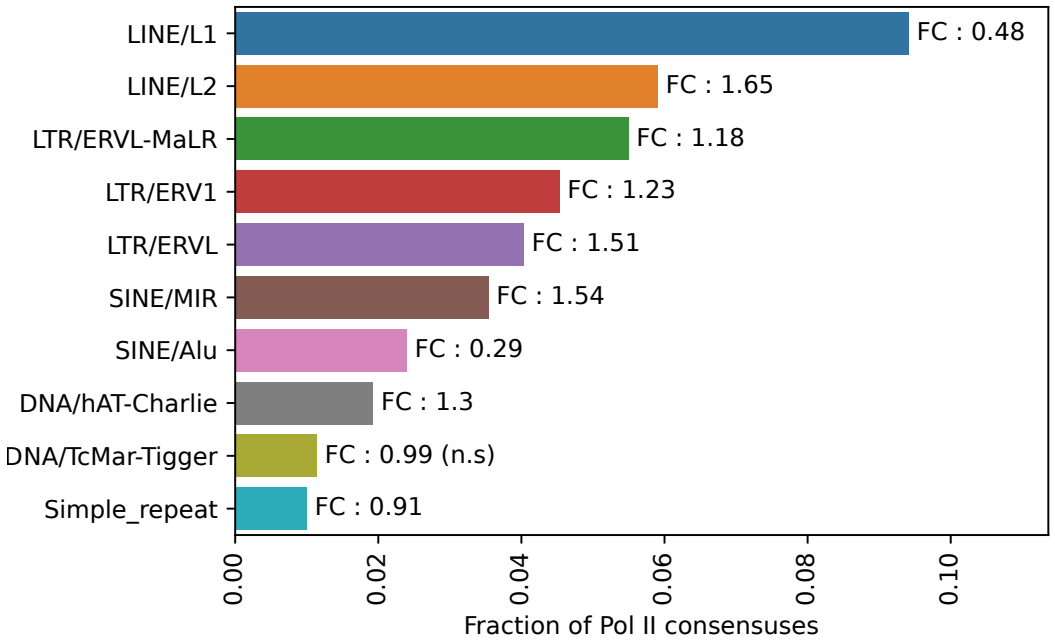


FIGURE 5.1 – Fraction des pics de consensus intersectés pour les dix familles d'éléments répétés les plus intersectées. FC correspond au Fold Change par rapport à des régions intergéniques aléatoires. Tous les Fold Change sont statistiquement significatifs (FDR < 0,05, test binomial) sauf indication contraire.

Nous avons observé que la Pol II avait une tendance à éviter les éléments répétés du génome (24% moins d'intersections qu'attendu, voir article Figure S6A). Toutefois, en observant les enrichissements par type de transposon (Figure 5.1), nous pouvons observer que certains transposons sont plus fréquemment observés qu'attendus, tels que les LINE L2, les ERVL ou les MIR ; d'autre part les séquences Alu et les LINE L1 sont fortement sous-représentés. Étonnamment, les MIRs et les Alus sont originellement transcrits par l'ARN Polymérase III car portant un promoteur de tRNA, et cependant les MIRs apparaissent sur-représentés au contraire des éléments Alu.

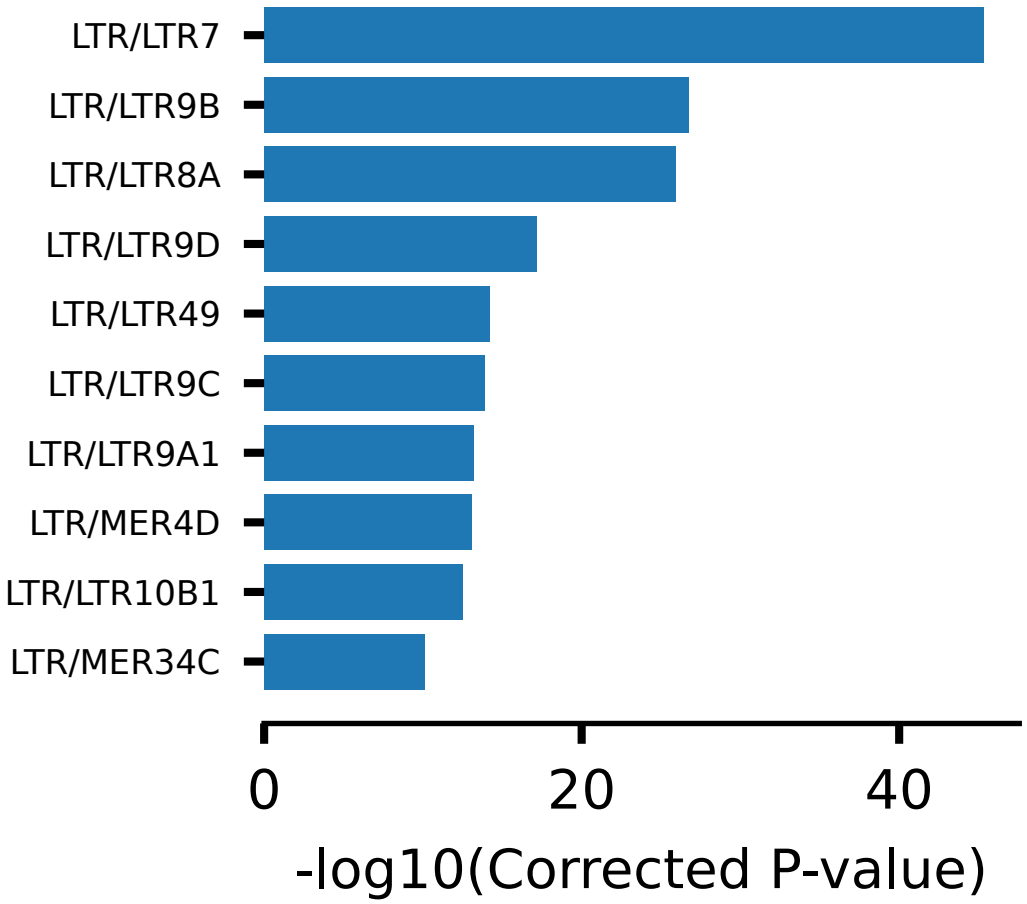


FIGURE 5.2 – Enrichissement en éléments répétés parmi les pics consensus actifs dans les cellules embryonnaires par rapport à tous les pics consensus. La p-value provient d'un test hypergéométrique, puis est corrigée par la méthode de Benjamini-Hochberg pour toutes les familles d'éléments répétés testés.

Dans les pics consensus du cluster identifié comme étant principalement actif dans les cellules embryonnaires ou pluripotentes, nous retrouvons un grand nombre d'éléments répétés LTR (Figure 5.2). Il a été montré que ces éléments étaient en effet largement transcrits dans les cellules embryonnaires humaines[67], et la famille LTR7 comprend l'un des lncRNA rétroviraux qui a été montré comme étant requis pour l'identité de cellule embryonnaire pluripotente[117].

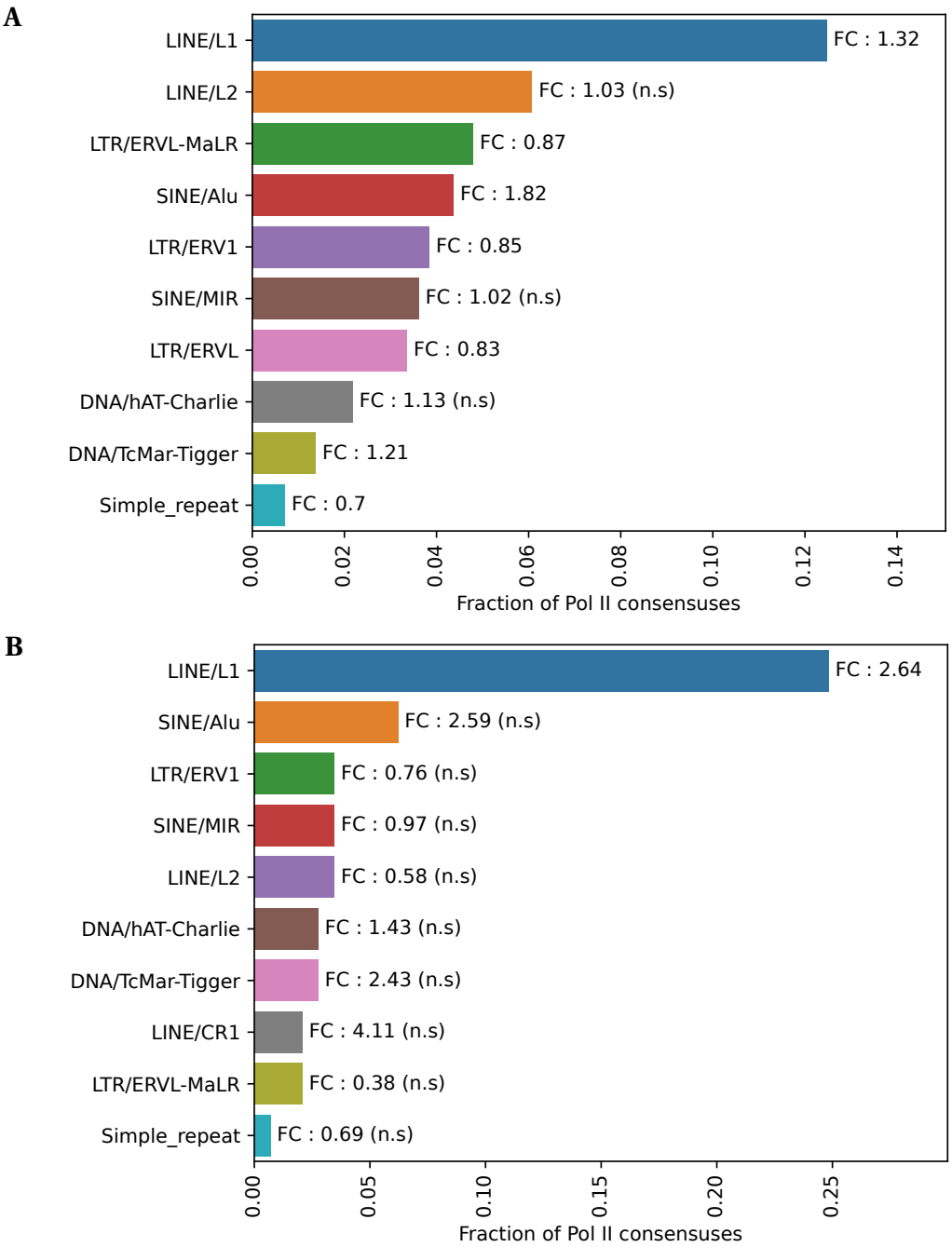


FIGURE 5.3 – Fraction des pics de consensus intersectés pour les dix familles d'éléments répétés les plus intersectées. FC correspond au Fold Change par rapport à tous les pics consensus. Les Fold Change sont statistiquement significatifs (FDR < 0,05, test hypergéométrique) sauf indication contraire. **A.** Parmi les marqueurs d'expression différentielle pluri-cancer entre tumeurs et tissus normaux. **B.** Parmi les marqueurs de survie pluri-cancer.

Dans les cancers, dans les 11 000 marqueurs d'expression différentielle observés dans un grand nombre de cancers, nous observons au contraire plus d'éléments LINE L1 et Alu qu'attendu (Figure 5.3). De même, parmi les 150 marqueurs de survie communs à plusieurs cancers, nous observons encore un fort enrichissement en éléments Alu et L1, avec jusqu'à près de 25% des marqueurs étant situés sur un élément L1. Il a été montré que l'expression de ces derniers corrèle avec des mutations sur la protéine TP53 dans les cancers, un des oncogènes les plus étudiés ; ils semblent également être étroitement liés au cycle cellulaire[127]. Les éléments Alu ont été relativement peu étudiés dans ce cadre, mais ils semblent porter des marques de méthylation anormales dans les cancers[89].

Ces résultats sont toutefois limités par l'utilisation de données short read à mapping unique. Ainsi, la présence d'éléments répétés sous représentés est potentiellement causée par la perte de reads du à un mapping non-ambigu difficile au niveau de certains éléments répétés. De plus, nous avons utilisé l'assemblage du génome hg38 ; tandis que l'assemblage T2T [140], publié début 2022, est plus précis, et ce particulièrement au niveau des éléments répétés du génome. Ainsi, cette version identifie 8.7% de séquences répétées en plus par rapport à la version hg38, avec 112.9% de repeats simples et 96.6% de repeats satellite additionnelles.

5.3 Etude de la transcription intergénique par séquençage 10X

J'ai également étudié si il était envisageable d'utiliser l'atlas Pol II pour échantillonner le signal transcriptionnel à l'échelle de la cellule unique dans les protocoles de séquençage 10X, en utilisant un jeu de données références comprenant plus de 10 000 cellules PBMC. Le signal est ici extrêmement faible (100 UMIs par cellule contre plus de 20 000 en utilisant les gènes). Lors d'une approche classique, utilisant les gènes comme référence, les profondeurs de séquençage faibles limitent la faisabilité de la distinction entre types cellulaires. Il est ainsi ici très difficile de différentier finement les groupes cellulaires, même s'il est possible d'identifier les groupes principaux (T-Cells, B-Cells, Monocytes et NK Cells, voir Figure 5.4).

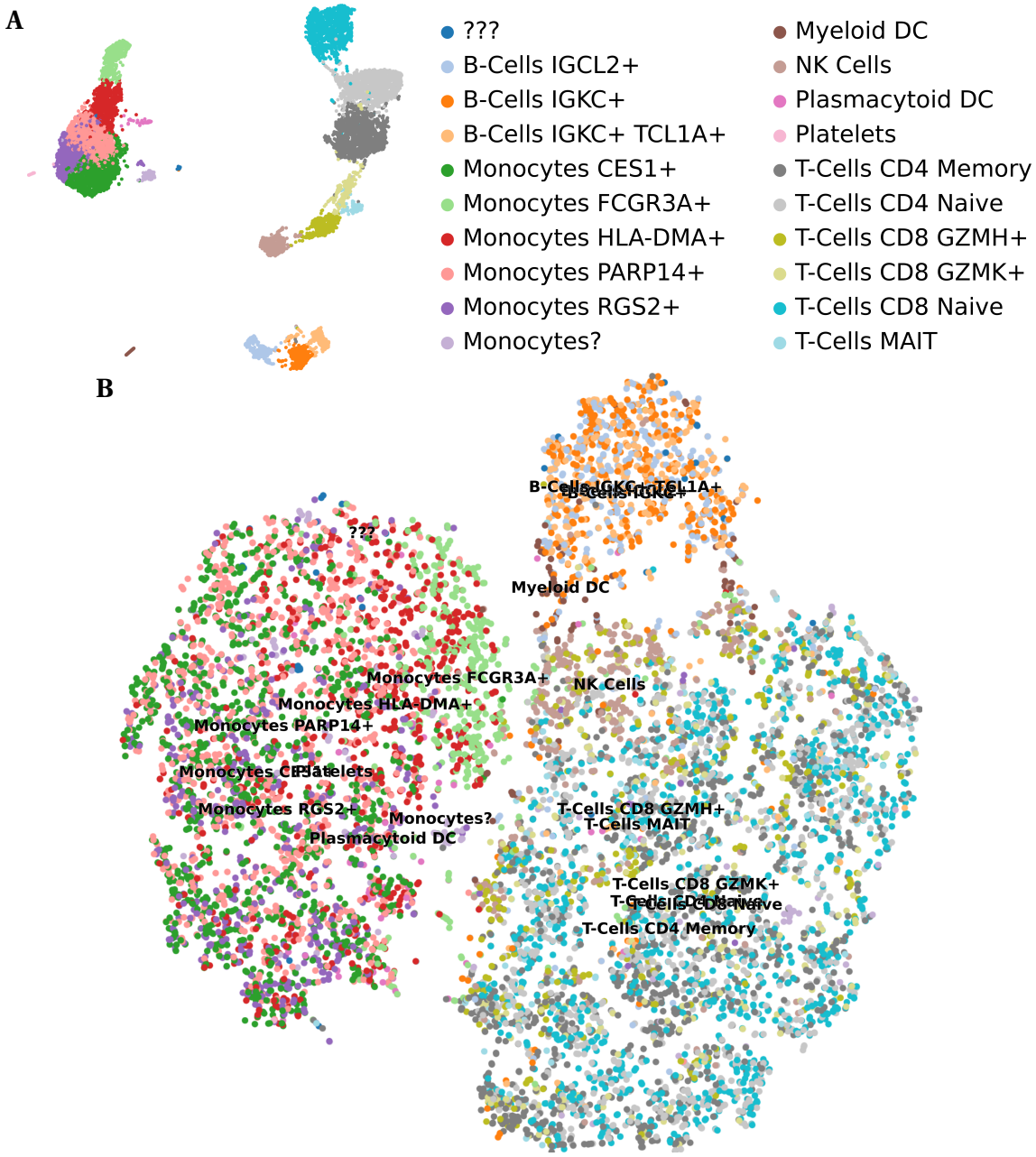


FIGURE 5.4 – Visualisation UMAP du jeu de données 10k PBMCs de 10X genomics. Les analyses ont étées réalisées à l'aide de MUFFIN (voir Section 6.3), aux paramètres par défaut. **A.** En utilisant la table de comptage des UMIs situés sur les gènes. **B.** En utilisant la table de comptage issue de l'atlas Pol II dans l'intergénique. Les labels des clusters ont été obtenus sur la table de comptage des gènes, et transférés par correspondance des Barcodes cellulaires.

6 Généralisation des méthodes développées

Sommaire

6.1 Identification de modules de régulation pour la base de données ReMap 148
6.2 Travail préliminaire sur l'étude combinatoire de Facteurs de Transcription160
6.3 Muffin : un package python contenant des outils génériques pour l'analyse de données de séquençage fonctionnel 163

6.1 Identification de modules de régulation pour la base de données ReMap

La base de données ReMap est une base de données publique recensant les sites de fixation des régulateurs transcriptionnels sur l'ADN, identifiés principalement par la technique ChIP-seq. Dans les données ReMap, les pics de TFs sont principalement concentrés dans des régions régulatrices, formant des régions à haute densité en TFs. L'utilisation de mon approche d'identification d'élément régulateurs par densité est alors particulièrement indiquée ici, permettant une identification de CRMs bien plus fine que la méthode précédemment utilisée, basée sur la fusion des intervalles génomiques des pics se chevauchant (Figure 6.1). Ma méthode a été intégrée au pipeline de génération des données de ReMap, et a permis d'identifier les CRMs dans les différentes espèces répertoriées.

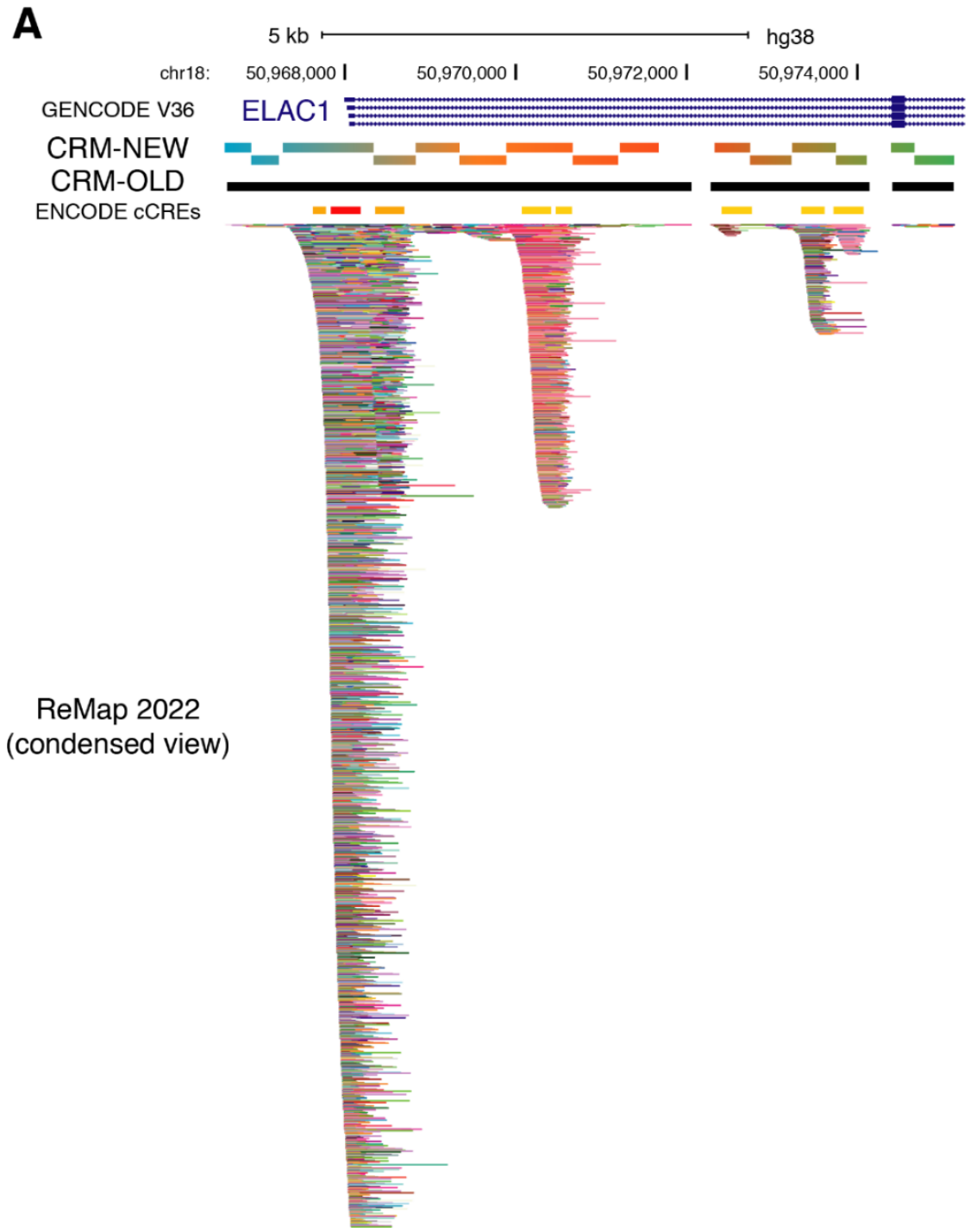


FIGURE 6.1 – Identification de CRMs. La région promotrice du gène ELAC1 est correctement découpée en plusieurs éléments régulateurs.

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

© The Author(s) 2021. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/nar/article/50/D1/D316/6423925 by guest on 23 October 2023

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

Downloaded from https://academic.oup.com/nar/article/50/D1/D316/6423925 by guest on 23 October 2023

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

Downloaded from https://academic.oup.com/nar/article/50/D1/D316/6423925 by guest on 23 October 2023

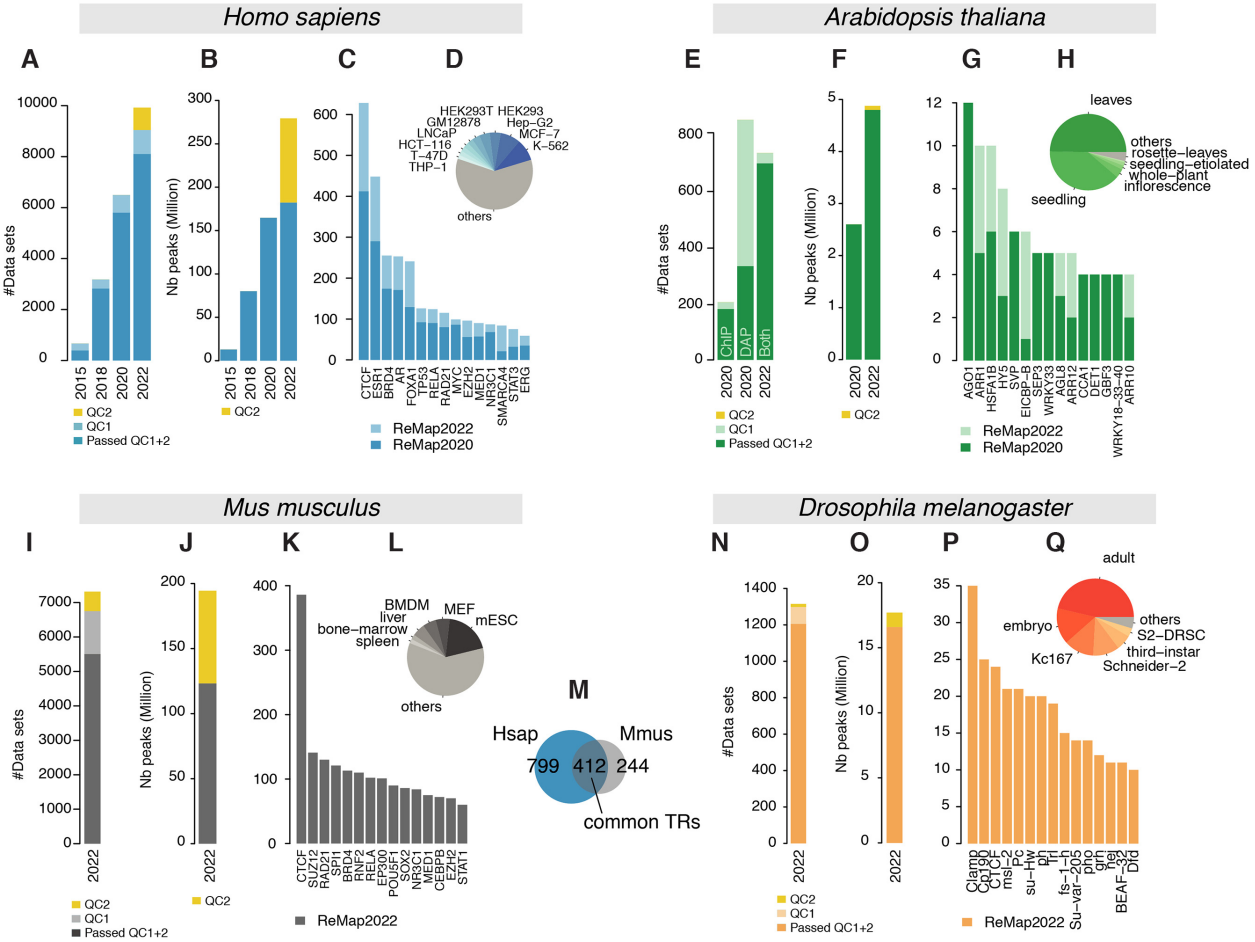


Figure 1. Overview of the ReMap 2022 database growth. (**A**) Analysed Human datasets growth in ReMap 2022 compared to 2020, 2018 and 2015, removed datasets from the new quality control (QC2) in yellow. (**B**) Human ChIP-seq peaks growth in ReMap 2022 compared to previous releases, yellow corresponds 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

Downloaded from https://academic.oup.com/nar/article/50/D1/D316/6423925 by guest on 23 October 2023

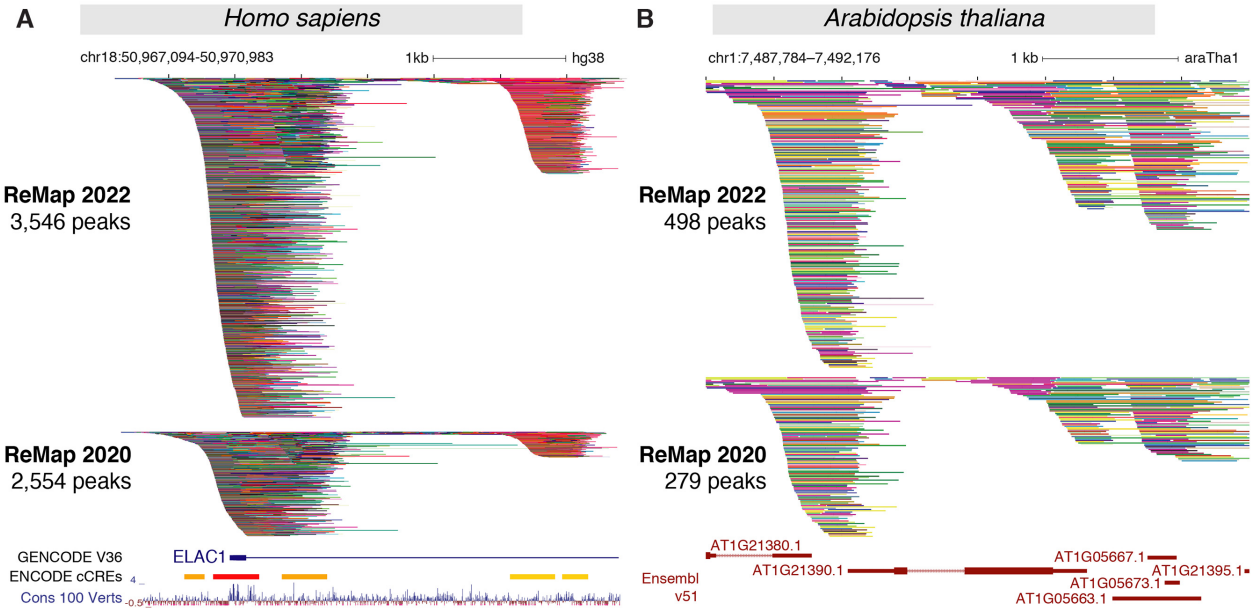


Figure 2. Updated ReMap 2022 regulatory atlas for Human and Plant. (**A**) ReMap 2022 Human DNA-protein binding pattern of 8103 valid datasets. This genome browser example of the DNA-binding peak depth of ReMap 2022 atlas is compared to ReMap 2020 at the vicinity of the ELAC1 promoter (chr18:50 967 094–50 970 983). 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

Downloaded from https://academic.oup.com/nar/article/50/D1/D316/6423925 by guest on 23 October 2023

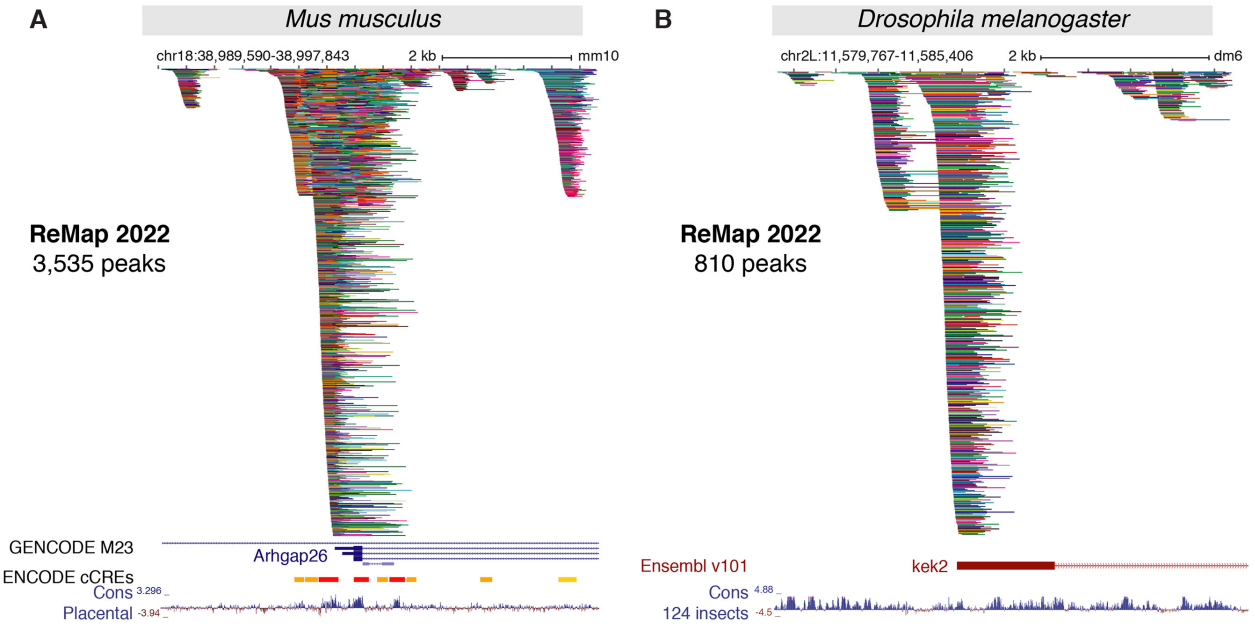


Figure 3. New ReMap 2022 regulatory atlas for *Mus musculus* and *Drosophila melanogaster*. (**A**) A genome browser view of the first Mouse ReMap 2020 atlas for transcriptional regulators at the vicinity of the Arhgap26 gene (chr18:38 989 590–38 997 843). Genome tracks correspond to the GENCODE M23 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.

Downloaded from https://academic.oup.com/nar/article/50/D1/D316/6423925 by guest on 23 October 2023

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)

Downloaded from https://academic.oup.com/nar/article/50/D1/D316/6423925 by guest on 23 October 2023

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.

Downloaded from https://academic.oup.com/nar/article/50/D1/D316/6423925 by guest on 23 October 2023

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.

Downloaded from https://academic.oup.com/nar/article/50/D1/D316/6423925 by guest on 23 October 2023

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.

Downloaded from https://academic.oup.com/nar/article/50/D1/D316/6423925 by guest on 23 October 2023

6.2 Travail préliminaire sur l'étude combinatoire de Facteurs de Transcription

A la manière de l'étude portant sur l'ARN Polymérase II, les CRMs identifiés pour la base de données ReMap peuvent être caractérisés par les TFs pouvant s'y fixer, car la fonction d'un CRM dépend principalement des TFs qu'il est capable de recruter. En étudiant les combinatoires de TFs, il pourrait être possible d'identifier des classes de CRMs régulés par des TFs similaires. Dans un premier travail préliminaire, j'ai utilisé un auto-encoder variationnel[96] pour ordonner les CRMs (filtrés de façon à obtenir au moins 10 TFs par CRM) par similarité dans un espace à une seule dimension pour visualiser rapidement leurs caractéristiques.

Ce réseau de neurones prend en entrée un CRM, chacun caractérisé par la présence ou absence de chacun d'un des 1100 TFs de ReMap sous la forme d'un vecteur binaire. Le réseau de neurones possède 3 couches de 512 unités avec une activation ReLu, puis une couche prédisant les paramètres d'une distribution normale univariée N(μ, σ) qui est échantillonnée pour le réseau d'encodage. Le décodeur possède également possède 3 couches de 512 unités avec une activation ReLu, puis la couche de sortie prédit la probabilité de présence ou non de chacun des 1100 TFs de ReMap (activation sigmoïde). La fonction de perte est la fonction de perte standard des auto-encoders variationnels, c'est à dire la divergence de Kullback-Leibler, plus le logarithme de vraisemblance (ici d'une distribution de bernoulli).

$$\begin{aligned} L(x,\hat{x},\sigma,\mu) &= -\sum_{i=1}^{TFs}\left[x_i\log(\hat{x}_i)+(1-x_i)\log(1-\hat{x}_i)\right]+\frac{1}{2}\left(1+\log(\sigma^2)-\mu^2-\sigma^2\right) \\ &= -\text{Binary Cross-Entropy}(x,\hat{x})+\text{KL Divergence}(q_\phi(z|x)\|p(z)) \end{aligned}$$

Le réseau de neurones est entraîné avec l'optimiseur Adam aux paramètres par défaut.

L'intérêt d'utiliser un réseau de neurones ici est principalement du à la large quantité de données couplée au temps d'entraînement des réseaux de neurones par descente de gradient qui a une complexité en temps linéaire et reste relativement rapide même pour de grands jeux de données (ici 2.7 millions de points). Sur la heatmap suivante (Figure 6.1), les CRMs sont ordonnés selon la valeur de μ, les TFs par clustering hiérarchique.

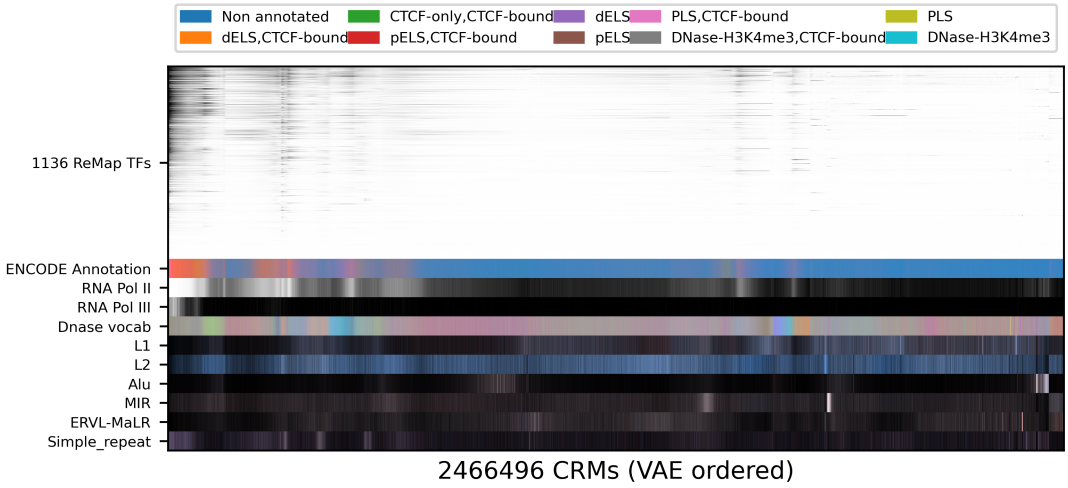


FIGURE 6.2 – Visualisation des CRMs. De haut en bas : Présence (noir) ou absence (blanc) des TFs (lignes) dans les CRMs (colonnes) ; Annotation ENCODE cCREs du CRM ; présence ou Absence de Pol II ; Présence ou absence de Pol III ; tissu préférentiel d'accessibilité de la chromatine determiné par DNase-seq (chaque couleur correspond à un tissu différent) ; présence d'un élément répété L1, L2, Alu, MIR, ERVL ou simple repeat, la couleur correspond à la position normalisée sur l'élément répété consensus (début 5' en rouge, fin 3' en bleu).

Cette étude préliminaire semble révéler des groupes de CRMs portant des TFs similaires, et également possédant des caractéristiques biologiques similaires. Certaines combinatoires de TFs semblent propres à certaines catégories d'éléments régulateurs identifiés par ENCODE (telles que les enhancers, promoteurs ou insulateurs CTCF), ou semblent être particulièrement capables de recruter la Pol II ou la Pol III. Au contraire, certains ne semblent pas avoir une forte affinité pour la Pol II ; il est possible que l'on ait des régions silencers ou enhancers. Certaines combinatoires semblent également associés une ouverture de la chromatine qui est tissu-spécifique. Enfin, certaines combinatoires précises semblent être caractéristiques d'éléments transposables. Des travaux futurs pourraient permettre permettre l'identification de clusters de combinatoires de TFs et leur caractérisation, à l'instar des analyses qui ont été effectuées sur le jeu de données ChIP-seq Pol II. Il pourrait également être intéressant d'étudier

la séquence de ces CRMs similaire pour y trouver des motifs, et explorer différentes méthodes d'analyse combinatoire de CRMs.

6.3 Muffin : un package python contenant des outils génériques pour l'analyse de données de séquençage fonctionnel

Les méthodes que j'ai utilisé et développé au cours de mon projet principal sont en partie inspirées de méthodes utilisées pour l'analyse de données single-cell, mais applicables à d'autres types de données. Afin de faciliter la réutilisation de ces méthodes, j'ai développé un package python, Muffin, intégrant des approches d'analyse de table de comptages issues de séquençage fonctionnel, un outil d'enrichissement fonctionnel sur régions génomiques, et mon outil d'identification de pics consensus. Celui-ci s'inscrit dans l'écosystème du package d'analyse single-cell Scanpy[194], afin de maximiser sa facilité d'utilisation. J'ai testé mon outil sur trois jeux de données variés : une étude à grande échelle de l'accessibilité de la chromatine dans les tumeurs par ATAC-seq, un jeu de données single-cell RNA-seq portant sur plus de 10 000 PBMCs réalisé par 10X, et enfin un jeu de données de ChIP-seq ciblant l'histone H3K4Me3 dans des cellules immunitaires provenant d'ENCODE.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.11.570597; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Journal Title, 2023, pp. 1–11

doi: BIORXIV/2023/570597
Advance Access Publication Date: Day Month Year
Paper

PAPER

MUFFIN : A suite of tools for the analysis of functional sequencing data

Pierre de Langen[1] and Benoit Ballester[1,*]

[1]Aix-Marseille Univ, INSERM, TAGC, Marseille, France
[*]Corresponding author. benoit.ballester@inserm.fr

Abstract

The large diversity of functional genomic assays allows for the characterization of non-coding and coding events at the tissue level or at a single-cell resolution. However, this diversity also leads to protocol differences, widely varying sequencing depths, substantial disparities in sample sizes, and number of features. In this work, we have specifically designed a suite of tools for exploring the non-coding genome, particularly for identifying consensus peaks in peak-called assays, as well as linking non-coding genomic regions to genes and performing Gene Set Enrichment Analyses. We demonstrate that a generic but flexible count modelling approach can be utilised to compare different conditions across a broad range of genomic assay such as ENCODE H3K4Me3 ChIP-seq, scRNA-seq and TCGA ATAC-seq. Our Python package, MUFFIN, offers a suite of tools to address common issues associated with high-dimensional genomic data, such as normalisation, count transformation, dimensionality reduction, differential expression, and clustering. Additionally, our tool integrates with the popular Scanpy ecosystem and is available on Conda and at https://github.com/pdelangen/Muffin.

Introduction

The diversity of whole-genome functionnal sequencing protocols allows us to measure a wide variety of regulatory activities across the genome. These range from transcriptomic signals, which can be detected using techniques like RNA sequencing or Cap Analysis of Gene Expression, to epigenetic signals, which can be identified with methods like ChIP-seq or ATAC-seq. The latter respectively quantify the binding of target proteins to the DNA and regions of open chromatin. The number of observations (e.g., cells or tissue samples) can also vary significantly, depending on the experimental design, from a standard two-condition triplicate protocol to thousands of observations in large-scale integrative analyses or sequencing at single-cell resolution.

These whole-genome sequencing protocols involve the mapping of sequencing reads onto a reference genome. These are then counted at specific genomic locations to effectively serve as molecular counters. The genomic locations typically originate either from reference databases (for gene-centric analyses) or are identified de-novo through peak-calling algorithms (for epigenetic or regulatory non-coding events analyses). The measured sequencing signal's intensity can also be highly diverse, ranging from a few thousand UMIs per observation in single-cell analyses to hundreds of millions of reads in bulk sequencing. However, a common aspect across all these whole-genome sequencing protocols is that the sequencing signal per observation ultimately gets quantified through counts in a large number of features (e.g., genes or genomic peaks).

We demonstrate that a generic methodology can be employed to conduct between-observation comparisons in a wide variety of experimental designs, with low to high observation counts, and shallow to deep sequencing. This is accomplished by extending state-of-the-art approaches developed for bulk and single-cell sequencing. Notably, we build on previous statistical models and extend them to account for sequencing protocols with input or control sequencing.

Furthermore, we offer tools to facilitate work with non-coding regions of the genome. The methods presented in this paper can be easily reused and are integrated into a modular Python package, MUFFIN. Our tool is fully interoperable with the functions of the popular Scanpy tool and its ecosystem, which was initially designed for single-cell analysis, but whose tools can be adapted for broader applications.

Methods

Sequencing signal in arbitrary genomic features

Our framework uses raw, non-normalized read or UMI count values over genomic features. For read-based data, we provide a simple wrapper to FeatureCounts [14] to retrieve counts in BAM files located in query genomic features provided in BED, GTF, or SAF format. We support any type of genomic feature, as the user can use genes or reference maps for regulatory elements such as the ENCODE CCREs [19] or DNase hypersensitive regions catalogue [18]. For the de novo discovery of genomic elements, typically done after using a

© The Author 2023. LaTeX template by Oxford University Press.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.11.570597; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

peak-calling algorithm, we provide a simple tool detailed in the next paragraph for the identification of consensus peaks across a large number of experiments. Alternatively, the user can provide their own count tables, which can be easily formatted in the standardised Python AnnData format [27] using the provided helper functions. An overview of our tools is available in Figure 1, and a detailed description of the provided tools is given in the methods below.

High resolution consensus peak identification in large datasets

When working with sequencing data that target regulatory elements, the signal typically manifests in the form of peaks in sequencing protocols such as ChIP-seq, ATAC-seq, or CAGE. Various peak-callers have been developed to identify these signal regions on genomic tracks [23, 29]. To integrate observation-level peaks into consensus peaks, which serve as key sampling points for sequencing reads and downstream analyses, we propose a simple tool. This tool, which we used in previous work, accepts per-experiment peak-calling results (from an external tool) as input and outputs consensus peaks based on the genome-wide peak density. While identifying consensus peaks based on simple overlaps is an acceptable approach when working with a small number of experiments, it becomes problematic with a larger number of experiments. In such cases, some regions of the genome can be almost entirely covered by peaks, making the high-resolution identification of consensus peaks impossible. To merge peaks into consensus peaks, we first compute the density function of the peak summits (the single base pair genomic location with the maximum signal of the peak, we use the centre position if not available) across each chromosome. To do this, we use a kernel density estimate, employing σ = Median peak size / 8 as the bandwidth. We then delineate consensus peaks at each local minimum of the density function. Abnormally small artifactual consensus peaks with a size smaller than the bandwidth are discarded, as well as irreproducible peaks (consensus peaks formed by only one experiment, can be modified). Finally, we restrict the boundaries of the newly defined consensus peak to those of the farthest peaks (Figure S1). When working with stranded peak data (for example, CAGE or RAMPAGE), we simply run these steps once for each strand.

Linking genomic regions to genes and functional annotations

Genes are much more studied and functionally annotated compared to regulatory elements of the genome. As regulatory elements are typically located near the gene they regulate, or as transcripts are frequently grouped into co-regulated clusters with similar functions, we propose tools to assist in functionally annotating regions of interest. We provide an utility to rename them according to their nearest gene, as well as a tool to perform Gene Set Enrichment Analysis based on genomic regions. We use a statistical framework that assumes the query regions are a subset of background regions (e.g., a set of differentially expressed/bound/accessible regions across two conditions is a subset of all the regions considered for differential testing). Our approach is similar to Chip-Enrich [26] and Poly-Enrich [13], which have shown that gene-wise modelling is required to reduce false discoveries. However, these two methods do not offer a model for the case where the query regions are a subset of a set of background regions, which is important as the background regions typically have a gene set bias. Following the conclusions of these two aforementioned methods, we do not use GREAT [17] as its two-tests approach saturates when querying too many regions, and as its binomial test alone is unreliable.

To assign query genomic regions to genes, we use the same heuristic as GREAT at default settings: a basal domain of 5 kb upstream and 1 kb downstream, extended in both directions up to 1 Mb or to the nearest basal domain (whichever is the closest). For each gene i, we obtain the number of genomic regions in its regulatory region, for all the background regions (n_i) and the subset of interest (k_i).

To compute Gene Set enrichments, we fit a Negative Binomial Generalized Linear Model per gene set s (e.g., the set of genes with the "B cell activation" GO term) , which predicts the expected number of genomic regions of the subset of interest $\mu_{i,s}$ in the regulatory region of a gene i : $ln(\mu_{i,s}) = \beta_{0,s} + \beta_{1,s} \times G_s + ln(E_i)$. Where G_s is equal to 1 if the studied gene belongs to the Gene Set s of interest and 0 otherwise. The term E_i multiplicatively corrects for the intersection bias of the background regions, with E_i being the expected number of hits for a particular gene: $E_i = n_i \times \frac{K}{N}$, with K being the number of query regions, and N the number of background regions. We test for each gene set s whether $\beta_{1,s}$ is greater than zero (i.e., whether genes in the gene set have more hits than those not in the gene set) using a Wald Test and apply the Benjamini-Hochberg FDR correction.

Additionally, to reduce redundancy between gene sets, gene sets can be clustered according to the similarities between their annotated genes. We use the leiden [25] algorithm on a kNN graph to cluster gene sets, which can be represented as a gene, gene set binary matrix of association.

Count modelling and transformation for between observation comparisons

Following previous work [2, 15], we assume that for an observation i (e.g., sample or cell), for a variable j (e.g., genomic region or genes), the observed number of counts $K_{i,j}$(reads or UMIs) can be modelled with a Negative Binomial (NB) distribution with mean $\mu_{i,j}$ and overdispersion α_j :

$$K_{i,j} \sim NB(\mu_{i,j}, \alpha_j)$$

This family of distribution is discrete, heteroscedastic, skewed, and with long tails which often cause issues with approaches that assume normality. To circumvent this issue, normalizing transformations have been developed, the most popular being the log1p transform (or logCPM), which typically appears under the form :

$$log_a(1 + c \times \frac{K_{i,j}}{s_i}),\ a\ and\ c\ being\ constants.$$

While this approach has been found to work well for large counts with few zeroes, it raises issues for small counts where it can distort the underlying distribution, especially between zeroes and small counts. Furthermore, the choice of the a and c constants is often arbitrary but does have a large effect on the balance between skewness and zero-distortion [24]. Recent work suggests using the residuals of a null (i.e constant expression)

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.11.570597; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

NB model that has its overdispersion parameter α_j being constrained to be a function of the mean [2, 7, 12]:

$$ln(\mu_{i,j}) = \beta_j X_i + ln(s_{i,j})$$

$$\alpha_j^{reg} = f(\mu_j)$$

With $s_{i,j}$ being the multiplicative normalization factor for variable j and observation i; α_j^{reg} the regressed model coefficients for variable j, X the design matrix containing variables to regress out (i.e. a single intercept if there is no unwanted variables to regress). Briefly, the idea behind this approach is to model what would be the expected count distribution of a variable according to its mean expression for a constant gene/feature, then compute how much each observed count $K_{i,j}$ is deviating from the expected fitted distribution $NB(\mu_{i,j}, \alpha_j^{reg})$ of the variable j.

In this work, we fit $f(\mu_{i,j})$ by randomly sampling up to 2000 variables (to speed up computations as there is no need to fit each variable to obtain a trendline), fit models using Maximum Likelihood Estimation with the Statsmodels [22] Nelder-Mead solver (without constraint on the overdispersion parameter). To obtain the regularized estimate of overdispersion we use a rolling median of the mean-sorted overdispersion values. Finally, we re-fit for each variable a new model with its regularised j using the Statsmodels IRLS solver, then compute residuals between the predicted distribution $NB(\mu_{i,j}, \alpha_j^{reg})$ and the observed counts $K_{i,j}$.

The popular SCTransform approach suggests using pearson residuals, which are a linear transformation of the observed counts :

$$r_{i,j}^p = \frac{K_{i,j} - \mu_{i,j}}{\sqrt{\mu_{i,j} + \alpha_j^{reg} \times \mu_{i,j}^2}}$$

Here, we use Anscombe residuals, as implemented in Statsmodels, which are a nonlinear transformation of the observed counts, and are asymptotically following a standard normal distribution for a properly specified model.

$$r_{i,j}^a = \frac{K_{i,j}^{2/3} \times H(-\alpha_j^{reg} \times K_{i,j}) - \mu_{i,j}^{2/3} \times H(-\alpha_j^{reg} \times \mu_{i,j})}{(\mu_{i,j} + \alpha_j^{reg} \times \mu_{i,j}^2)^{1/6}}$$

Where $H(x) = H_2F_1(2/3, 1/3, 5/3, x)$, H_2F_1 being the Gauss Hypergeometric function.

We found Pearson residuals to be much more sensitive to outliers, and as they are still skewed and heavy-tailed due to the linear nature of the transformation, their values have to be clipped to prevent large values from driving a majority of the variance. The clipping value is a sensitive hyperparameter and has to be adjusted per dataset, with too large values being sensitive to outliers, and too small values removing biological signal. Additionally, SCTransform uses an arbitrary lower bound on the variance estimate to prevent weakly expressed genes to drive a majority of the variance. We found that Anscombe residuals do not require clipping (Figure S2), give low weights to weakly expressed features (Figure S4 A), as well as having better known theoretical statistical properties concerning normality. In our package, we implement both methods and use Anscombe residuals as the default transform.

Between sample normalisation

We implement a few popular methods of normalisation on top of library size normalisation, such as DESeq2[15] median of ratios, Upper Quartile normalisation, and implement the scran pooling and deconvolution method [10] through a rpy2 wrapper to the scran library. The median of ratios approach is recommended for data with large counts, while the scran approach is better suited for small counts and a large number of observations. A generic method of normalisation that is independent of sample size and sequencing depth is still an area of open research. We note that our model for count distribution allows not only for a per observation multiplicative size factor but also for a per feature, per observation normalisation factor ($s_{i,j}$). This could be exploited to use normalisation strategies that aim to correct sequence-content bias in the model. Here, we use this property to extend the model to sequencing data with input counts.

Extending the model to sequencing data with input counts

Input sequencing is used to estimate sequence-content biases, such as PCR bias, sonication, mappability, or CNV alterations, and is used widely, for example in ChIP-seq. In these experiments, the log of the fold change (LFC), or the enrichment p-value against the input is typically used as metrics for the signal. However, the LFC does not take into account the significance of the enrichment, and on the other hand, the p-value is difficult to interpret nor displays the effect size of the enrichment; both measurements can be difficult to use for between sample comparisons. Most current methods simply ignore the amount of Input [6] and assume the sequence bias is the same across all conditions, which can be incorrect in case of differential chromatin accessibility or difference in the immunoprecipitation effiency. As the amount of input has a multiplicative effect on the observed counts, the null count model simply becomes :

$$ln(\mu_{i,j}) = \beta_j X_i + ln(Normalized\ Input_{i,j})$$

However, there is a need to normalize the input counts, and previous research has found that library size normalisation is insufficient, as the observed (ChIP) signal contains both true signal and noise reads. The signal-to-noise ratio has also been found to fluctuate strongly between experiments, mainly due to varying immunoprecipitation quality. We present a novel two-step approach, which consists of centering and then scaling the input counts. The centering step aims to find the zero of the log fold change, i.e., finds a multiplicative factor at the observation level which normalizes background regions with no signal between the ChIP and input. Here, we use the Signal Extraction Scaling [4] (SES) approach on counts sampled from random regions of the genome (10,000 per default) to estimate a centering factor c_i. While the centering step is sufficient to perform peak calling, another scaling step is required to normalize fold changes, as samples with a more successful immunoprecipitation will have larger fold changes. We compute the per observation, median centered, signal enrichment factor $SEFM_i$ as :

$$SEF_i = \frac{\sum_{j\ \epsilon\ variables} K_{i,j}}{\sum_{j\ \epsilon\ variables} c_i \times I_{i,j}}$$

$$SEFM_i = SEF_i / median(SEF)$$

I being the input count matrix (Note that here, we do not use the counts at random genomic regions to compute SEF_i, but

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.11.570597; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

the actual count matrices K and I).

Intuitively, this means that, for example, for an observation i with a two times stronger signal enrichment factor, with an observed fold change over input of 1.5 at feature j, its input should be scaled to obtain a fold change of 1.25. Formally, this translates into :

$$fc_{i,j} = \frac{K_{i,j}}{I_{i,j} \times c_i}$$

$$r_{i,j} = SEFM_i \times (fc_{i,j} - 1) + 1, \ fc_{i,j} >= 1$$

$$r_{i,j} = \frac{1}{SEFM_i \times (1/fc_{i,j} - 1) + 1}, \ fc_{i,j} < 1$$

$$Normalized\ Input_{i,j} = I_{i,j} \times c_i \times \frac{fc_{i,j}}{r_{i,j}}$$

Additionally, to avoid removing features having at least a single zero as input counts (as it causes a division by 0 or taking the log of zero in the model), we replace zeroes by the mean input count for this feature (taking into account the input library size). This avoids dropping an extremely large number of variables when dealing with a large number of observations while keeping a meaningful value for the input counts.

Note that we considered ChIP-seq as an example in this paragraph, but our framework should work with experimental designs that are using a control track with an expected multiplicative effect on the resulting counts. Figure S3 highlights the effect of the two steps of our input-based normalisation.

Differential expression

We implement DESeq2 with Log Fold Change (LFC) shrinkage [30] via an rpy2 wrapper to conduct differential expression between two conditions, as it uses a statistical model similar to ours, supporting a per feature, per observation normalisation factor. However, DESeq2 can be computationally intensive for large datasets. When working with more than 50 samples, which provides sufficient statistical power, we transition to a Welch's t-test performed on the NB residuals, which has been shown to work well for single-cell data [2]. It should be noted, though, that the t-test offers less statistical power than DESeq2.

Feature selection for downstream tasks

Feature selection in scRNA-seq (single-cell RNA sequencing) is a step that facilitates the elimination of a substantial portion of likely uninformative variables, i.e., those with very low expression that are mostly contaminated by technical noise, or those with ubiquitous expression, which do not provide insight into the sample/cell biology [12, 24]. Removing these can enhance the quality of downstream tasks, such as classification, clustering, or visualisation. Typically, around 2000 to 3000 genes are retained in scRNA-seq experiments, but this number is generally hand-tuned for each experiment. We use the sum of squared residuals (SSR) as a criterion for selecting variables, which serves as an indicator of the goodness of fit for the null NB model. Large residuals suggest a poor fit of the null model of constant expression, likely caused by differential expression between biological conditions that are not observed in the null model. In Figure S4 A, we show that the Anscombe residuals have a higher variance in sufficiently expressed, highly variable features.

Optimal number of components for principal component analysis

To automatically identify the optimal number of Principal Components, we utilize Horn's Parallel Permutation Analysis, which has proven to be one of the most effective methods to determine the number of components in factor analysis [1]. This approach involves generating row-wise permutations for each feature, computing PCA on these permuted datasets, and selecting the number of components at the threshold where the eigenvalues from the randomized dataset exceed those from the actual dataset. By default, we use the residuals as input to the PCA, and due to the computational cost of this approach, we only perform three permutations by default. This is justifiable, as the randomized eigenvalues are very stable on large matrices (Figure S4 D). To compute PCAs, we use the fast "randomized" solver from the Python sklearn library. By default, we scale PCA components to unit variance (whitening) to give an equal weight to each significant component.

Graph Clustering

We implement the Shared Nearest Neighbour (SNN) Graph Clustering approach to identify clusters. This approach is common in single-cell RNA sequencing (scRNA-seq) analyses to identify clusters of cells without a priori on the number of clusters. To scale to a large number of points to cluster, we use an Approximate Nearest Neighbour (ANN) method to build the NN graph (python library PyNNDescent [5]). This approach avoids the quadratic time complexity of building exact nearest neighbours, can use any metric and runs in an almost linear time complexity. By default, we use the PCA representation with pearson correlation as the metric to build the NN graph. In the SNN graph, vertices are weighted by the number of shared nearest neighbours between the two nodes. To identify communities in the SNN graph, we used the Leiden graph clustering algorithm [25] implemented in the Python Leidenalg library.

Data visualisation

For visualising the similarity among observations in datasets with a large number of features and observations, dimensionality reduction methods such as UMAP (Uniform Manifold Approximation and Projection) or t-SNE[16] (t-Distributed Stochastic Neighbor Embedding) have become standard, particularly for the analysis of single-cell data. Here, we provide a wrapper for UMAP, which is, by default, performed on the PCA representation. By default, Pearson correlation is used as the metric when there are more than 10 input dimensions; otherwise, the Euclidean distance is used.

Analysis of benchmark datasets

We manually retrieved matching peak calling BED files, ChIP BAM files, and Input BAM files from ENCODE [19] (see Supplementary Tables). We selected only those H3K4Me3 ChIP-seq experiments that were mapped onto hg38, and removed samples with any audit errors. In cases where multiple replicates of the input file were present, we retained only the most deeply sequenced one and used it across all biological replicates. We used a coarse annotation (the same as in the ENCODE data portal) instead of the detailed cell types as there is a large number of precise cell types with no replicates. The list of genes associated with Gene Ontology terms was retrieved from the g:ProfileR website[21].

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.11.570597; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

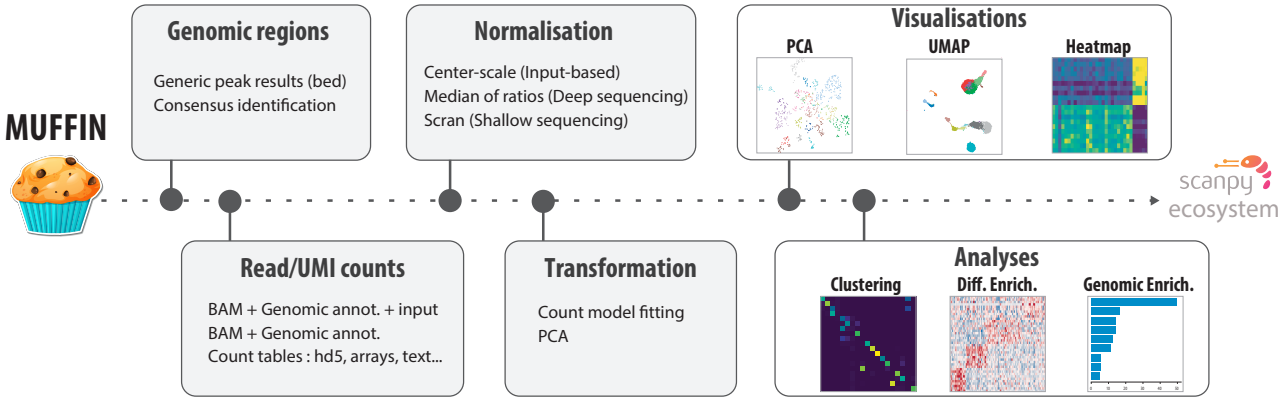


Fig. 1. Overview of the MUFFIN suite of tools.
MUFFIN covers multiple aspects of the analysis of several types of functional sequencing data, from the identification of consensus peaks, to differential expression and Gene Set Enrichment Analysis.

We obtained the filtered 10k PBMC scRNA-seq count tables from 10X Genomics (https://www.10xgenomics.com/resources/datasets/10k-human-pbmcs-3-ht-v3-1-chromium-x-3-1-high). We also analyzed the dataset using Seurat with SCTransform V2 as a reference, with the default settings from its tutorial vignette. We compared the similarity between our method and SCTransform using the jaccard index of similarities between clusters, and using Adjusted Mutual Information (AMI) to compare similarities across all clusters.

TCGA ATAC-seq[3] count tables were obtained from the NIH (https://gdc.cancer.gov/about-data/publications/ATACseq-AWG). The list of candidate genes associated with specific cancer hallmarks was obtained from the CHG database[28].

All analyses were conducted using the default settings, with the exception of the normalisation method: for the single-cell dataset, we employed scran normalisation; for the TCGA ATAC-seq dataset, we applied DESeq's median of ratios; and for the ChIP-seq dataset, we used our custom normalisation method. Furthermore, we utilised logistic regression via Scanpy to identify markers across multiple classes in both the single-cell and the ATAC datasets.

Results

MUFFIN Suite: applications and validation in diverse biological contexts

Here we introduce the comprehensive MUFFIN suite, encompassing a diverse array of tools tailored for analyses across a spectrum of experimental designs, accommodating varying observation counts and sequencing depths. The origin of this suite can be traced back to the creation and analysis of an RNA Polymerase II Atlas, involving the development and use of various tools in the context of 900 ChIP-seq experiments and approximately 28,000 RNA-seq samples [11]. The suite's functionalities include the analysis of sequencing signals, in the form of raw non-normalized read or UMI counted over a range of genomic features such as genes, or reference maps for regulatory elements like ENCODE CCREs or DNase hypersensitive regions. Additionally, the suite offers high-resolution consensus peak identification for large datasets, integrating observation-level peaks into consensus peaks—a crucial feature when dealing with a larger number of experiments. Tools are also provided for linking genomic regions to genes and functional annotations, including utilities for renaming regions based on their nearest gene and performing Gene Set Enrichment Analysis. Furthermore, the suite encompasses a variety of tools for functional sequencing analyses, covering count modeling for between-observation comparisons, differential expression analysis, feature selection for downstream tasks, determination of the optimal number of components for principal component analysis, graph clustering, and data visualization (1). A detailed description of the tools is available in the Methods. To validate the suite's efficacy, it was tested in three largely different datasets : i) the ENCODE Epigenetic Atlas of Immune Cells using H3K4Me3 ChIP-seq, ii) 10,000 PBMCs sequenced with 10X scRNA-seq, and iii) 400 cancer samples sequenced via ATAC-seq originating from TCGA. These applications showcase the versatility and robustness of the MUFFIN suite across diverse biological and technical contexts.

ENCODE epigenetic atlas of Immune cells through H3K4Me3 ChIP-seq

In order to evaluate the general applicability of our proposed methodology, we utilised the H3K4Me3 ChIP-seq dataset from the comprehensive ENCODE epigenetic atlas of immune cells. H3K4me3, an epigenetic modification known for marking active promoters, is a crucial mediator of transcriptional activity and plays a significant role in cellular differentiation and function. Here, multiple difficulties are linked to this dataset : there are no reference regions to sample sequencing signals from, the sample sizes are rather small, and the sequencing protocol uses a control track.

Our first task was to identify which regions to consider for quantifying H3K4Me3 occupancy. For this purpose, we employed our density-based approach to identify consensus peaks from per-experiment peak calling results. Our approach

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.11.570597; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

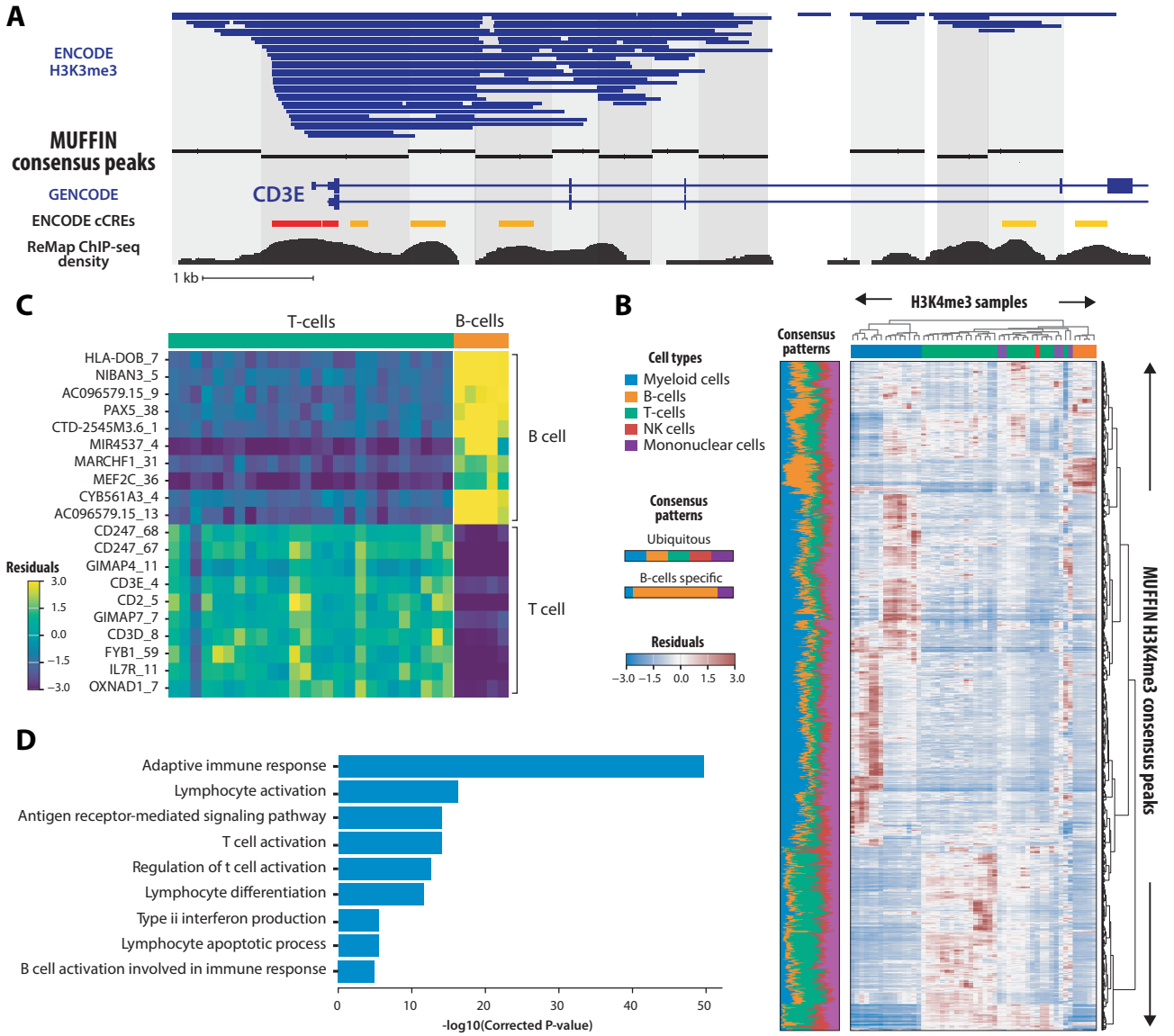


Fig. 2. Analysis of the ENCODE epigenetic atlas of Immune cells through H3K4Me3 ChIP-seq.
A. Result of the consensus peak identification in a 10kb region around the CD3E gene. Top to bottom rows indicates: ReMap Transcription Factor binding density, Consensus peaks, Peaks from all experiments aggregated, GENCODE V43 annotation. The alternating colors highlight the coverage of each consensus peak. **B.** Heatmap of the Anscombe residuals. Rows correspond to samples and columns to consensus peaks. Bottom panel indicates the weighted proportion of the total Sum of Squared Residuals carried by each class. **C.** Heatmap of the Anscombe residuals, for the top 10 most differentially marked consensus peaks (each renamed according to its nearest gene) in either B-Cells or T-Cells. **D.** Clustered GO terms enrichments of genes near differentially marked consensus peaks between B-Cells and T-Cells. Terms are clustered by gene similarity and only the term with the strongest enrichment is displayed.

identifies the consensus peaks at high resolution in dense regions, and correlates with the Transcription Factor binding density from the external ReMap database[8] (Figure 2A). Subsequently, we counted reads in immuno-precipitated and input sample pairs at consensus peak locations. Ultimately, the counts are normalised using our centering and scaling approach (Methods), transformed using Anscombe residuals, and can be processed through a visualisation or dimensionality reduction routine. As shown in Figure 2 B, a heatmap view shows a clear distinction between immune cell types and reveals distinct H3K4Me3 occupancy across different cell types.

Next, we sought to identify regions differentially marked by H3K4Me3 between B-cells (n=5) and T-cells (n=26). We found that the regions most differentially bound in B-cells are indeed located near B-cell marker genes such as HLA-DOB, which is expressed in Antigen Presenting Cells, or PAX5, a transcription factor involved in B-Cell differentiation. Conversely, for T-cells, we also retrieved T-cell markers like CD247 (T-cell surface glycoprotein CD3 zeta chain) or CD3E (CD3 Epsilon Subunit Of T-Cell Receptor Complex). Using our tool of Genomic Regions Enrichment Analysis, we find that these differentially occupied regions are located near genes annotated with immune and lymphocyte-related Gene Ontology terms (Figure 2 C-D). As a whole, this confirms that our methods are able to analyze ChIP-seq data and are likely

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.11.570597; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

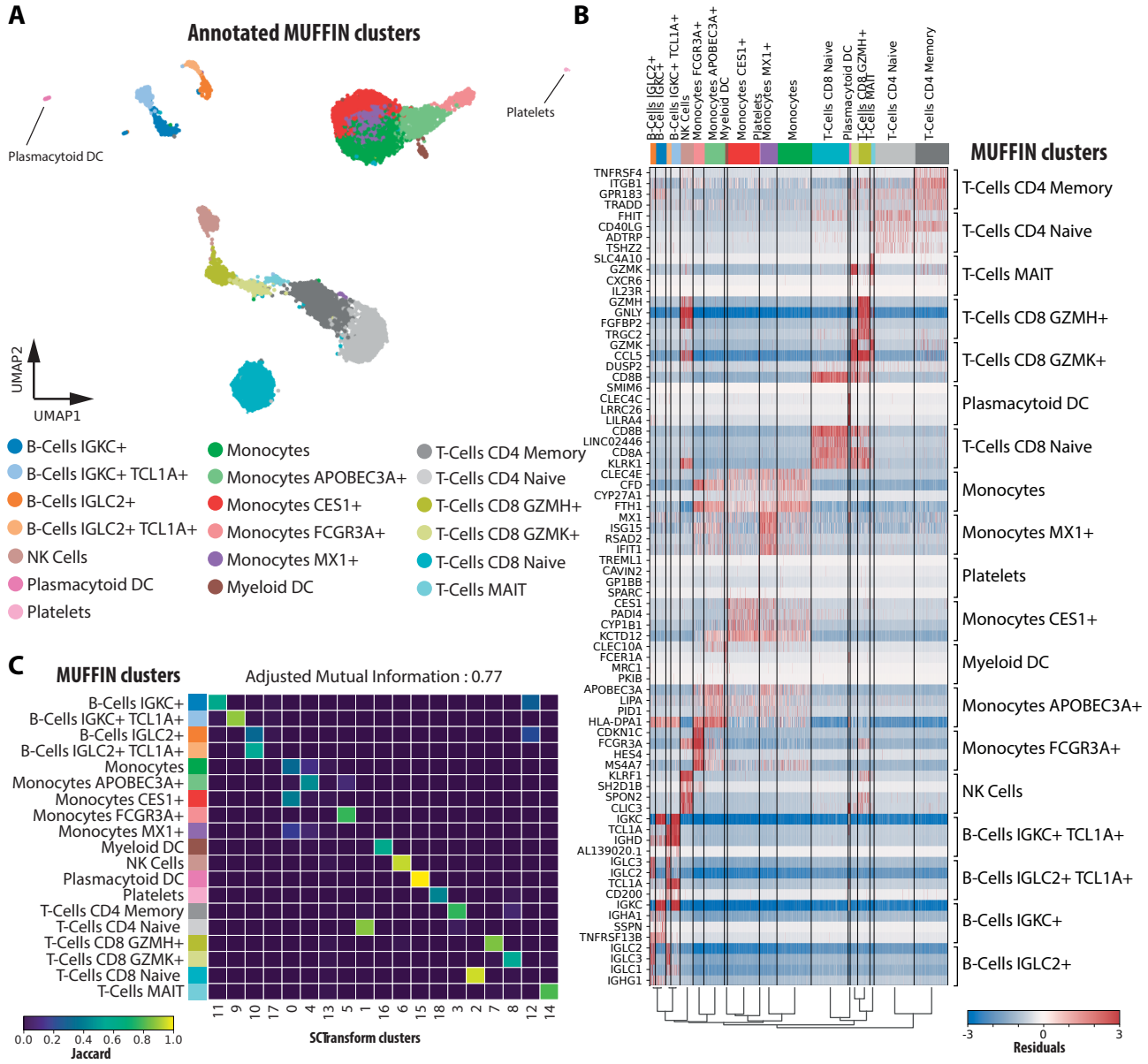


Fig. 3. Analysis of a 10k PBMCs single-cell RNA-seq dataset.
A. UMAP visualisation of the annotated clustered cells. **B.** Heatmap of the Anscombe residuals, per cell, for the 4 best markers of each cluster, identified via the coefficients of a multivariate, multi-class logistic regression. **C.** Heatmap of the Jaccard indexes between clusters identified via SCTransform and MUFFIN.

to work well with other similar datasets that are using control sequencing.

Cell-type clustering of 10,000 PBMCs sequenced by 10X scRNA-seq

After investigating the differential occupancy of H3K4Me3 across various immune cell types, we sought to evaluate the robustness and versatility of our methodology using another widely adopted dataset. Hence, we turned our focus to a single-cell RNA-sequencing (scRNA-seq) dataset, which offers a distinct set of challenges due to the large number of observations, inherent sparsity, shallow sequencing depth and high dimensionality of the data. The 10X Genomics scRNA-seq dataset of Peripheral Blood Mononuclear Cells (PBMCs), a mixed population of various immune cells, is often used as a benchmark case study. The PBMC dataset encompasses numerous immune cell types, allowing us to assess the efficacy of our approach in identifying and distinguishing between different cellular phenotypes.

Here, count quantification per gene, per cell has already been performed. We used UMAP to visualise the transcriptomic similarities between individual cells and applied graph clustering to pinpoint cell types (Figure 3 A). We clearly identify canonical cell types, such as Monocytes, B-Cells, T-Cells or platelets. Their markers, identified via the coefficients of a logistic regression, are also coherent (Figure 3 B), with canonical markers such as CD8B for CD8+ T-Cells, or GP1BB

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.11.570597; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

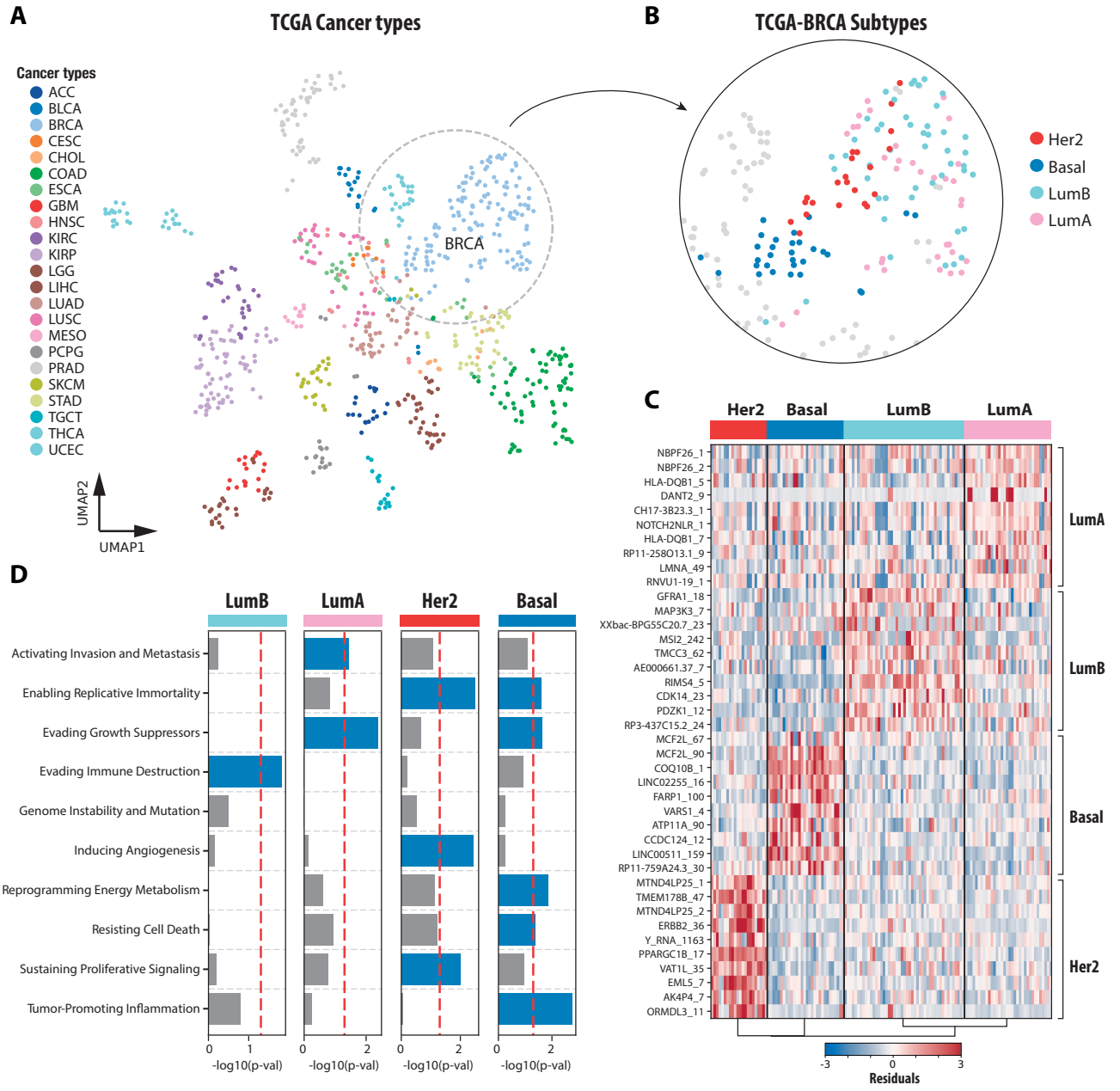


Fig. 4. Analysis of cancer samples sequenced with ATAC-seq
A. UMAP visualisation of each tissue sample, annotated by cancer type, as well as subtype for the breast cancer samples. **B.** Heatmap of the Anscombe residuals, per sample, for the 10 best markers of each breast cancer subtype, identified via the coefficients of a multivariate, multi-class logistic regression. **C.** Enrichment (-log10(p-value)) in cancer hallmarks for genes nearby the 5% most discriminative ATAC peaks for each breast cancer subtype, identified via the coefficients of a multivariate, multi-class logistic regression.

for platelets. Furthermore, we are able to identify cell types at a fine resolution, highlighting clusters of cells with similar transcriptomes, but with a few very specific markers, especially in the B-Cell clusters and the monocyte clusters. Overall, our methodology was able to provide clusters similar to those obtained with the reference SCTransform method (Figure 3 C). Together, these results show that our approaches are also well-suited for the analysis of datasets with low sequencing depth that are typically found in single-cell sequencing.

The Cancer Genome Atlas open chromatin landscape of cancers via ATAC-seq

Finally, we studied the open chromatin landscape in cancers using Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) in tissues. ATAC-seq is a technique used to assess the chromatin accessibility, thereby providing insights into the regulatory regions active in a particular cell type, tissue or condition. This sequencing protocol typically generates deeply sequenced datasets with an extremely large number of features.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.11.570597; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The read count table per ATAC peak, per sample was already provided, but a consensus peak identification as performed with the ChIP-seq dataset could have been performed. In Figure 4 A, using UMAP, we can clearly distinguish the different cancer types, and even distinguish between different breast cancer molecular subtypes. When focusing solely on breast cancer samples, we can identify markers for each subtype (Figure 4 B). While Luminal A and B subtypes do not exhibit strongly characteristic signatures, the Basal and Her2 subtypes display characteristic regions of open chromatin. Notably, in the Her2 subtype, an open chromatin region located near the main marker of this cancer, ERBB2 (or Her2) can be identified. Using our tool of Genomic Regions Enrichment Analysis, we observe that characteristic open chromatin regions for each breast cancer subtype locate near cancer hallmark genes (Figure 4 C). Notably, the Her2+ breast cancer subtype is known to have elevated levels of Vascular Endothelial Growth Factor (VEGF) gene expression, inducing angiogenesis [9, 20].

Discussion

In conclusion, our proposed framework, MUFFIN, represents a valuable and generic toolset for the processing and analysis of count data derived from a variety of high-throughput sequencing experiments. We have demonstrated its performance across diverse datasets, from epigenetic to transcriptomic, and from bulk to single-cell resolution. The underlying model is highly flexible and can adapt to a wide range of experimental designs, and relies on already well-tested, state-of-art statistical approaches.

One of MUFFIN's key strengths lies in its ability to address multiple facets of count-based genomic assays in a generic way with minimal hand-tuning : normalization, count transformation, feature selection, dimensionality reduction, visualization, differential expression, and clustering. Additionally, it offers specialized tools for analyzing data that do not rely on gene annotations. This includes a tool for identifying consensus peaks at a high resolution, and another one for performing functional enrichment of genes located near genomic regions of interest. Furthermore, MUFFIN seamlessly integrates in the existing Python Scanpy ecosystem and its diverse range of tools. The modular architecture of MUFFIN could also allow future packages to be built on top of it.

However, it's worth noting a potential limitation of MUFFIN, which is its compatibility with extremely large, sparse datasets. As its count transformation does not support sparse data formats, this may restrict its use in situations where memory capacity is a significant concern, such as in the analysis of extremely large integrative single-cell datasets consisting of millions of cells.

In summary, MUFFIN offers generic tools to analyze high-throughput sequencing count data and complements the existing tools available in Scanpy and the Python ecosystem. We anticipate that it will be of strong interest for bioinformaticians working with functional genomic data, at the tissue or single-cell level.

Competing interests

No competing interest is declared.

Author contributions statement

P.D.L. took the lead in developing the methods, coding, and analyzing the datasets. Both P.D.L. and B.B. contributed to the writing and reviewing of the manuscript. All authors edited and approved the article.

Acknowledgements

We would like to thank Dr. Lionel Spinelli for engaging in discussions that have contributed positively to our work.

This work was supported with ; PhD Fellowship to P.D.L. from the French Ministry of Higher Education and Research (MESR); Institut National de la Santé et de la Recherche Médicale (INSERM); The Core Cluster of the Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013) for granting access to its high performance computing resources. The results shown here are based upon data generated by the TCGA Research Network, the ENCODE Consortium and the ENCODE production laboratories.

References

1. Buja, A. & Eyuboglu, N. Remarks on Parallel Analysis. *Multivariate Behavioral Research* **27.** Publisher: Routledge _eprint: https://doi.org/10.1207/s15327906mbr2704_2, 509–540. ISSN: 0027-3171 (Oct. 1992).
2. Choudhary, S. & Satija, R. Comparison and evaluation of statistical error models for scRNA-seq. en. *Genome Biology* **23.** Number: 1 Publisher: BioMed Central, 1–20. ISSN: 1474-760X (Dec. 2022).
3. Corces, M. R., Granja, J. M., Shams, S., *et al.* The chromatin accessibility landscape of primary human cancers. eng. *Science (New York, N.Y.)* **362,** eaav1898. ISSN: 1095-9203 (Oct. 2018).
4. Diaz, A., Park, K., Lim, D. A., *et al.* Normalization, bias correction, and peak calling for ChIP-seq. *Statistical Applications in Genetics and Molecular Biology* **11,** 10.1515/1544–6115.1750 /j/sagmb.2012.11.issue–3/1544–6115.1750/1544–6115.1750.xml. ISSN: 1544-6115 (Mar. 2012).
5. Dong, W., Moses, C. & Li, K. *Efficient k-nearest neighbor graph construction for generic similarity measures* in *Proceedings of the 20th international conference on World wide web* (Association for Computing Machinery, New York, NY, USA, Mar. 2011), 577–586. ISBN: 978-1-4503-0632-4.
6. Eder, T. & Grebien, F. Comprehensive assessment of differential ChIP-seq tools guides optimal algorithm selection. en. *Genome Biology* **23.** Number: 1 Publisher: BioMed Central, 1–27. ISSN: 1474-760X (Dec. 2022).
7. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. en. *Genome Biology* **20.** Number: 1 Publisher: BioMed Central, 1–15. ISSN: 1474-760X (Dec. 2019).
8. Hammal, F., de Langen, P., Bergon, A., *et al.* ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.11.570597; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Research* **50,** D316–D325. ISSN: 0305-1048 (Jan. 2022).

9. Kumar, R. & Yarmand-Bagheri, R. The role of HER2 in angiogenesis. eng. *Seminars in Oncology* **28,** 27–32. ISSN: 0093-7754 (Oct. 2001).
10. L. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology* **17,** 75. ISSN: 1474-760X (Apr. 2016).
11. Langen, P. d., Hammal, F., Guéret, E., *et al.* Characterizing intergenic transcription at RNA polymerase II binding sites in normal and cancer tissues. English. *Cell Genomics* **3.** Publisher: Elsevier. ISSN: 2666-979X (Oct. 2023).
12. Lause, J., Berens, P. & Kobak, D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. en. *Genome Biology* **22.** Number: 1 Publisher: BioMed Central, 1–20. ISSN: 1474-760X (Dec. 2021).
13. Lee, C. T., Cavalcante, R. G., Lee, C., *et al.* Poly-Enrich: count-based methods for gene set enrichment testing with genomic regions. *NAR Genomics and Bioinformatics* **2,** lqaa006. ISSN: 2631-9268 (Mar. 2020).
14. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930. ISSN: 1367-4803 (Apr. 2014).
15. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. eng. *Genome Biology* **15,** 550. ISSN: 1474-760X (2014).
16. Maaten, L. v. d. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9,** 2579–2605. ISSN: 1533-7928 (2008).
17. McLean, C. Y., Bristor, D., Hiller, M., *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28,** 495–501. ISSN: 1087-0156 (May 2010).
18. Meuleman, W., Muratov, A., Rynes, E., *et al.* Index and biological spectrum of human DNase I hypersensitive sites. en. *Nature* **584.** Number: 7820 Publisher: Nature Publishing Group, 244–251. ISSN: 1476-4687 (Aug. 2020).
19. Moore, J. E., Purcaro, M. J., Pratt, H. E., *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. en. *Nature* **583.** Number: 7818 Publisher: Nature Publishing Group, 699–710. ISSN: 1476-4687 (July 2020).
20. Nasir, A., Holzer, T. R., Chen, M., *et al.* Differential expression of VEGFR2 protein in HER2 positive primary human breast cancer: potential relevance to anti-angiogenic therapies. *Cancer Cell International* **17,** 56. ISSN: 1475-2867 (May 2017).
21. Raudvere, U., Kolberg, L., Kuzmin, I., *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* **47,** W191–W198. ISSN: 0305-1048 (July 2019).
22. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference.* Conference Name: Proceedings of the 9th Python in Science Conference, 92–96 (2010).
23. Tarbell, E. D. & Liu, T. HMMRATAC: a Hidden Markov ModeleR for ATAC-seq. *Nucleic Acids Research* **47,** e91. ISSN: 0305-1048 (Sept. 2019).
24. Townes, F. W., Hicks, S. C., Aryee, M. J., *et al.* Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. en. *Genome Biology* **20.** Number: 1 Publisher: BioMed Central, 1–16. ISSN: 1474-760X (Dec. 2019).
25. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. en. *Scientific Reports* **9.** Number: 1 Publisher: Nature Publishing Group, 5233. ISSN: 2045-2322 (Mar. 2019).
26. Welch, R. P., Lee, C., Imbriano, P. M., *et al.* ChIP-Enrich: gene set enrichment testing for ChIP-seq data. *Nucleic Acids Research* **42,** e105. ISSN: 0305-1048 (Sept. 2014).
27. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. en. *Genome Biology* **19.** Number: 1 Publisher: BioMed Central, 1–5. ISSN: 1474-760X (Dec. 2018).
28. Zhang, D., Huo, D., Xie, H., *et al.* CHG: A Systematically Integrated Database of Cancer Hallmark Genes. *Frontiers in Genetics* **11.** ISSN: 1664-8021 (2020).
29. Zhang, Y., Liu, T., Meyer, C. A., *et al.* Model-based Analysis of ChIP-Seq (MACS). en. *Genome Biology* **9.** Number: 9 Publisher: BioMed Central, 1–9. ISSN: 1474-760X (Nov. 2008).
30. Zhu, A., Ibrahim, J. G. & Love, M. I. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* **35,** 2084–2092. ISSN: 1367-4803 (June 2019).

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.11.570597; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supplementary Figures

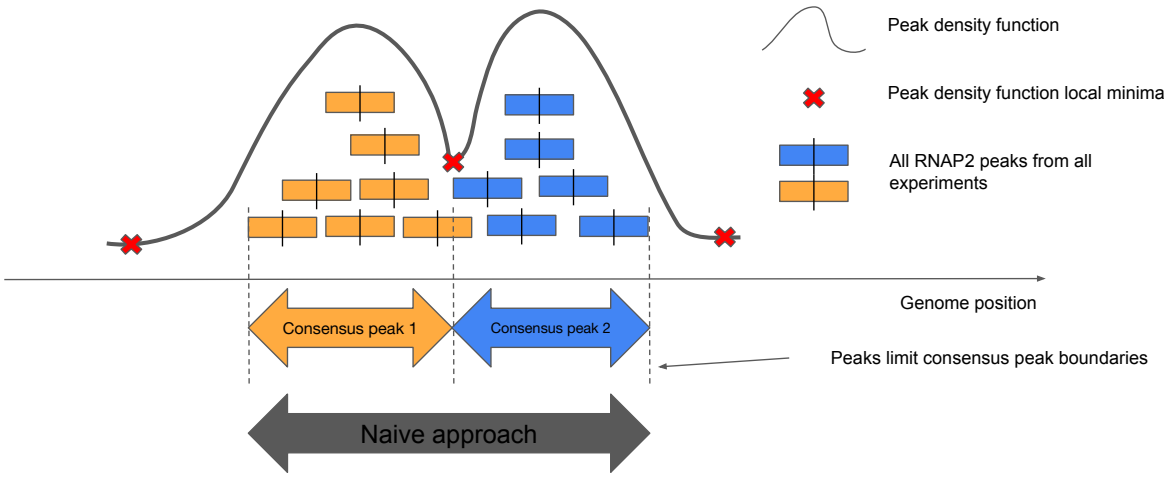
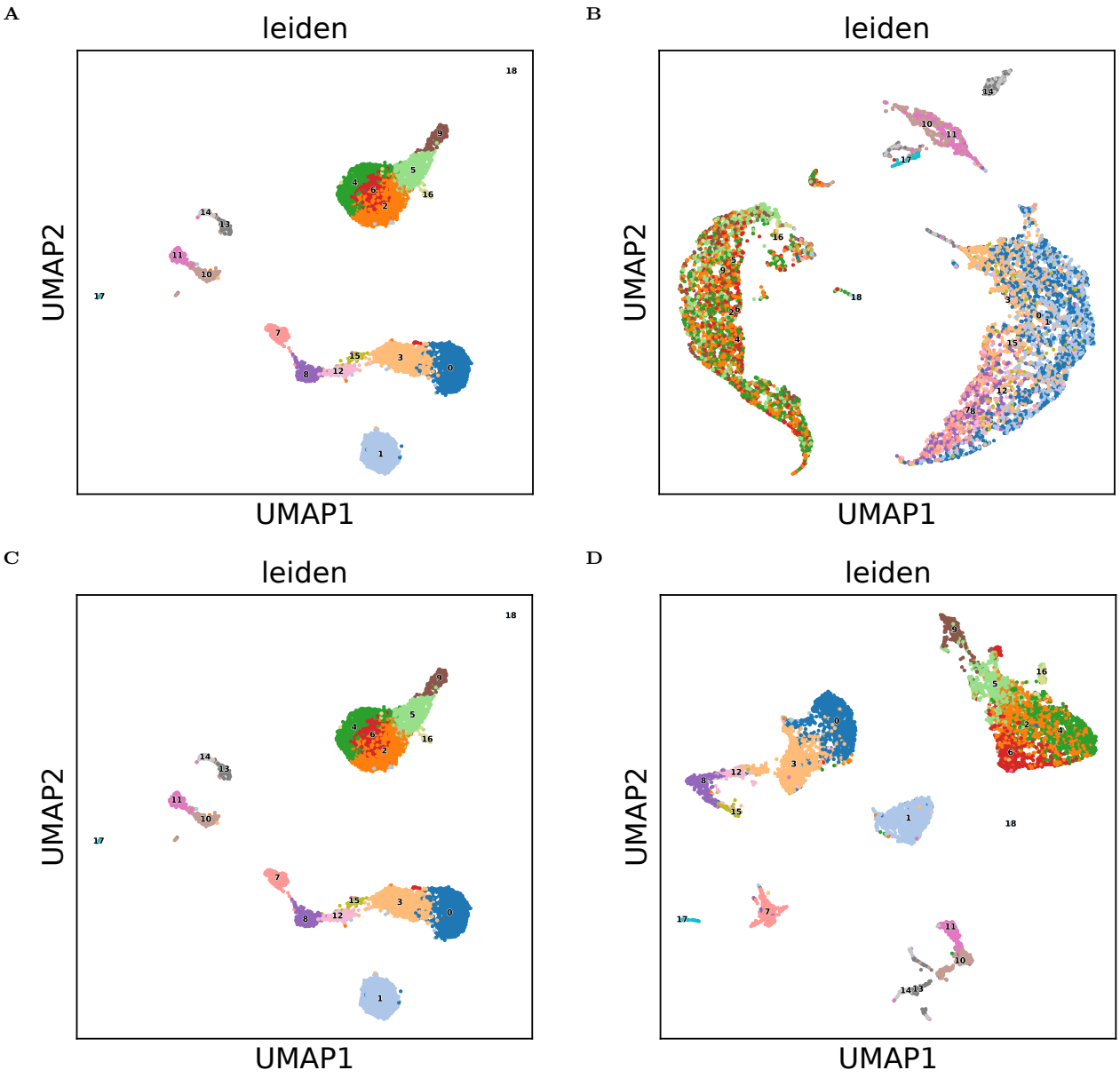


Fig. S1. Identification of consensus peaks
Identification of consensus peaks via a peak density-based approach. Naive approach refers to a simple merge on overlap.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.11.570597; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2023.12.11.570597; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

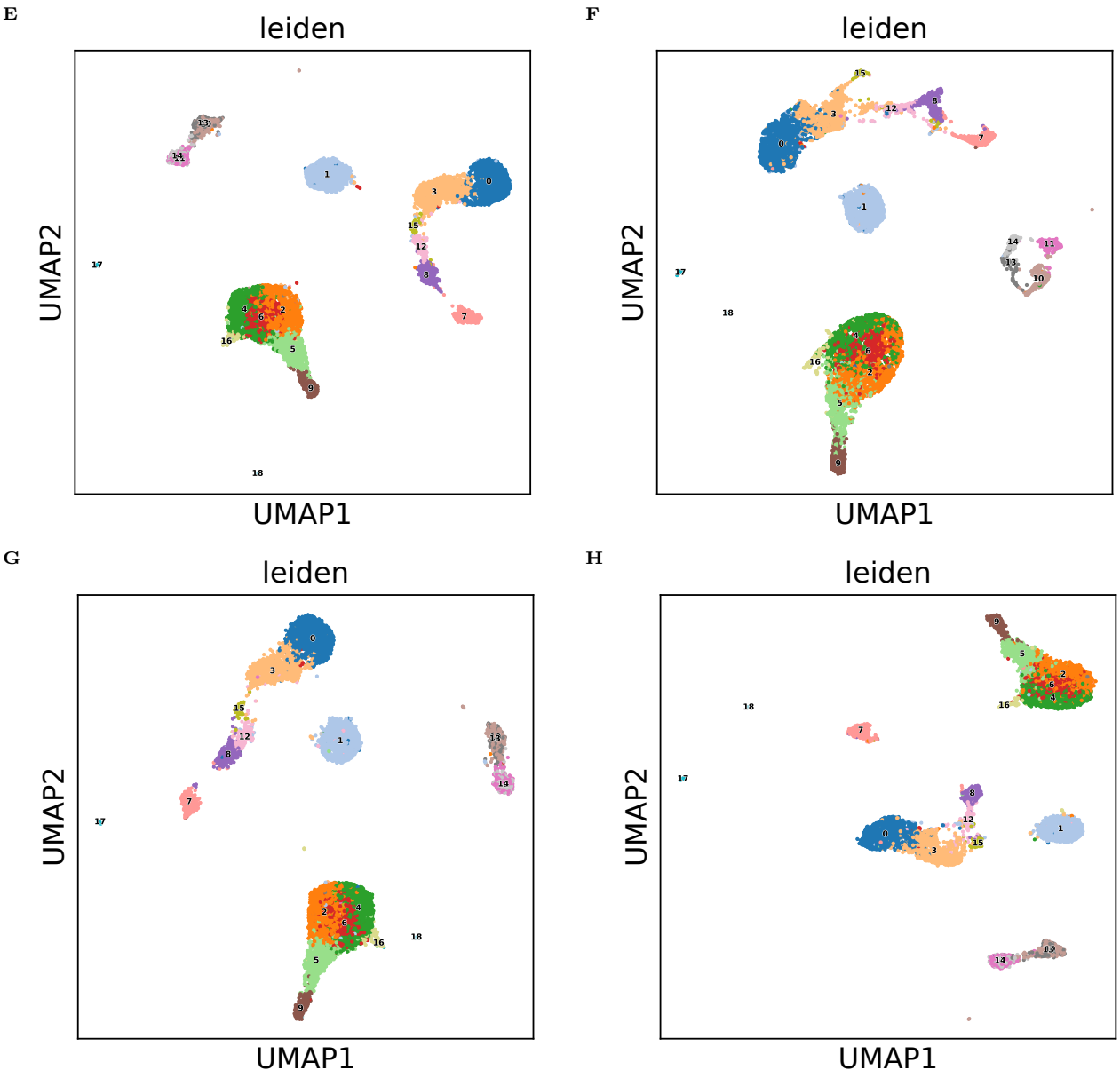


Fig. S2. Anscombe Residuals do not require value clipping
UMAP visualization, using default MUFFIN cluster labels of the PCA representation obtained after : **A.** Anscombe residuals transform, no value clipping. **B.** Pearson Residuals transform, no value clipping. **C.** Anscombe residuals transform, clipped to +-sqrt(n) (n being the number of observations/cells, SCTransform V1 default). **D.** Pearson Residuals, clipped to +-sqrt(n). **E.** Anscombe residuals transform, clipped to +-sqrt(n/30) (SCTransform V2 default). **F.** Pearson Residuals, clipped to +-sqrt(n/30). **G.** Anscombe residuals transform, clipped to +-3. **H.** Pearson Residuals, clipped to +-3. Large clipping values can produce incoherent clusters with pearson residuals and low values can eliminate biological signal with both methods, especially in the B-Cell clusters (10,11,13,14).

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.11.570597; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

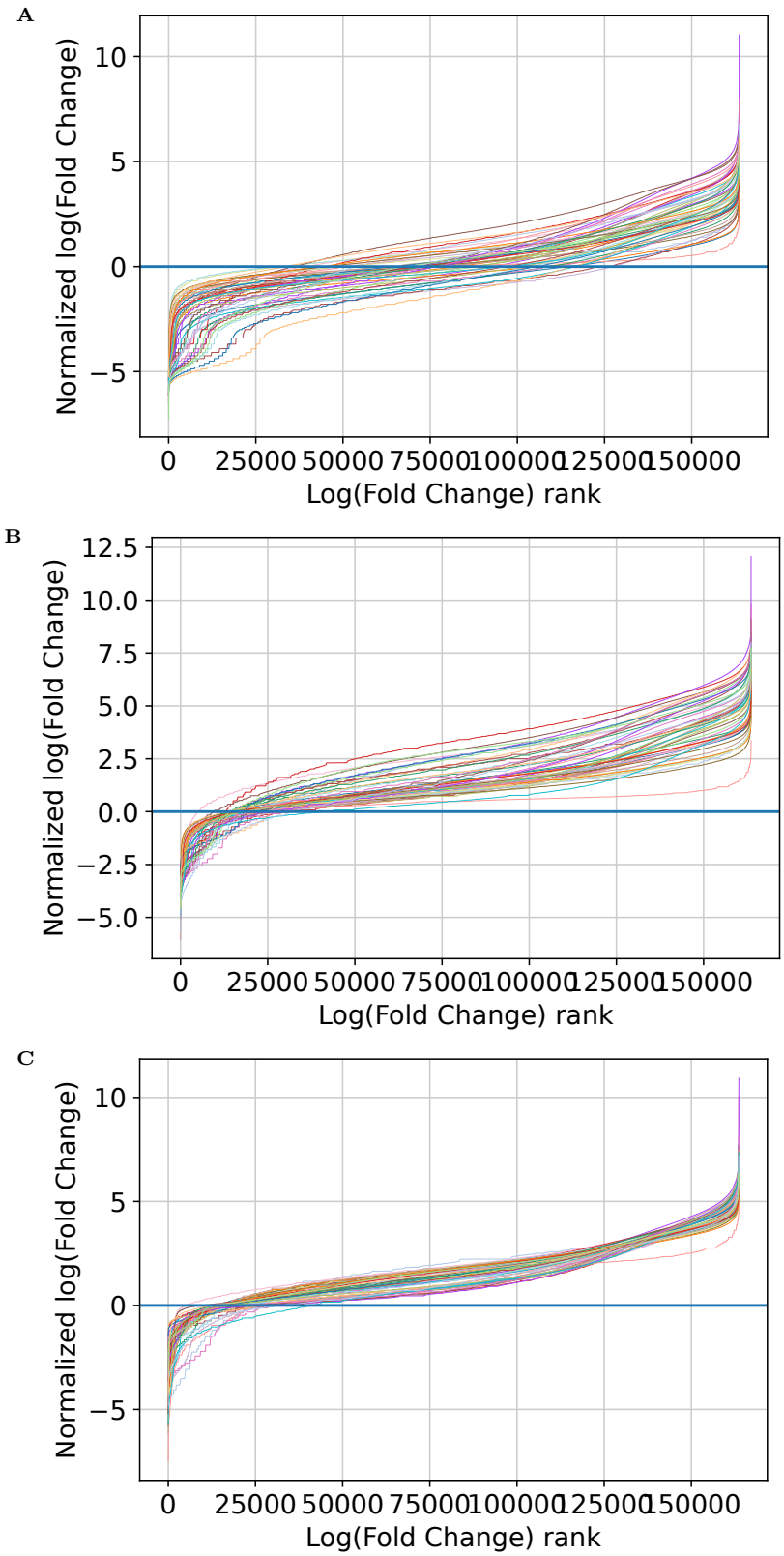


Fig. S3. Center-and-scale normalization for input-based sequencing data
A. Ranked, non normalized log fold change over input curves. Each curve correspond to one experiment in the H3K4Me3 ENCODE Immune cell atlas. **B.** Ranked, centered log fold change over input curves. **C.** Ranked, centered and scaled log fold change over input curves.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.11.570597; this version posted December 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

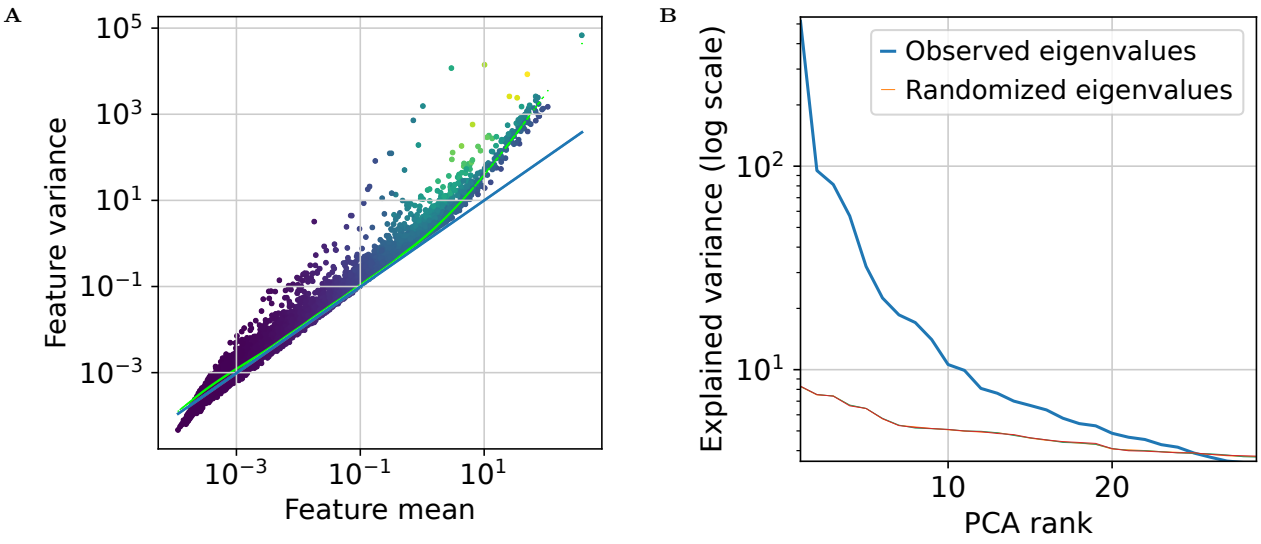


Fig. S4. Feature weighting of Anscombe residuals, Permutation PCA.
In the 10k PBMCs single-cell dataset : **A.** Mean-variance relationship of gene-expression (scran-normalized UMIs). Blue line indicates a Poisson mean-variance relationship, Green line indicates fitted mean-variance relationship. The color of each dot corresponds to its standard deviation (Yellow = larger). A lower standard deviation will result in less weight being given to the feature in downstream tasks. **B.** Observed and expected variance in each PCA component. Here 3 permutation steps have been performed (stacked Randomized eigenvalues curves).

Discussion et perspectives

Un atlas des régions intergénique potentiellement transcrites chez l'humain

Notre travail sur la transcription non-codante a permis d'établir un catalogue nouveau de régions potentiellement transcrites par l'ARN Polymérase II dans les régions intergéniques du génome. L'intérêt principal de notre approche par rapport aux travaux existants est l'utilisation d'une méthode indirecte ciblant l'enzyme responsable de la transcription, permettant de s'affranchir de la faible stabilité des transcrits non-codants. Bien que des protocoles permettent une identification sensible de ces transcrits non-codants[44], très peu de données sont disponibles pour ceux-ci, tandis qu'une grande quantité de données ChIP-seq ciblant la Pol II est disponible.

Nous avons mis en évidence que ces régions potentiellement transcrites possédaient en majorité des signatures épigénétiques enhancer, et étaient précisément régulées en étant fixées par la Pol II dans un ensemble de lignées cellulaires et tissus propre à chacune. A l'aide d'une approche de clustering, nous avons groupé et donc annoté ces régions en fonction des conditions biologiques dans lesquelles elles sont actives. Ce travail a permis d'établir une carte annotée de la fixation par la Pol II du génome intergénique, qui pourrait être utilisée dans de futurs travaux afin d'aider à la caractérisation d'éléments régulateurs ou de variants impliqués dans des maladies humaines.

A l'aide de ce catalogue de régions potentiellement transcrites, nous avons analysé plus 28 000 expériences RNA-seq, et validé que nos régions étaient en effet préférentiellement transcrites, mais également que malgré un signal faible, il était possible de discerner les différentes conditions biologiques en se basant uniquement sur la transcription intergénique située sur notre atlas. Cette analyse à également indiqué que l'expression tissu-spécifique corrélait avec l'occupation tissu spécifique observée dans le jeu de données Pol II, malgré des protocoles et un matériel biologique différent. Dans de prochains travaux, mon atlas pourrait être intégré aux analyses RNA-seq, qui habituellement ne ciblent que les gènes. Intégrer mon atlas permettrait de découvrir sans coût additionnel des éléments régulateurs différentiellement exprimés entre deux conditions, et identifier de nouveaux marqueurs dans des maladies autres que le cancer. Un travail de ré-analyse uniforme des données RNA-seq publiques, utilisant l'atlas Pol II est une piste de

travaux futurs.

En appliquant cette approche sur les données RNA-seq de cancers du TCGA, j'ai pu mettre en évidence qu'il était possible discriminer les différents cancers en utilisant uniquement cette transcription intergénique, et que cette observation est également valide pour les sous-types de cancers. J'ai pu ainsi identifier pour chaque cancer et par la suite à travers plusieurs cancers, des régions du génome dont la transcription est caractéristique de l'état tumoral du tissu, mais également de la survie du patient et donc de la sévérité de la tumeur. Ces régions pourraient potentiellement servir de nouveaux marqueurs ou de cibles thérapeutiques, et améliorer la compréhension que l'on a des cancers pour à terme, améliorer la prise en charge des patients. Il pourrait notamment être intéressant de valider expérimentalement ces régions par KO, puis observer des variations dans le phénotype de lignées tumorales. Une grande partie de ces régions semble être associée au cycle cellulaire, il serait donc intéressant d'étudier si le KO de ces régions à un effet sur la prolifération de cellules tumorales.

D'autre part, au cours de mes analyses, j'ai pu constater qu'une grande quantité de pics consensus Pol II se situaient en fin de gène, et chevauchaient des éléments régulateurs enrichis en motifs CTCF, qui est l'une des protéines responsable de la conformation de la chromatine. Ainsi, il serait intéressant d'étudier ces régions en particulier pour identifier quels mécanismes biologiques de régulation sont en jeu ici. Une des pistes d'étude possibles serait d'analyser des jeux de données Hi-C pour étudier la conformation de la chromatine en fin de gène, et identifier à quelles autres régions les fins de gènes sont spatialement proches ; ainsi que d'étudier les profils STARR-seq afin d'observer une activité potentiellement enhancer de ces régions.

De même, certains pics consensus Pol II (5%) ne chevauchent aucune annotation de référence du génome (GENCODE, CAGE par FANTOM5, cCREs d'ENCODE, lncipedia, ReMap). Ceux-ci pourraient correspondre à de nouveaux éléments régulateurs ou potentiellement être de faux positifs. Ainsi, il serait possible d'étudier si ceux-ci sont fortement reproductibles (c'est à dire présents dans un grand nombre d'expériences de l'atlas Pol II), ainsi que leur caractéristiques, telles que leur localisation génomique ou leur séquence.

De plus, j'ai remarqué que la Pol II semblait avoir une forte affinité pour certains éléments répétés du génome. Cette analyse est limitée d'une part par l'utilisation de données à read courts, rendant l'alignement non-ambigu des reads sur les séquences répétées difficile ; et par l'utilisation de la version du génome humain hg38, incomplète et réalisée avec des techniques de séquençages permettant difficilement l'assemblage des régions répétées du génome. Une étude de données transcriptomiques à read long sur le génome T2T, permettrait d'étudier plus finement et avec moins de biais la transcription non-codante au niveau des éléments répétés, en réduisant l'ambiguïté de mapping des reads. Nous avions également fait le choix d'éliminer les reads multi-mapping, il pourrait également être intéressant d'essayer

d'inclure ces reads dans les analyses quantifiant le signal au niveau des éléments répétés.

De nouvelles méthodes pour l'analyse du séquençage fonctionnel

Les méthodes que j'ai développé au cours de ce travail sont généralisables à d'autres jeux de données, et pour faciliter leur réutilisation, j'ai créé un package python, MUFFIN, contenant ma méthode d'identification de pics consensus, mon outil d'enrichissement fonctionnel pour régions génomiques non-géniques, et mes outils d'analyse de tables de comptages. Afin de maximiser sa facilité de d'utilisation, MUFFIN est disponible via conda, afin d'une part faciliter le versionnage pour la reproductibilité, et d'autre part résoudre aisément les conflits de dépendances. MUFFIN s'inscrit également dans l'écosystème de Scanpy et de son format de données standardisé AnnData, spécialement conçu pour le traitement de tables de comptages de génomique fonctionnelle. Cela permet d'insérer aisément MUFFIN dans un pipeline d'analyse de données de génomique fonctionnelle utilisant les outils de Scanpy ou de son écosystème.

Dans un premier temps, ma méthode d'identification de pics consensus pourrait être utilisée pour de futures analyses intégratives générant des données génomiques sous formes de pics, telles que le ChIP-seq, le CAGE ou le STARR-seq. Elle a notamment déjà été intégrée au pipeline de traitement des données de ReMap, afin d'identifier des modules de régulation parmi les 170 millions de pics ChIP-seq chez l'humain, mais également pour Mus Musculus, Drosophila Melanogaster et Arabidopsis Thaliana.

Relier des éléments régulateurs à des fonctions biologiques est essentiel pour l'interprétation de leur fonction. Cependant, le plus souvent, seuls les gènes possèdent des annotations standardisées. MUFFIN permet d'effectuer des enrichissement sur sets de gènes à partir de régions génomique en associant celles-ci aux gènes, et ce de façon plus statistiquement robuste par rapport aux méthodes précédentes. Cet facette de MUFFIN pourrait être utilisée dans de futurs travaux cherchant à caractériser les fonctions biologiques d'ensembles de régions génomiques, pouvant être par exemple générées par des protocoles tels que l'ATAC-seq ou le CUT and TAG, en bulk ou par single cell.

Enfin, MUFFIN contient des outils pour l'analyse de tables de comptages issues de séquençage fonctionnel que j'avais initialement développé pour l'analyse du RNA-seq dans l'intergénique, possédant un signal faible. Bien qu'issues en partie du monde du single cell, j'ai montré que ces méthodes étaient très flexibles et pouvaient être utilisées une grande variété de protocoles de séquençage, notamment l'ATAC-seq, le ChIP-seq et le scRNA-seq. Par de légers changements de méthodologie, j'ai rendu

ces méthodes moins sensibles au paramètres spécifiés par l'utilisateur, pouvant parfois grandement changer les conclusions d'une analyse. Ces méthodes étaient également uniquement disponibles dans l'écosystème R, mais pas dans celui python de Scanpy, ni adaptées au traitement de larges comptes. Cet aspect de MUFFIN permet d'obtenir un traitement plus générique des données issues de protocoles séquençage fonctionnel et pouvant être résumées à travers une table de comptage, ainsi que d'ouvrir les outils de Scanpy à une plus grande variété de protocoles de séquençage.

Perspectives sur l'utilisation des données génomiques

Mon travail a mis en évidence que l'énorme et grandissante quantité de données présente dans les entrepôts de données n'est pas pleinement exploitée et peut toujours permettre de nouvelles avancées. Outre un travail d'annotation du génome, ces données pourraient nous aider à mieux comprendre et modéliser les processus de régulation de la transcription, et ce à différents niveaux : non seulement au niveau de l'initiation, mais également de la pause proximale, de l'épissage ou encore de la terminaison.

L'utilisation de modèles de Deep Learning inspirés des LLMs (Large Language Models) tels que Basenji [93] ou ENFORMER [10], appliqués à la prédiction du signal de près de 7,000 expériences de séquençage fonctionnel (comprenant entre autres ChIP-seq d'histone, CAGE, DNase-seq) provenant d'une grande partie d'ENCODE à partir de la séquence seule, permet déjà d'inférer in silico le signal d'initiation de la transcription obtenu par CAGE avec une bonne précision, et d'interpréter quelles séquences expliquent potentiellement cette transcription. Ces modèles pourraient potentiellement être appliqués pour bien d'autres questions biologiques, dont par exemple prédire quelles séquences ont une forte affinité pour la Pol II ou influencent son état de phosphorylation, en entraînant un nouveau modèle, ou en procédant à un fine-tuning [113] de modèles existants. Ces modèles possèdent toutefois des limitations, notamment au niveau de la taille de séquence pouvant être modélisée. Les interactions entre éléments régulateurs pouvant se produire jusqu'à un million de paires de bases ou plus, les interactions longue distance restent difficiles à intégrer dans ces modèles, car augmentant la taille du modèle. Afin d'améliorer la précision de ces modèles complexes en augmentant leur nombre de paramètres, une grande quantité de données est également nécessaire, tout comme une très grande puissance de calcul pour entraîner ces modèles.

Bibliographie

1. ADELMAN, K. & LIS, J. T. Promoter-proximal pausing of RNA polymerase II : emerging roles in metazoans. en. *Nature Reviews Genetics* **13,** 720-731. ISSN : 1471-0056, 1471-0064 (oct. 2012) (cf. p. 25).
2. AKIRAV, E. M., RUDDLE, N. H. & HEROLD, K. C. The role of AIRE in human autoimmune disease. en. *Nature Reviews Endocrinology* **7.** Number : 1 Publisher : Nature Publishing Group, 25-33. ISSN : 1759-5037 (jan. 2011) (cf. p. 42).
3. ALIPOUR, E. & MARKO, J. F. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Research* **40,** 11202-11212. ISSN : 0305-1048 (déc. 2012) (cf. p. 29).
4. ANDERSON, S. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Research* **9,** 3015-3027. ISSN : 0305-1048 (juill. 1981) (cf. p. 45).
5. ANDERSSON, R., GEBHARD, C., MIGUEL-ESCALADA, I. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455-461. ISSN : 0028-0836 (mars 2014) (cf. p. 21, 31, 63).
6. ANDERSSON, R. & SANDELIN, A. Determinants of enhancer and promoter activities of regulatory elements. en. *Nature Reviews Genetics* **21,** 71-87. ISSN : 1471-0064 (fév. 2020) (cf. p. 38).
7. ARNOLD, C. D., GERLACH, D., SPIES, D. *et al.* Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. en. *Nature Genetics* **46.** Number : 7 Publisher : Nature Publishing Group, 685-692. ISSN : 1546-1718 (juill. 2014) (cf. p. 74).
8. ARNOLD, C. D., GERLACH, D., STELZER, C. *et al.* Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* **339.** Publisher : American Association for the Advancement of Science, 1074-1077 (mars 2013) (cf. p. 74).
9. *ATAC-seq Data Standards and Processing Pipeline – ENCODE* (cf. p. 68).
10. AVSEC, Ž., AGARWAL, V., VISENTIN, D. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. en. *Nature Methods* **18,** 1196-1203. ISSN : 1548-7105 (oct. 2021) (cf. p. 23, 182).
11. BANKEVICH, A., NURK, S., ANTIPOV, D. *et al.* SPAdes : A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* **19,** 455-477. ISSN : 1066-5277 (mai 2012) (cf. p. 50).

12. BANNISTER, A. J. & KOUZARIDES, T. Regulation of chromatin by histone modifications. en. *Cell Research* **21,** 381-395. ISSN : 1748-7838 (mars 2011) (cf. p. 38).

13. BAO, W., KOJIMA, K. K. & KOHANY, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6,** 11. ISSN : 1759-8753 (juin 2015) (cf. p. 36).

14. BARANELLO, L., KOUZINE, F., SANFORD, S. *et al.* ChIP bias as a function of cross-linking time. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* **24,** 175-181. ISSN : 0967-3849 (mai 2016) (cf. p. 73).

15. BARSKI, A., CUDDAPAH, S., CUI, K. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129,** 823-837. ISSN : 0092-8674 (mai 2007) (cf. p. 70).

16. BARTKOWIAK, B. & GREENLEAF, A. L. Phosphorylation of RNAPII. *Transcription* **2,** 115-119. ISSN : 2154-1264 (mai 2011) (cf. p. 27).

17. BEAGAN, J. A. & PHILLIPS-CREMINS, J. E. On the existence and functionality of topologically associating domains. en. *Nature Genetics* **52,** 8-16. ISSN : 1546-1718 (jan. 2020) (cf. p. 32).

18. BELL, S. P. & DUTTA, A. DNA replication in eukaryotic cells. eng. *Annual Review of Biochemistry* **71,** 333-374. ISSN : 0066-4154 (2002) (cf. p. 19).

19. BENTLEY, D. R., BALASUBRAMANIAN, S., SWERDLOW, H. P. *et al.* Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry. *Nature* **456,** 53-59. ISSN : 0028-0836 (nov. 2008) (cf. p. 47).

20. BERNSTEIN, B. E., STAMATOYANNOPOULOS, J. A., COSTELLO, J. F. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology* **28,** 1045-1048. ISSN : 1087-0156 (oct. 2010) (cf. p. 39).

21. BIRNEY, E. The International Human Genome Project. *Human Molecular Genetics* **30,** R161-R163. ISSN : 0964-6906 (oct. 2021) (cf. p. 44).

22. BLONDEL, V. D., GUILLAUME, J.-L., LAMBIOTTE, R. *et al.* Fast unfolding of communities in large networks. en. *Journal of Statistical Mechanics : Theory and Experiment* **2008,** P10008. ISSN : 1742-5468 (oct. 2008) (cf. p. 80).

23. BOLGER, A. M., LOHSE, M. & USADEL, B. Trimmomatic : a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114-2120. ISSN : 1367-4803 (août 2014) (cf. p. 50).

24. BOURQUE, G., BURNS, K. H., GEHRING, M. *et al.* Ten things you should know about transposable elements. en. *Genome Biology* **19,** 1-12. ISSN : 1474-760X (déc. 2018) (cf. p. 34).

25. BOYLE, A. P., DAVIS, S., SHULHA, H. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. eng. *Cell* **132,** 311-322. ISSN : 1097-4172 (jan. 2008) (cf. p. 66).

26. BRAY, N. L., PIMENTEL, H., MELSTED, P. *et al.* Near-optimal probabilistic RNA-seq quantification. en. *Nature Biotechnology* **34,** 525-527. ISSN : 1546-1696 (mai 2016) (cf. p. 60).

27. BUENROSTRO, J., WU, B., CHANG, H. *et al.* ATAC-seq : A Method for Assaying Chromatin Accessibility Genome-Wide. en. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* **109,** 21.29.1 (2015) (cf. p. 66).

28. BUENROSTRO, J. D., GIRESI, P. G., ZABA, L. C. *et al.* Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. en. *Nature Methods* **10,** 1213-1218. ISSN : 1548-7105 (déc. 2013) (cf. p. 66).

29. BUJA, A. & EYUBOGLU, N. Remarks on Parallel Analysis. *Multivariate Behavioral Research* **27.** Publisher : Routledge _eprint : https ://doi.org/10.1207/s15327906mbr2704_-2, 509-540. ISSN : 0027-3171 (oct. 1992) (cf. p. 80).

30. BULIK-SULLIVAN, B. K., LOH, P.-R., FINUCANE, H. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. en. *Nature Genetics* **47,** 291-295. ISSN : 1546-1718 (mars 2015) (cf. p. 54).

31. BYCROFT, C., FREEMAN, C., PETKOVA, D. *et al.* The UK Biobank resource with deep phenotyping and genomic data. eng. *Nature* **562,** 203-209. ISSN : 1476-4687 (oct. 2018) (cf. p. 91).

32. BYERS, D. L. in *Population and quantitative genetics* Nature Education 1(1) :29 (2008) (cf. p. 54).

33. CHEN, F. X., SMITH, E. R. & SHILATIFARD, A. Born to run : control of transcription elongation by RNA polymerase II. en. *Nature Reviews Molecular Cell Biology* **19,** 464-478. ISSN : 1471-0080 (juill. 2018) (cf. p. 24).

34. CHEN, H., LI, C., PENG, X. *et al.* A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. *Cell* **173,** 386-399.e12. ISSN : 0092-8674 (avr. 2018) (cf. p. 43, 92).

35. CHOUDHARY, S. & SATIJA, R. Comparison and evaluation of statistical error models for scRNA-seq. en. *Genome Biology* **23.** Number : 1 Publisher : BioMed Central, 1-20. ISSN : 1474-760X (déc. 2022) (cf. p. 79).

36. CHUONG, E. B., ELDE, N. C. & FESCHOTTE, C. Regulatory activities of transposable elements : from conflicts to benefits. en. *Nature Reviews Genetics* **18,** 71-86. ISSN : 1471-0064 (fév. 2017) (cf. p. 34).

37. CHURCHMAN, L. S. & WEISSMAN, J. S. Native elongating transcript sequencing (NET-seq). eng. *Current Protocols in Molecular Biology* **Chapter 4,** Unit 4.14.1-17. ISSN : 1934-3647 (avr. 2012) (cf. p. 64).

38. CLOUGH, E. & BARRETT, T. The Gene Expression Omnibus database. *Methods in molecular biology (Clifton, N.J.)* **1418,** 93-110. ISSN : 1064-3745 (2016) (cf. p. 85, 89, 92).

39. COMMINS, J., TOFT, C. & FARES, M. A. Computational Biology Methods and Their Application to the Comparative Genomics of Endocellular Symbiotic Bacteria of Insects. *Biological Procedures Online* **11,** 52-78. ISSN : 1480-9222 (mars 2009) (cf. p. 50).

40. CONAWAY, R. C. & CONAWAY, J. W. The hunt for RNA polymerase II elongation factors : a historical perspective. en. *Nature Structural & Molecular Biology* **26.** Number : 9 Publisher : Nature Publishing Group, 771-776. ISSN : 1545-9985 (sept. 2019) (cf. p. 25).

41. CONESA, A., MADRIGAL, P., TARAZONA, S. *et al.* A survey of best practices for RNA-seq data analysis. en. *Genome Biology* **17,** 1-19. ISSN : 1474-760X (déc. 2016) (cf. p. 59).

42. CORE, L. & ADELMAN, K. Promoter-proximal pausing of RNA polymerase II : a nexus of gene regulation. *Genes & Development* **33,** 960-982. ISSN : 0890-9369 (août 2019) (cf. p. 25).

43. CORE, L. J., MARTINS, A. L., DANKO, C. G. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature genetics* **46,** 1311-1320. ISSN : 1061-4036 (déc. 2014) (cf. p. 26, 64, 65).

44. CORE, L. J., WATERFALL, J. J. & LIS, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science (New York, N.Y.)* **322,** 1845-1848. ISSN : 0036-8075 (déc. 2008) (cf. p. 26, 64, 179).

45. DAVIDSON, I. F. & PETERS, J.-M. Genome folding through loop extrusion by SMC complexes. en. *Nature Reviews Molecular Cell Biology* **22.** Number : 7 Publisher : Nature Publishing Group, 445-464. ISSN : 1471-0080 (juill. 2021) (cf. p. 31, 32).

46. DAVIS, C. A., HITZ, B. C., SLOAN, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE) : data portal update. eng. *Nucleic Acids Research* **46,** D794-D801. ISSN : 1362-4962 (jan. 2018) (cf. p. 16).

47. DI TOMMASO, P., CHATZOU, M., FLODEN, E. W. *et al.* Nextflow enables reproducible computational workflows. en. *Nature Biotechnology* **35.** Number : 4 Publisher : Nature Publishing Group, 316-319. ISSN : 1546-1696 (avr. 2017) (cf. p. 84).

48. DIAZ, A., PARK, K., LIM, D. A. *et al.* Normalization, bias correction, and peak calling for ChIP-seq. *Statistical Applications in Genetics and Molecular Biology* **11,** 10.1515/1544–6115.1750 /j/sagmb.2012.11.issue–3/1544–6115.1750/1544-6115.1750.xml. ISSN : 1544-6115 (mars 2012) (cf. p. 72).

49. DJEBALI, S., DAVIS, C. A., MERKEL, A. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101-108. ISSN : 0028-0836 (sept. 2012) (cf. p. 92).

50. DOBIN, A., DAVIS, C. A., SCHLESINGER, F. *et al.* STAR : ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15-21. ISSN : 1367-4803 (jan. 2013) (cf. p. 59, 78).

51. DONG, W., MOSES, C. & LI, K. *Efficient k-nearest neighbor graph construction for generic similarity measures* in *Proceedings of the 20th international conference on World wide web* (Association for Computing Machinery, New York, NY, USA, mars 2011), 577-586. ISBN : 978-1-4503-0632-4 (cf. p. 80).

52. DURBIN, R. M., ALTSHULER, D., DURBIN, R. M. *et al.* A map of human genome variation from population-scale sequencing. en. *Nature* **467,** 1061-1073. ISSN : 1476-4687 (oct. 2010) (cf. p. 44, 53).

53. EATON, J. D., FRANCIS, L., DAVIDSON, L. *et al.* A unified allosteric/torpedo mechanism for transcriptional termination on human protein-coding genes. en. *Genes & Development* **34.** Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab, 132-145. ISSN : 0890-9369, 1549-5477 (jan. 2020) (cf. p. 26).

54. EATON, J. D. & WEST, S. Termination of Transcription by RNA Polymerase II : BOOM! en. *Trends in Genetics* **36,** 664-675. ISSN : 01689525 (sept. 2020) (cf. p. 27).

55. EDER, T. & GREBIEN, F. Comprehensive assessment of differential ChIP-seq tools guides optimal algorithm selection. en. *Genome Biology* **23.** Number : 1 Publisher : BioMed Central, 1-27. ISSN : 1474-760X (déc. 2022) (cf. p. 72).

56. EICK, D. & GEYER, M. The RNA Polymerase II Carboxy-Terminal Domain (CTD) Code. *Chemical Reviews* **113.** Publisher : American Chemical Society, 8456-8490. ISSN : 0009-2665 (nov. 2013) (cf. p. 29).

57. ENCODE PROJECT CONSORTIUM. An integrated encyclopedia of DNA elements in the human genome. eng. *Nature* **489,** 57-74. ISSN : 1476-4687 (sept. 2012) (cf. p. 21).

58. ERNST, J. & KELLIS, M. Chromatin-state discovery and genome annotation with ChromHMM. en. *Nature Protocols* **12,** 2478-2492. ISSN : 1750-2799 (déc. 2017) (cf. p. 39).

59. FELSENFELD, G. & GROUDINE, M. Controlling the double helix. en. *Nature* **421.** Number : 6921 Publisher : Nature Publishing Group, 448-453. ISSN : 1476-4687 (jan. 2003) (cf. p. 20).

60. FINUCANE, H. K., BULIK-SULLIVAN, B., GUSEV, A. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. en. *Nature Genetics* **47,** 1228-1235. ISSN : 1546-1718 (nov. 2015) (cf. p. 55).

61. FRANKISH, A., DIEKHANS, M., FERREIRA, A.-M. *et al.* GENCODE reference annotation for the human and mouse genomes. eng. *Nucleic Acids Research* **47,** D766-D773. ISSN : 1362-4962 (jan. 2019) (cf. p. 21, 60).

62. FURLONG, E. E. M. & LEVINE, M. Developmental enhancers and chromosome topology. *Science (New York, N.Y.)* **361,** 1341-1345. ISSN : 0036-8075 (sept. 2018) (cf. p. 39).

63. GALLEGO ROMERO, I. & LEA, A. J. Leveraging massively parallel reporter assays for evolutionary questions. en. *Genome Biology* **24,** 1-18. ISSN : 1474-760X (déc. 2023) (cf. p. 74, 75).

64. GATES, L. A., FOULDS, C. E. & O'MALLEY, B. W. Histone Marks in the 'Driver's Seat' : Functional Roles in Steering the Transcription Cycle. English. *Trends in Biochemical Sciences* **42.** Publisher : Elsevier, 977-989. ISSN : 0968-0004 (déc. 2017) (cf. p. 39).

65. GAZZOLI, I., PULYAKHINA, I., VERWEY, N. E. *et al.* Non-sequential and multi-step splicing of the dystrophin transcript. *RNA Biology* **13,** 290-305. ISSN : 1547-6286 (déc. 2015) (cf. p. 21).

66. GIORDANO, F., AIGRAIN, L., QUAIL, M. A. *et al.* De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. en. *Scientific Reports* **7,** 3935. ISSN : 2045-2322 (juin 2017) (cf. p. 51).

67. GÖKE, J., LU, X., CHAN, Y.-S. *et al.* Dynamic Transcription of Distinct Classes of Endogenous Retroviral Elements Marks Specific Populations of Early Human Embryonic Cells. *Cell Stem Cell* **16,** 135-141. ISSN : 1934-5909 (fév. 2015) (cf. p. 144).

68. GRANDI, F. C., MODI, H., KAMPMAN, L. *et al.* Chromatin accessibility profiling by ATAC-seq. en. *Nature Protocols* **17.** Number : 6 Publisher : Nature Publishing Group, 1518-1552. ISSN : 1750-2799 (juin 2022) (cf. p. 67).

69. GREENBERG, M. V. C. & BOURC'HIS, D. The diverse roles of DNA methylation in mammalian development and disease. en. *Nature Reviews Molecular Cell Biology* **20,** 590-607. ISSN : 1471-0080 (oct. 2019) (cf. p. 37).

70. GREGERSEN, L. H., MITTER, R., UGALDE, A. P. *et al.* SCAF4 and SCAF8, mRNA Anti-Terminator Proteins. *Cell* **177,** 1797-1813.e18. ISSN : 0092-8674 (juin 2019) (cf. p. 27).

71. GUO, Y., ZHAO, S., SHENG, Q. *et al.* RNAseq by Total RNA Library Identifies Additional RNAs Compared to Poly(A) RNA Library. *BioMed Research International* **2015,** 862130. ISSN : 2314-6133 (2015) (cf. p. 21).

72. HABERLE, V. & STARK, A. Eukaryotic core promoters and the functional basis of transcription initiation. en. *Nature Reviews Molecular Cell Biology* **19,** 621-637. ISSN : 1471-0080 (oct. 2018) (cf. p. 23-25, 29, 30).

73. HAHN, S. Structure and mechanism of the RNA polymerase II transcription machinery. en. *Nature Structural & Molecular Biology* **11.** Number : 5 Publisher : Nature Publishing Group, 394-403. ISSN : 1545-9985 (mai 2004) (cf. p. 28).

74. HAMMAL, F., de LANGEN, P., BERGON, A. *et al.* 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,** D316-D325. ISSN : 0305-1048 (jan. 2022) (cf. p. 72, 90).

75. HEINZ, S., BENNER, C., SPANN, N. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. eng. *Molecular Cell* **38,** 576-589. ISSN : 1097-4164 (mai 2010) (cf. p. 68, 72).

76. HENRIQUES, T., SCRUGGS, B. S., INOUYE, M. O. *et al.* Widespread transcriptional pausing and elongation control at enhancers. en. *Genes & Development* **32.** Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab, 26-41. ISSN : 0890-9369, 1549-5477 (jan. 2018) (cf. p. 41).

77. HIRABAYASHI, S., BHAGAT, S., MATSUKI, Y. *et al.* NET-CAGE characterizes the dynamics and topology of human transcribed cis-regulatory elements. en. *Nature Genetics* **51.** Number : 9 Publisher : Nature Publishing Group, 1369-1379. ISSN : 1546-1718 (sept. 2019) (cf. p. 64).

78. *https ://www.genome.gov/Images/EdKit/bio2j_large.gif* (cf. p. 22).

79. HUSSAIN, S., SADOUNI, N., van ESSEN, D. *et al.* Short tandem repeats are important contributors to silencer elements in T cells. *Nucleic Acids Research* **51,** 4845-4866. ISSN : 0305-1048 (mars 2023) (cf. p. 75).

80. *Illumina adapter portfolio - https ://knowledge.illumina.com/library-preparation/general/library-preparation-general-reference_material-list/000003275* en (cf. p. 47).

81. INOUE, F. & AHITUV, N. Decoding enhancers using massively parallel reporter assays. *Genomics. Recent advances in functional assays of transcriptional enhancers* **106,** 159-164. ISSN : 0888-7543 (sept. 2015) (cf. p. 74, 76).

82. INSTITUTE, N. H. G. R. *DNA, alternative splicing* (cf. p. 22).

83. Intergenic RNAP2 Atlas : output data - https ://zenodo.org/records/7740073. en (cf. p. 93).

84. Intergenic RNAPII Atlas : input data - https ://zenodo.org/records/7785393. en (cf. p. 93).

85. ISLAM, S., ZEISEL, A., JOOST, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. en. *Nature Methods* **11.** Number : 2 Publisher : Nature Publishing Group, 163-166. ISSN : 1548-7105 (fév. 2014) (cf. p. 77).

86. JIANG, Y., SHEN, H., LIU, X. *et al.* Genetic Variants at 1p11.2 and Breast Cancer Risk : A Two-Stage Study in Chinese Women. en. *PLOS ONE* **6.** Publisher : Public Library of Science, e21563. ISSN : 1932-6203 (2011) (cf. p. 42).

87. JOHNSON, D. S., MORTAZAVI, A., MYERS, R. M. *et al.* Genome-wide mapping of in vivo protein-DNA interactions. eng. *Science (New York, N.Y.)* **316,** 1497-1502. ISSN : 1095-9203 (juin 2007) (cf. p. 70).

88. JÖNSSON, M. E., GARZA, R., JOHANSSON, P. A. *et al.* Transposable Elements : A Common Feature of Neurodevelopmental and Neurodegenerative Disorders. English. *Trends in Genetics* **36.** Publisher : Elsevier, 610-623. ISSN : 0168-9525 (août 2020) (cf. p. 35).

89. JORDÀ, M., DÍEZ-VILLANUEVA, A., MALLONA, I. *et al.* The epigenetic landscape of Alu repeats delineates the structural and functional genomic architecture of colon cancer cells. *Genome Research* **27,** 118-132. ISSN : 1088-9051 (jan. 2017) (cf. p. 146).

90. KAIKKONEN, M. U., SPANN, N. J., HEINZ, S. *et al.* Remodeling of the Enhancer Landscape during Macrophage Activation Is Coupled to Enhancer Transcription. *Molecular Cell* **51,** 310-325. ISSN : 1097-2765 (août 2013) (cf. p. 40).

91. KAYA-OKUR, H. S., JANSSENS, D. H., HENIKOFF, J. G. *et al.* Efficient low-cost chromatin profiling with CUT&Tag. *Nature protocols* **15,** 3264-3283. ISSN : 1754-2189 (oct. 2020) (cf. p. 73).

92. KAYA-OKUR, H. S., WU, S. J., CODOMO, C. A. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications* **10,** 1930. ISSN : 2041-1723 (avr. 2019) (cf. p. 73).

93. KELLEY, D. R., RESHEF, Y. A., BILESCHI, M. *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research* **28,** 739-750. ISSN : 1088-9051 (mai 2018) (cf. p. 23, 182).

94. KEREN, H., LEV-MAOR, G. & AST, G. Alternative splicing and evolution : diversification, exon definition and function. en. *Nature Reviews Genetics* **11,** 345-355. ISSN : 1471-0064 (mai 2010) (cf. p. 21).

95. KIM, T.-K., HEMBERG, M., GRAY, J. M. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465,** 182-187. ISSN : 0028-0836 (mai 2010) (cf. p. 31).

96. KINGMA, D. P. & WELLING, M. *Auto-Encoding Variational Bayes* en. Déc. 2013 (cf. p. 160).

97. KIVIOJA, T., VÄHÄRAUTIO, A., KARLSSON, K. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. en. *Nature Methods* **9.** Number : 1 Publisher : Nature Publishing Group, 72-74. ISSN : 1548-7105 (jan. 2012) (cf. p. 77).

98. KLEIN, J. C., AGARWAL, V., INOUE, F. *et al.* A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nature methods* **17,** 1083-1091. ISSN : 1548-7091 (nov. 2020) (cf. p. 76).

99. KLOC, M., FOREMAN, V. & REDDY, S. A. Binary function of mRNA. *Biochimie. "Coding or Non-coding : need they be exclusive ?"* **93,** 1955-1961. ISSN : 0300-9084 (nov. 2011) (cf. p. 40).

100. KUNDAJE, A., MEULEMAN, W., ERNST, J. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317-330. ISSN : 0028-0836 (fév. 2015) (cf. p. 39, 70).

101. KWAK, H., FUDA, N. J., CORE, L. J. *et al.* Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. eng. *Science (New York, N.Y.)* **339,** 950-953. ISSN : 1095-9203 (fév. 2013) (cf. p. 64).

102. L. LUN, A. T., BACH, K. & MARIONI, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. en. *Genome Biology* **17.** Number : 1 Publisher : BioMed Central, 1-14. ISSN : 1474-760X (déc. 2016) (cf. p. 79).

103. LANDER, E. S., LINTON, L. M., BIRREN, B. *et al.* Initial sequencing and analysis of the human genome. eng. *Nature* **409,** 860-921. ISSN : 0028-0836 (fév. 2001) (cf. p. 44).

104. LANDT, S. G., MARINOV, G. K., KUNDAJE, A. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research* **22,** 1813-1831. ISSN : 1088-9051 (sept. 2012) (cf. p. 71, 72).

105. LANGMEAD, B. & SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. en. *Nature Methods* **9,** 357-359. ISSN : 1548-7105 (avr. 2012) (cf. p. 63, 68).

106. LEE, T. I. & YOUNG, R. A. Transcriptional Regulation and its Misregulation in Disease. *Cell* **152,** 1237-1251. ISSN : 0092-8674 (mars 2013) (cf. p. 42).

107. LI, J. & LIU, C. Coding or Noncoding, the Converging Concepts of RNAs. *Frontiers in Genetics* **10.** ISSN : 1664-8021 (2019) (cf. p. 40).

108. LI, W., NOTANI, D., MA, Q. *et al.* Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. en. *Nature* **498,** 516-520. ISSN : 1476-4687 (juin 2013) (cf. p. 40).

109. LIANG, K. & KELEŞ, S. Normalization of ChIP-seq data with control. *BMC Bioinformatics* **13,** 199. ISSN : 1471-2105 (août 2012) (cf. p. 71, 72).

110. LIAO, Y., SMYTH, G. K. & SHI, W. featureCounts : an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923-930. ISSN : 1367-4803 (avr. 2014) (cf. p. 60, 68).

111. LIEBERMAN-AIDEN, E., van BERKUM, N. L., WILLIAMS, L. *et al.* Comprehensive mapping of long range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)* **326,** 289-293. ISSN : 0036-8075 (oct. 2009) (cf. p. 32).

112. LITTLEWOOD, T. D., KREUZALER, P. & EVAN, G. I. All Things to All People. *Cell* **151,** 11-13. ISSN : 0092-8674 (sept. 2012) (cf. p. 42).

113. LIU, H., TAM, D., MUQEETH, M. *et al.* Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning. en (cf. p. 182).

114. LOGSDON, G. A., VOLLGER, M. R. & EICHLER, E. E. Long-read human genome sequencing and its applications. en. *Nature Reviews Genetics* **21,** 597-614. ISSN : 1471-0064 (oct. 2020) (cf. p. 51).

115. LONSDALE, J., THOMAS, J., SALVATORE, M. *et al.* The Genotype-Tissue Expression (GTEx) project. en. *Nature Genetics* **45,** 580-585. ISSN : 1546-1718 (juin 2013) (cf. p. 43, 90).

116. LOVE, M. I., HUBER, W. & ANDERS, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. eng. *Genome Biology* **15,** 550. ISSN : 1474-760X (2014) (cf. p. 60, 68, 79).

117. LU, X., SACHS, F., RAMSAY, L. *et al.* The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. en. *Nature Structural & Molecular Biology* **21.** Number : 4 Publisher : Nature Publishing Group, 423-425. ISSN : 1545-9985 (avr. 2014) (cf. p. 143, 144).

118. LUN, A. T. L. & SMYTH, G. K. csaw : a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. eng. *Nucleic Acids Research* **44,** e45. ISSN : 1362-4962 (mars 2016) (cf. p. 68, 72).

119. LUO, Y., HITZ, B. C., GABDANK, I. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. eng. *Nucleic Acids Research* **48,** D882-D889. ISSN : 1362-4962 (jan. 2020) (cf. p. 89, 92).

120. MAATEN, L. v. d. & HINTON, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9,** 2579-2605. ISSN : 1533-7928 (2008) (cf. p. 80).

121. MAHAT, D. B., KWAK, H., BOOTH, G. T. *et al.* Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). en. *Nature Protocols* **11,** 1455-1476. ISSN : 1750-2799 (août 2016) (cf. p. 64).

122. MARGULIES, M., EGHOLM, M., ALTMAN, W. E. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. en. *Nature* **437,** 376-380. ISSN : 1476-4687 (sept. 2005) (cf. p. 47).

123. MARTIN, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. en. *EMBnet.journal* **17,** 10-12. ISSN : 2226-6089 (mai 2011) (cf. p. 50).

124. MARTINEZ, T. F., CHU, Q., DONALDSON, C. *et al.* Accurate annotation of human protein-coding small open reading frames. en. *Nature Chemical Biology* **16,** 458-468. ISSN : 1552-4469 (avr. 2020) (cf. p. 40).

125. MARTÍNEZ-GARCÍA, P. M., GARCÍA-TORRES, M., DIVINA, F. *et al.* Genome-wide prediction of topoisomerase IIbeta binding by architectural factors and chromatin accessibility. en. *PLOS Computational Biology* **17.** Publisher : Public Library of Science, e1007814. ISSN : 1553-7358 (jan. 2021) (cf. p. 73).

126. MCINNES, L., HEALY, J., SAUL, N. *et al.* UMAP : Uniform Manifold Approximation and Projection. en. *Journal of Open Source Software* **3,** 861. ISSN : 2475-9066 (sept. 2018) (cf. p. 80).

127. MCKERROW, W., WANG, X., MENDEZ-DORANTES, C. *et al.* LINE-1 expression in cancer correlates with p53 mutation, copy number alteration, and S phase checkpoint. *Proceedings of the National Academy of Sciences* **119.** Publisher : Proceedings of the National Academy of Sciences, e2115999119 (fév. 2022) (cf. p. 146).

128. MCLEAN, C. Y., BRISTOR, D., HILLER, M. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28,** 495-501. ISSN : 1087-0156 (mai 2010) (cf. p. 68).

129. MEDINA-RIVERA, A., SANTIAGO-ALGARRA, D., PUTHIER, D. *et al.* Widespread Enhancer Activity from Core Promoters. eng. *Trends in Biochemical Sciences* **43,** 452-468. ISSN : 0968-0004 (juin 2018) (cf. p. 31).

130. MELNIKOV, A., MURUGAN, A., ZHANG, X. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. en. *Nature Biotechnology* **30.** Number : 3 Publisher : Nature Publishing Group, 271-277. ISSN : 1546-1696 (mars 2012) (cf. p. 74).

131. MEULEMAN, W., MURATOV, A., RYNES, E. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. en. *Nature* **584,** 244-251. ISSN : 1476-4687 (août 2020) (cf. p. 30, 66).

132. MIKHAYLICHENKO, O., BONDARENKO, V., HARNETT, D. *et al.* The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. en. *Genes & Development* **32,** 42-57. ISSN : 0890-9369, 1549-5477 (jan. 2018) (cf. p. 31).

133. MÖLDER, F., JABLONSKI, K. P., LETCHER, B. *et al. Sustainable data analysis with Snakemake* en. Rapp. tech. 10 :33. Type : article (F1000Research, jan. 2021) (cf. p. 84).

134. MOORE, L. D., LE, T. & FAN, G. DNA Methylation and Its Basic Function. en. *Neuropsychopharmacology* **38,** 23-38. ISSN : 1740-634X (jan. 2013) (cf. p. 37).

135. MORTAZAVI, A., WILLIAMS, B. A., MCCUE, K. *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq. en. *Nature Methods* **5,** 621-628. ISSN : 1548-7105 (juill. 2008) (cf. p. 58).

136. MUERDTER, F., BORYŃ, Ł. M., WOODFIN, A. R. *et al.* Resolving systematic errors in widely-used enhancer activity assays in human cells. *Nature methods* **15,** 141-149. ISSN : 1548-7091 (fév. 2018) (cf. p. 76).

137. NAYAN, V., ONTERU, S. K. & SINGH, D. Reproduction and nutriment–nurture crosstalk : epigenetic perspectives. *Journal of Reproductive Health and Medicine* **1,** 50-59. ISSN : 2214-420X (juill. 2015) (cf. p. 38).

138. NEALE, B. M. Introduction to Linkage Disequilibrium, the HapMap, and Imputation. en. *Cold Spring Harbor Protocols* **2010,** pdb.top74. ISSN : 1940-3402, 1559-6095, 1559-6095 (mars 2010) (cf. p. 53).

139. NTRANOS, V., YI, L., MELSTED, P. *et al.* A discriminative learning approach to differential expression analysis for single-cell RNA-seq. en. *Nature Methods* **16.** Number : 2 Publisher : Nature Publishing Group, 163-166. ISSN : 1548-7105 (fév. 2019) (cf. p. 80).

140. NURK, S., KOREN, S., RHIE, A. *et al.* The complete sequence of a human genome. *Science* **376,** 44-53 (avr. 2022) (cf. p. 36, 51, 146).

141. OGBOURNE, S. & ANTALIS, T. M. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochemical Journal* **331,** 1-14. ISSN : 0264-6021 (avr. 1998) (cf. p. 29).

142. OHMIYA, H., VITEZIC, M., FRITH, M. C. *et al.* RECLU : a pipeline to discover reproducible transcriptional start sites and their alternative regulation using capped analysis of gene expression (CAGE). *BMC Genomics* **15,** 269. ISSN : 1471-2164 (avr. 2014) (cf. p. 63).

143. ØROM, U. A., DERRIEN, T., BERINGER, M. *et al.* Long Noncoding RNAs with Enhancer-like Function in Human Cells. *Cell* **143,** 46-58. ISSN : 0092-8674 (oct. 2010) (cf. p. 40).

144. PARK, P. J. ChIP–seq : advantages and challenges of a maturing technology. en. *Nature Reviews Genetics* **10,** 669-680. ISSN : 1471-0064 (oct. 2009) (cf. p. 70, 71, 73).

145. PATRO, R., DUGGAL, G., LOVE, M. I. *et al.* Salmon : fast and bias-aware quantification of transcript expression using dual-phase inference. *Nature methods* **14,** 417-419. ISSN : 1548-7091 (avr. 2017) (cf. p. 60).

146. PATWARDHAN, R. P., HIATT, J. B., WITTEN, D. M. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. en. *Nature Biotechnology* **30.** Number : 3 Publisher : Nature Publishing Group, 265-270. ISSN : 1546-1696 (mars 2012) (cf. p. 74).

147. PAUL, J. & DUERKSEN, J. D. Chromatin-associated RNA content of heterochromatin and euchromatin. eng. *Molecular and Cellular Biochemistry* **9,** 9-16. ISSN : 0300-8177 (oct. 1975) (cf. p. 40).

148. PEASTON, A. E., EVSIKOV, A. V., GRABER, J. H. *et al.* Retrotransposons Regulate Host Genes in Mouse Oocytes and Preimplantation Embryos. *Developmental Cell* **7,** 597-606. ISSN : 1534-5807 (oct. 2004) (cf. p. 34).

149. PIERRE, d. L. & BENOIT, B. *A high-resolution atlas of intergenic transcription in nearly 29,000 normal and cancer samples, https ://github.com/benoitballester/Pol2Atlas* Issue : 1 Pages : 1–10 Publication Title : Journal Title Volume : 1 original-date : 2023-01-04T09 :25 :19Z. Sept. 2023 (cf. p. 93).

150. POTT, S. & LIEB, J. D. What are super-enhancers ? eng. *Nature Genetics* **47,** 8-12. ISSN : 1546-1718 (jan. 2015) (cf. p. 30).

151. QUAIL, M. A., SMITH, M., COUPLAND, P. *et al.* A tale of three next generation sequencing platforms : comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13,** 341. ISSN : 1471-2164 (juill. 2012) (cf. p. 47, 49).

152. RAUDVERE, U., KOLBERG, L., KUZMIN, I. *et al.* g :Profiler : a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research* **47,** W191-W198. ISSN : 0305-1048 (juill. 2019) (cf. p. 60).

153. RAY, D., KAZAN, H., COOK, K. B. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499,** 172-177. ISSN : 0028-0836 (juill. 2013) (cf. p. 40).

154. RHEE, H. S. & PUGH, B. F. ChIP-exo : A Method to Identify Genomic Location of DNA-binding proteins at Near Single Nucleotide Accuracy. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* **0 21,** 10.1002/0471142727.mb2124s100. ISSN : 1934-3639 (oct. 2012) (cf. p. 73).

155. RITCHIE, M. E., PHIPSON, B., WU, D. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43,** e47. ISSN : 0305-1048 (avr. 2015) (cf. p. 60).

156. SAHU, B., HARTONEN, T., PIHLAJAMAA, P. *et al.* Sequence determinants of human gene regulatory elements. en. *Nature Genetics* **54,** 283-294. ISSN : 1546-1718 (mars 2022) (cf. p. 23, 24, 30).

157. SANGER, F. & COULSON, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94,** 441-448. ISSN : 0022-2836 (mai 1975) (cf. p. 44).

158. SANTA, F. D., BAROZZI, I., MIETTON, F. *et al.* A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. en. *PLOS Biology* **8,** e1000384. ISSN : 1545-7885 (2010) (cf. p. 31, 70).

159. SANTONI, F. A., GUERRA, J. & LUBAN, J. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. eng. *Retrovirology* **9,** 111. ISSN : 1742-4690 (déc. 2012) (cf. p. 34).

160. SARTORELLI, V. & LAUBERTH, S. M. Enhancer RNAs are an important regulatory layer of the epigenome. en. *Nature Structural & Molecular Biology* **27,** 521-528. ISSN : 1545-9985 (juin 2020) (cf. p. 31, 40).

161. SATIJA, R., FARRELL, J. A., GENNERT, D. *et al.* Spatial reconstruction of single-cell gene expression data. en. *Nature Biotechnology* **33.** Number : 5 Publisher : Nature Publishing Group, 495-502. ISSN : 1546-1696 (mai 2015) (cf. p. 79).

162. SCHAUKOWITCH, K., JOO, J.-Y., LIU, X. *et al.* Enhancer RNA Facilitates NELF Release from Immediate Early Genes. *Molecular Cell* **56,** 29-42. ISSN : 1097-2765 (oct. 2014) (cf. p. 40).

163. SCHWALB, B., MICHEL, M., ZACHER, B. *et al.* TT-seq maps the human transient transcriptome. *Science* **352,** 1225-1228 (juin 2016) (cf. p. 64).

164. SEGERT, J. A., GISSELBRECHT, S. S. & BULYK, M. L. Transcriptional silencers : driving gene expression with the brakes on. *Trends in genetics : TIG* **37,** 514-527. ISSN : 0168-9525 (juin 2021) (cf. p. 30).

165. SEILA, A. C., CALABRESE, J. M., LEVINE, S. S. *et al.* Divergent transcription from active promoters. *Science (New York, N.Y.)* **322,** 1849-1851. ISSN : 0036-8075 (déc. 2008) (cf. p. 26).

166. SEILA, A. C., CORE, L. J., LIS, J. T. *et al.* Divergent transcription : A new feature of active promoters. *Cell Cycle* **8.** Publisher : Taylor & Francis _eprint : https ://doi.org/10.4161/cc.8.16.9305, 2557-2564. ISSN : 1538-4101 (août 2009) (cf. p. 26).

167. *Seurat - Guided Clustering Tutorial - https ://satijalab.org/seurat/articles/pbmc3k_tutorial* en (cf. p. 79).

168. SHENDURE, J., BALASUBRAMANIAN, S., CHURCH, G. M. *et al.* DNA sequencing at 40 : past, present and future. en. *Nature* **550,** 345-353. ISSN : 1476-4687 (oct. 2017) (cf. p. 45, 46, 49).

169. SHIRAKI, T., KONDO, S., KATAYAMA, S. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences* **100,** 15776-15781 (déc. 2003) (cf. p. 62).

170. SHLYUEVA, D., STAMPFEL, G. & STARK, A. Transcriptional enhancers : from properties to genome-wide predictions. en. *Nature Reviews Genetics* **15,** 272-286. ISSN : 1471-0064 (avr. 2014) (cf. p. 29).

171. SIMON, J. M., GIRESI, P. G., DAVIS, I. J. *et al.* Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. eng. *Nature Protocols* **7,** 256-267. ISSN : 1750-2799 (jan. 2012) (cf. p. 66).

172. SLATKIN, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. en. *Nature Reviews Genetics* **9,** 477-485. ISSN : 1471-0064 (juin 2008) (cf. p. 53).

173. Smit, AFA, Hubley, R and Green, P. RepeatMasker Open-4.0. 2013-2015 <http ://www.repeatmasker.org>. (Cf. p. 36).

174. SONG, L. & CRAWFORD, G. E. DNase-seq : a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor protocols* **2010,** pdb.prot5384. ISSN : 1940-3402 (fév. 2010) (cf. p. 66).

175. STARK, R., GRZELAK, M. & HADFIELD, J. RNA sequencing : the teenage years. en. *Nature Reviews Genetics* **20,** 631-656. ISSN : 1471-0064 (nov. 2019) (cf. p. 58, 59).

176. SZABO, Q., BANTIGNIES, F. & CAVALLI, G. Principles of genome folding into topologically associating domains. en. *Science Advances* **5,** eaaw1668. ISSN : 2375-2548 (avr. 2019) (cf. p. 31, 33).

177. TAKAHASHI, H., KATO, S., MURATA, M. *et al.* CAGE- Cap Analysis Gene Expression : a protocol for the detection of promoter and transcriptional networks. *Methods in molecular biology (Clifton, N.J.)* **786,** 181-200. ISSN : 1064-3745 (2012) (cf. p. 62).

178. TARBELL, E. D. & LIU, T. HMMRATAC : a Hidden Markov ModeleR for ATAC-seq. *Nucleic Acids Research* **47,** e91. ISSN : 0305-1048 (sept. 2019) (cf. p. 68).

179. THOMAS, M. C. & CHIANG, C.-M. The general transcription machinery and general cofactors. eng. *Critical Reviews in Biochemistry and Molecular Biology* **41,** 105-178. ISSN : 1040-9238 (2006) (cf. p. 23).

180. THURMAN, R. *hotspot* juill. 2022 (cf. p. 68).

181. TRAAG, V. A., WALTMAN, L. & van ECK, N. J. From Louvain to Leiden : guaranteeing well-connected communities. en. *Scientific Reports* **9.** Number : 1 Publisher : Nature Publishing Group, 5233. ISSN : 2045-2322 (mars 2019) (cf. p. 80).

182. VALENTE, A., VIEIRA, L., SILVA, M. J. *et al.* The Effect of Nanomaterials on DNA Methylation : A Review. en. *Nanomaterials* **13.** Number : 12 Publisher : Multidisciplinary Digital Publishing Institute, 1880. ISSN : 2079-4991 (jan. 2023) (cf. p. 37).

183. VOLDERS, P.-J., HELSENS, K., WANG, X. *et al.* LNCipedia : a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Research* **41,** D246-D251. ISSN : 0305-1048 (jan. 2013) (cf. p. 21).

184. WAN, Y., KERTESZ, M., SPITALE, R. C. *et al.* Understanding the transcriptome through RNA structure. *Nature reviews. Genetics* **12,** 10.1038/nrg3049. ISSN : 1471-0056 (août 2011) (cf. p. 40).

185. WANG, K. C. & CHANG, H. Y. Molecular mechanisms of long noncoding RNAs. *Molecular cell* **43,** 904-914. ISSN : 1097-2765 (sept. 2011) (cf. p. 40).

186. WANG, Z., GERSTEIN, M. & SNYDER, M. RNA-Seq : a revolutionary tool for transcriptomics. en. *Nature Reviews Genetics* **10,** 57-63. ISSN : 1471-0064 (jan. 2009) (cf. p. 58, 59).

187. WARD, L. D. & KELLIS, M. Interpreting noncoding genetic variation in complex traits and human disease. en. *Nature Biotechnology* **30.** Number : 11 Publisher : Nature Publishing Group, 1095-1106. ISSN : 1546-1696 (nov. 2012) (cf. p. 43, 90).

188. WATSON, J. D. & COOK-DEEGAN, R. M. Origins of the Human Genome Project. eng. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **5,** 8-11. ISSN : 0892-6638 (jan. 1991) (cf. p. 44).

189. WATSON, J. D. & CRICK, F. H. C. Molecular Structure of Nucleic Acids : A Structure for Deoxyribose Nucleic Acid. en. *Nature* **171,** 737-738. ISSN : 1476-4687 (avr. 1953) (cf. p. 18).

190. WEINSTEIN, J. N., COLLISSON, E. A., MILLS, G. B. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. en. *Nature Genetics* **45,** 1113-1120. ISSN : 1546-1718 (oct. 2013) (cf. p. 90).

191. WEST, S., GROMAK, N. & PROUDFOOT, N. J. Human 5 → 3 exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. en. *Nature* **432.** Number : 7016 Publisher : Nature Publishing Group, 522-525. ISSN : 1476-4687 (nov. 2004) (cf. p. 26).

192. WHYTE, W. A., ORLANDO, D. A., HNISZ, D. *et al.* Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* **153,** 307-319. ISSN : 0092-8674 (avr. 2013) (cf. p. 29).

193. WILKS, C., ZHENG, S. C., CHEN, F. Y. *et al.* recount3 : summaries and queries for large-scale RNA-seq expression and splicing. en. *Genome Biology* **22,** 1-40. ISSN : 1474-760X (déc. 2021) (cf. p. 58).

194. WOLF, F. A., ANGERER, P. & THEIS, F. J. SCANPY : large-scale single-cell gene expression data analysis. en. *Genome Biology* **19.** Number : 1 Publisher : BioMed Central, 1-5. ISSN : 1474-760X (déc. 2018) (cf. p. 80, 163).

195. WOLFF, J., BATUT, B. & RASCHE, H. *Mapping* en. text. Publisher : Galaxy Training Network. 0 (cf. p. 51).

196. WOYCHIK, N. A. & HAMPSEY, M. The RNA Polymerase II Machinery : Structure Illuminates Function. English. *Cell* **108,** 453-463. ISSN : 0092-8674, 1097-4172 (fév. 2002) (cf. p. 27).

197. WRAY, N. R. & VISSCHER, P. M. in *Population and quantitative genetics* Nature Education 1(1) :29 (2008) (cf. p. 53, 54).

198. YAN, F., POWELL, D. R., CURTIS, D. J. *et al.* From reads to insight : a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology* **21,** 22. ISSN : 1474-760X (fév. 2020) (cf. p. 68, 69).

199. YOUNG, R. S., KUMAR, Y., BICKMORE, W. A. *et al.* Bidirectional transcription initiation marks accessible chromatin and is not specific to enhancers. en. *Genome Biology* **18.** Number : 1 Publisher : BioMed Central, 1-11. ISSN : 1474-760X (déc. 2017) (cf. p. 26).

200. YU, N. Y.-L., HALLSTRÖM, B. M., FAGERBERG, L. *et al.* Complementing tissue characterization by integrating transcriptome profiling from the Human Protein Atlas and from the FANTOM5 consortium. *Nucleic Acids Research* **43,** 6787-6798. ISSN : 0305-1048 (août 2015) (cf. p. 62).

201. *Zenodo - https ://zenodo.org/* (cf. p. 85).

202. ZHANG, H., RIGO, F. & MARTINSON, H. G. Poly(A) Signal-Dependent Transcription Termination Occurs through a Conformational Change Mechanism that Does Not Require Cleavage at the Poly(A) Site. *Molecular Cell* **59,** 437-448. ISSN : 1097-2765 (août 2015) (cf. p. 26).

203. ZHANG, Y., LIU, T., MEYER, C. A. *et al.* Model-based Analysis of ChIP-Seq (MACS). en. *Genome Biology* **9,** 1-9. ISSN : 1474-760X (nov. 2008) (cf. p. 68, 72).

204. ZHAO, J., OHSUMI, T. K., KUNG, J. T. *et al.* Genome-wide identification of Polycomb-associated RNAs by RIP-seq. *Molecular cell* **40,** 939-953. ISSN : 1097-2765 (déc. 2010) (cf. p. 40).

205. ZHAO, Y., WANG, L., REN, S. *et al.* Activation of P-TEFb by Androgen Receptor-Regulated Enhancer RNAs in Castration-Resistant Prostate Cancer. *Cell Reports* **15,** 599-610. ISSN : 2211-1247 (avr. 2016) (cf. p. 40).

206. ZHENG, G. X. Y., TERRY, J. M., BELGRADER, P. *et al.* Massively parallel digital transcriptional profiling of single cells. en. *Nature Communications* **8,** 14049. ISSN : 2041-1723 (jan. 2017) (cf. p. 77, 78).

207. ZHOU, V. W., GOREN, A. & BERNSTEIN, B. E. Charting histone modifications and the functional organization of mammalian genomes. en. *Nature Reviews Genetics* **12,** 7-18. ISSN : 1471-0064 (jan. 2011) (cf. p. 19, 38).