Découverte et caractérisation des éléments silencers par des approches à haut débit

Nori Sadouni

Aix*Marseille université
Socialement engagée

NNT/NL : 2022AIXM0184/039ED62

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 24 mai 2022 par

Nori SADOUNI

Découverte et caractérisation des éléments silencers par des approches à haut débit

Discipline
Biologie

Spécialité
Génomique & Bioinformatique

École doctorale
École Doctorale 62

Laboratoire
Theories and Approaches of Genomic Complexity - TAGC
INSERM U1090e

Composition du jury

Touati BENOUKRAF Memorial University of Newfoundland, St. John's,NL, Canada	Rapporteur
Veronique ADOUE INFINITY-U1291	Rapporteure
Christelle CAYROU Centre de Recherche en Cancérologie de Marseille (CRCM)	Examinatrice
Charles-Henri LECELLIER IGMM/LIRMM/IMAG	Président du jury
Salvatore SPICUGLIA TAGC INSERM U-1090	Directeur de thèse

Faculté des Sciences
Aix*Marseille Université

ÉCOLE DOCTORALE AMU Sciences de la Vie et de la Santé

Inserm

Table des matières

Table des figures

Liste des tableaux

Affidavit

Je soussigné, Nori SADOUNI, déclare par la présente que le travail présenté dans ce manuscrit est mon propre travail, réalisé sous la direction scientifique de Salvatore SPICUGLIA, 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 16/03/2022

Research Publications

Journal Articles

1. Hussain, S., **Sadouni**, **N.**, van Essen, D., Lopez, P., Dao, L., Charbonnier, G., … Spicuglia, S. (Submitted). Znf263 tandem repeats are important contributors of silencer elements in t cells. *Nucleic Acids Research*.

2. Santiago-Algarra, D., Souaid, C., Singh, H., Dao, L., **Sadouni**, **N.**, Hussain, S., … Charbonnier, G. et al. (2021). Epromoters function as a hub to recruit key transcription factors required for the inflammatory response. *Nature communications*, *12*(1), 1–18.

3. Alomairi, J., Molitor, A. M., **Sadouni**, **N.**, Hussain, S., Torres, M., Saadi, W., … Andrau, J. C. et al. (2020). Integration of high-throughput reporter assays identify a critical enhancer of the ikzf1 gene. *PloS one*, *15*(5), e0233191.

4. Ferré, Q., Charbonnier, G., **Sadouni**, **N.**, Lopez, F., Kermezli, Y., Spicuglia, S., … Puthier, D. (2019). OLOGRAM: determining significance of total overlap length between genomic regions sets. *Bioinformatics*, *36*(6), 1920–1922. doi:10.1093/bioinformatics/btz810. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/6/1920/32915137/btz810.pdf

5. Cherfils-Vicini, J., Iltis, C., Cervera, L., Pisano, S., Croce, O., **Sadouni**, **N.**, … Rey-Millet, M. et al. (2019). Cancer cells induce immune escape via glycocalyx changes controlled by the telomeric protein trf 2. *The EMBO journal*, *38*(11), e100012.

Conference Proceedings

1. **Sadouni**, **N.**, Fassy, J., Modelska, A., Manosalva, I., Spicuglia, S., Verhoeyen, E., … Roulland, S. (2021), In *The crispr screen action, a structuring action from the canceropôle paca (poster)*, St-Raphael, France.

2. **Sadouni**, **N.**, Hussain, S., Charbonnier, G., Torres, M., Saccani, S., van Essen, D., & Spicuglia, S. (2021), In *Identification and analysis of silencer elements in t cells (poster)*, 28th Annual Meeting of Doctoral School, Marseille.

3. **Sadouni**, **N.**, Hussain, S., Charbonnier, G., Torres, M., Saccani, S., van Essen, D., & Spicuglia, S. (2018), In *Identification and analysis of silencer elements in t cells (poster)*, St-Raphael, France.

Remerciements

Je remercie tout d'abord les membres de mon jury, les Docteurs : Touati BENOUKRAF, Véronique ADOUE, Charles-Henri LECELLIER et Christelle CAYROU, pour avoir consacré de leur temps afin d'évaluer mon travail de thèse. Double mention à Charles-Henri LECELLIER qui a également fait partie de mon comité de suivi de thèse.
Par la suite, je souhaite remercier chaleureusement le Dr.Salvatore SPICUGLIA qui m'a encadré durant toute ma thèse. Il a fait preuve de patience, pédagogie, et m'a beaucoup inspiré durant toute cette période. Je garde de très bons souvenirs en tant qu'étudiant sous sa direction. J'ai rarement vu un directeur de thèse aussi impliqué dans le travail de son étudiant. En plus de ses compétences de chercheur et de chef d'équipe, il s'est montré extrêmement chaleureux et humain. Merci pour tout Salva.

Je souhaitais également remercier l'Université d'Aix-Marseille, de m'avoir accueilli comme étudiant de thèse. La bourse AMIDEX, Cancéropôle et MarMaRa qui ont financé mon projet de thèse.
Je remercie également tous les membres du TAGC pour les bons moments qu'on a pu partager, notamment Benoît, Lucie, Thomas mon co-organisateur de pot de thèse !

Un grand merci aux anciens membres du TAGC, le Dr.Quentin FERRE, le Dr.Guillaume CHARBONNIER, qui m'ont beaucoup apporté tant sur le plan professionnel que personnel. Ils se sont révélés être de vrais amis, surtout lors de ces derniers mois particulièrement éprouvants. Ils ont su être là pour m'écouter, accueillir mes émotions, et m'aider. Ensemble, nous avons vécu des moments intenses sur Snakemake (Guillaume tu viens plus aux soirées ? !).
Un immense merci à Iris pour son soutien durant toute ma thèse et les bons moments qu'on a pu passer ensemble. Merci à Saadat, membre iconique du Saadat Mustache Club, alias la moustache la plus épaisse d'orient. We shared a lot of suppo together haha ! Et bien évidemment merci à Juliette, mercy d'avoare korri j'ai mets non breuze phote, pour les bons moments qu'on a pu partager, les nombreuses critiques cinématographiques, tu as réussi à me faire changer d'avis sur Harry Potter. À l'heure où j'écris ces lignes, il te reste 114 jours avant le 2 septembre 2022 afin d'être prête pour la série du Seigneur des anneaux ! Et surtout n'oublie pas que pour chaque situation dans la vie, il existe quelque part, un GIF de Ron Swanson afin d'y répondre. Merci également à Francesco, Antoinette, José David et tous les autres membres de l'équipe.

Je remercie ma famille, mes parents qui ont su m'apporter des fondations solides. Maintenant que je suis papa, je me rends compte des efforts que vous avez dû faire pour moi. Un grand merci à Machou. Sincèrement je te dois beaucoup. Si j'en suis là aujourd'hui c'est aussi grâce à toi.
Petite dédicace à la famille Guemar pour les soirées jeux de sociétés (et surtout Huveaune). Ainsi qu'à David Kun & Raphael Kun pour les discussions sympas sur des thèmes très originaux.

Et bien évidemment je remercie mon épouse, Lilia qui a su me soutenir durant ma thèse. Je ne compte plus les sacrifices que tu fais pour notre famille.Et que dire de mes filles ? Vous êtes ma vraie fierté, Hanna & Louisa je vous aime plus que tout au monde. Vous êtes ma force, mon moteur, ma motivation.

À mes filles Hanna & Louisa.

Abstract

Control of gene expression is a complex mechanism that use cis-regulatory element as keystone. The control of this mechanism is fundamental to mammalian cell life. Gene expression is controlled at different levels such as DNA compaction, binding of transcription factors with cofactors making regulatory complexes or epigenetic modifications. Gene regulatory elements, including promoters, enhancers, silencers, etc., control transcriptional programs in a spatiotemporal manner. These elements are able to control gene expression by increasing or decreasing transcription. The community focused mainly in the study of enhancer elements which amplify transcription initiation, while silencers, which repress gene expression, were sidelined.

Recently, thanks to the improvement of technologies allowing a better analysis of the genome and to the advances in computing, few studies successfully characterized silencers elements in several organisms including human, mouse and drosophilia. They were carried out by efficient functional assays at different scale contributing to the understanding of their regulatory mechanisms and bringing silencers into the spotlight.

However, lot of questions remains unanswered, making silencers poorly understood. Indeed, their genome-wide distribution and mechanisms of action are largely unknown. In the lab we repurposed an existing high-throughput reporter assay in order to functionally identify silencers genome-wide. As a newly developed technology, the first objective was to build a bioinformatic pipeline to statistically identify genomic regions displaying silencer activity. Subsequently, I performed a comprehensive genomic and epigenomic characterization of the identified silencers, including : motif enrichment, genomic distribution, histone marks enrichment, repetitive elements enrichment, etc. Multi-omics integration of identified silencers with transcriptomic and epigenomic resources provided a comprehensive catalogue of silencer elements in the genome. My results also shed-light into the mechanism of regulation of silencer elements, which have been experimentally validated in the lab. These results have been published in a scientific article (Article : 2.4 [Hussain et al.,]). The pipeline developed is available on my GitHub account Norisad and has been used in two other projects, on the analysis of E-promoters (Article : 2.5.2 [Santiago-Algarra et al., 2021]) and enhancers (Article : 2.5.1 [Alomairi et al., 2020]), which made it possible to observe the overlap between the positive signal of the regions with the presence of active transcriptions factor binding site or simply the processing of CapSTARR-seq data.

Keywords : Computational genomic,CapSTARR-seq, Machine Learning, regulatory element, silencers, genomic characterizations.

Résumé

Le controle de l'expression des gènes est un mécanisme complexe qui utilise les éléments Cis-régulateur comme clé de voûte. Le contrôle de ce mécanisme est fondamental pour la vie cellulaire des mammifères. L'expression des gènes est contrôlée à différents niveaux tels que la compaction de l'ADN, la liaison des facteurs de transcription avec des cofacteurs créant des complexes régulateurs ou des modifications épigénétiques. Les éléments régulateurs des gènes, notamment les promoteurs, les enhancers, les silencers, etc., contrôlent les programmes transcriptionnels de manière spatio-temporelle.

Ces éléments peuvent contrôler l'expression des gènes en augmentant ou en diminuant la transcription. La communauté s'est principalement concentrée sur l'étude des éléments enhancers qui amplifient l'initiation de la transcription, tandis que les silencers, qui répriment l'expression des gènes, ont été mis de côtés. Récemment, grâce à l'amélioration des technologies permettant l'analyse du génome et aux progrès informatiques, quelques études ont réussi à caractériser les éléments silencers chez plusieurs organismes dont l'homme, souris et la drosophile. Celles-ci ont été réalisées par des tests fonctionnels efficaces à différentes échelles contribuant à la compréhension de leurs mécanismes de régulation et mettant les silencers sous les projecteurs.

Cependant, plusieurs questions restent sans réponses, rendant les silencers incompris. En effet, leurs distributions à l'échelle du génome et leurs mécanismes d'actions sont peu connus. En laboratoire, nous avons réutilisé un test de gène rapporteur à grande échelle afin d'identifier de manière fonctionnelle les silencers à l'échelle du génome. En tant que technologie nouvellement développée, le premier objectif était de construire un programme bioinformatique pour identifier statistiquement les régions génomiques affichant une activité silencer. Par la suite, j'ai effectué une caractérisation génomique et épigénomique complète des silencers identifiés, incluant : l'enrichissement des sites de fixations de facteurs de transcriptions, la distribution génomique, l'enrichissement en différentes marques d'histones, l'enrichissement en éléments répétitifs, etc. L'intégration multi-omique des silencers identifiés avec des ressources transcriptomiques et épigénomiques a fourni un catalogue d'éléments silencers dans le génome. Mes résultats ont également mis en lumière les mécanismes de régulations des éléments silencers, qui ont été validés expérimentalement en laboratoire. Ces résultats ont été mis en forme dans un article scientifique (Article : 2.4 [Hussain et al.,]).

Le programme informatique développé est disponible sur mon compte GitHub Norisad et a été utilisé dans deux autres projets, sur l'analyse des E-promoteurs (Article : 2.5.2 [Santiago-Algarra et al., 2021]) et d'enhancers (Article : 2.5.1 [Alomairi et al., 2020]), ce qui a permis d'observer la colocalisation du signal positif de la région avec la présence de site de fixation de facteurs de transcription actifs ou tout simplement le traitement des données CapSTARR-seq.

Mots clés : Bioinformatique génomique, CapSTARR-seq, Machine learning, éléments régulateurs, silencers, répression, caractérisations génomique.

1 Introduction

1.1 Organisation du génome eucaryote

Le génome représente l'ensemble du matériel génétique d'un organisme, stocké sous forme d'acide désoxyribonucléique (ADN). Chez les eucaryotes, celui-ci est conservé à l'intérieur du noyau (Figure :1.1), à l'exception près du génome mitochondrial [Boursot and Bonhomme, 1986], ainsi que du génome plastidial des végétaux [Sato, 2007]. C'est principalement cette structure qui les distingue des procaryotes (comme les bactéries) qui sont pour leurs parts dépourvu de ces structures.

C'est dans ce noyau que le génome nucléaire est compacté et empaqueté dans un complexe macromoléculaire appelé chromatine. Celle-ci est constituée d'ADN, d'ARN et de protéines dont les structures, puis les fonctions, sont présentées dans ce chapitre.

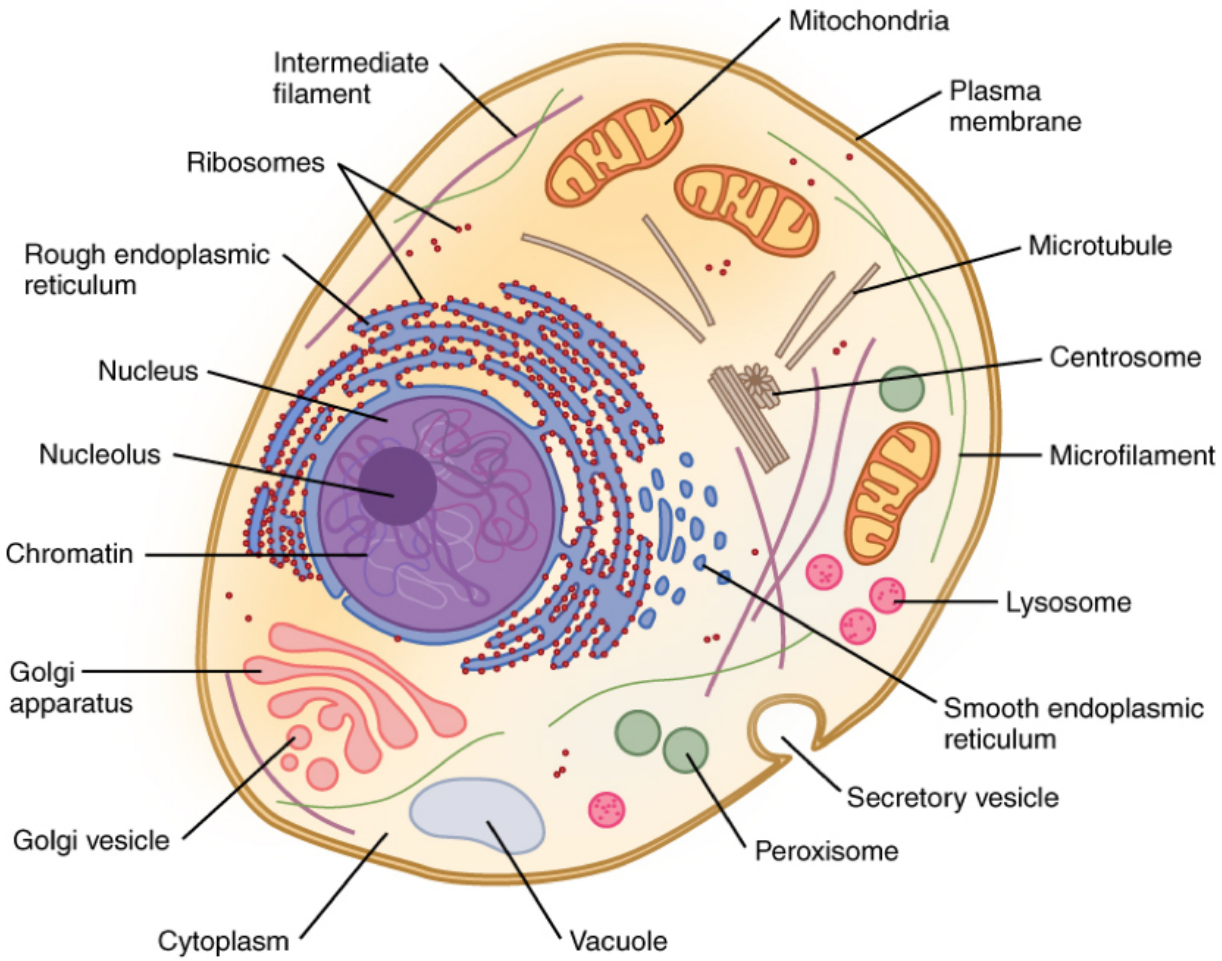


FIGURE 1.1 – Schéma d'une cellule eucaryote reprenant les différents compartiments et composants cellulaires. Richard Wheeler (Zephyris) / English Wikipedia / CC BY-SA 3.0

1.1.1 Configuration de la chromatine

1.1.1.1 Structure de l'ADN à différentes échelles

L'ADN a été décrite pour la première fois dès 1869 par Friedrich Miescher. Sa structure la plus commune en double hélice a été décrite par Watson et Crick en 1953. L'ADN est un biopolymère linéaire constitué de désoxyribonucléotides. Ces derniers sont composés d'un sucre (b-D-2'-désoxyribose), un groupement phosphate et une base nucléique. Quatre bases nucléiques se retrouvent dans l'ADN : l'adénine (A), la thymine (T), guanine (G) et cytosine (C). Ces bases peuvent former des liaisons hydrogènes entre deux brins d'ADN pour former des structures en double hélice communes in vivo (Figure 1.2). L'ADN va constituer l'unité de base du support de l'information génétique. Cette structure va nécessiter une organisation particulière afin de pouvoir être contenue dans le noyau des cellules.

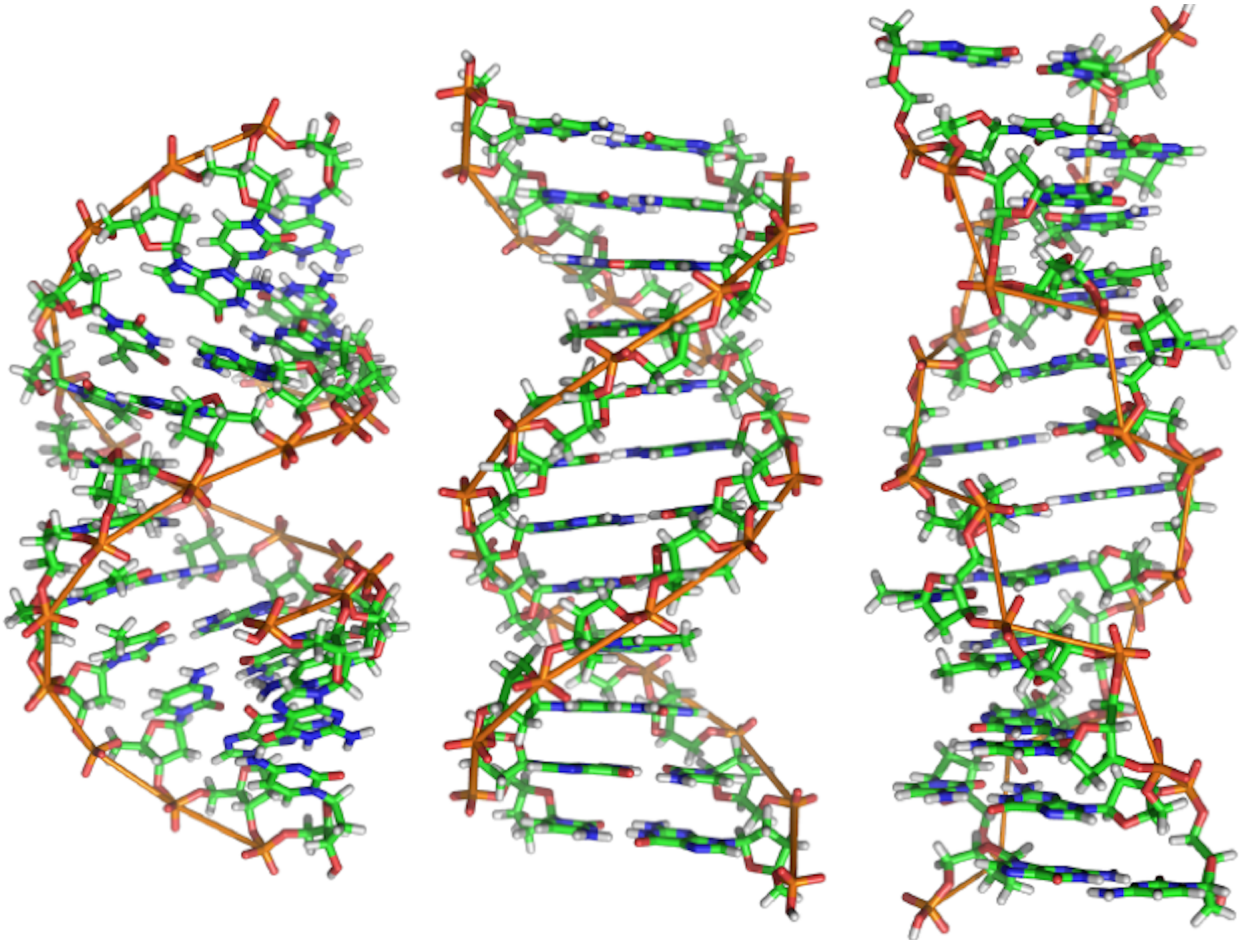


FIGURE 1.2 – Structure de l'ADN en double hélices. Richard Wheeler (Zephyris) / English Wikipedia / CC BY-SA 3.0

Le premier niveau de compaction de l'ADN est appelé chromatine. Elle est composée d'un assemblage d'ADN et de protéines. Pour donner une idée de la compression de l'ADN, le génome humain est composé de 3 milliards de nucléotides et mesure entre 2 et 3 mètres bout à bout [Reece et al., 2012, Manion, 2002] dans un noyau de cellule ne mesurant que quelques micromètres. Ainsi l'ADN va nécessiter une organisation draconienne afin de pouvoir être contenu dans le noyau cellulaire. Pour assurer ce niveau de compaction, l'ADN est enroulé autour de nucléosomes, complexes formés de protéines histones. Le niveau de compactage de la chromatine est un facteur important dans la transcription car il permet de rendre accessible ou non les gènes aux protéines régulatrices.

Le nucléosome est l'unité de base de compaction de l'ADN chez les eucaryotes. C'est un complexe formé d'ADN et d'un "coeur" de protéines. Environ 146 paires de bases d'ADN peuvent s'enrouler autour du "coeur" du nucléosome. L'ADN enroulé autour de ces nucléosomes sont six fois plus compactes qu'un fragment d'ADN nu comportant le même nombre de nucléotides [Kaplan et al., 2009]. Les nucléosomes sont ensuite regroupés par six pour former un solénoïde. Les solénoïdes sont à leurs tours enroulés sur une protéine d'échafaudage (Scaffold proteins), eux-mêmes enroulés pour former la matrice chromosomique (Figure :1.3). C'est la distance entre les nucléosomes qui détermine l'accessibilité à l'ADN. La distance entre nucléosomes n'est pas figée. Elle peut considérablement varier au cours de la vie de la cellule. Cette variation est, en grande partie, due aux histones (protéines composant le nucléosome).

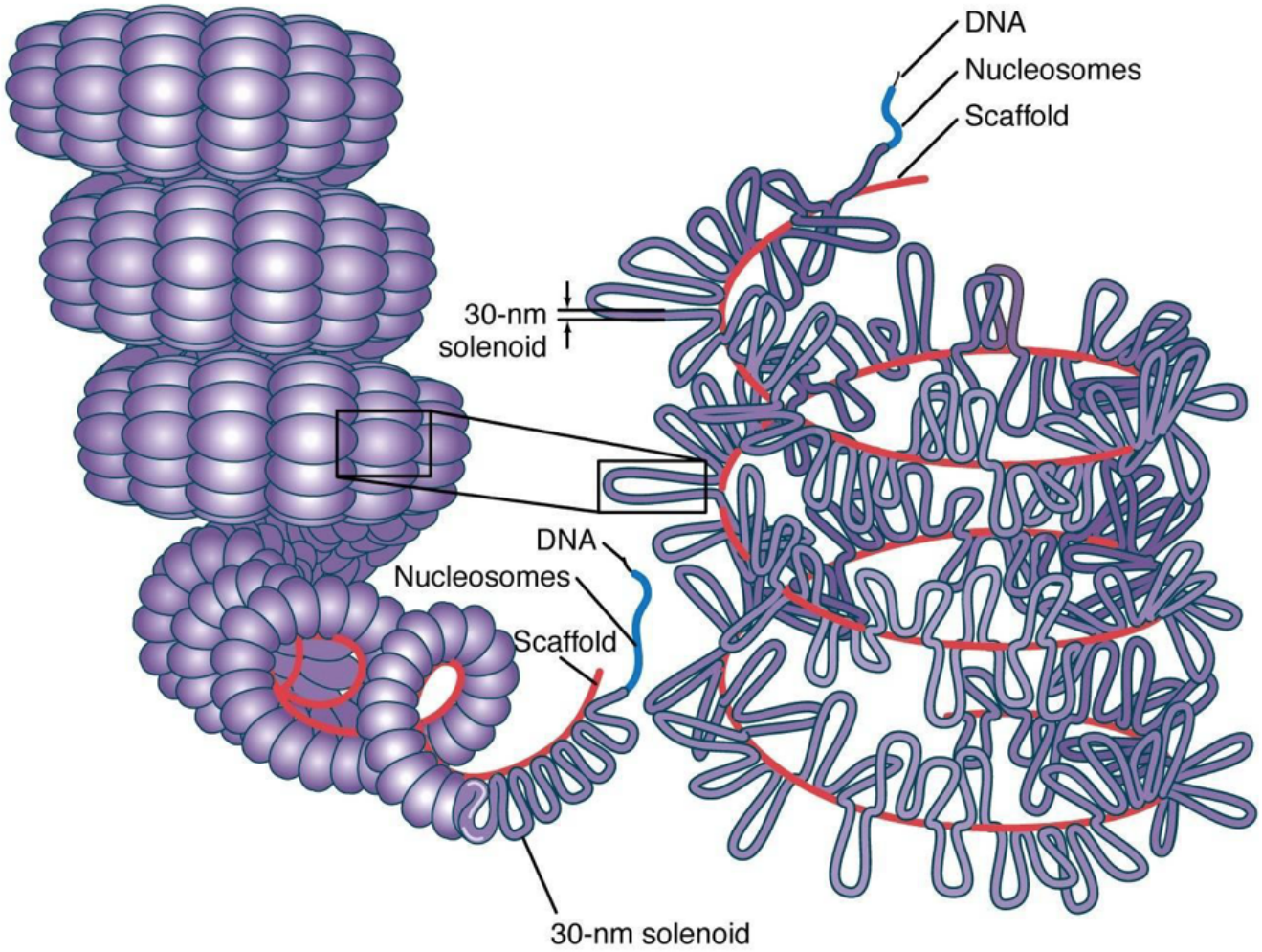


FIGURE 1.3 – L'ADN est enroulé autour de nucléosomes qui forment des brins compacts de chromatine appelés solénoïdes. Les solénoïdes sont maintenus entre eux tous les 30 nanomètres par une structure protéique appelée Scaffold [Griffiths et al., 2000]

.

Les histones sont des protéines qui forment, avec l'ADN, le nucléosome. Les principales histones appartiennent aux classes : H2A, H2B, H3, H4 et constituent le "coeur" du nucléosome. Le nucléosome est composé de deux histones de ces classes formant un octamère (Figure 1.4). Les histones sont des protéines alcalines (charge positive des acides aminés qui la composent), ce qui permet l'interaction avec les molécules d'ADN (le phosphate portant des charges négatives).

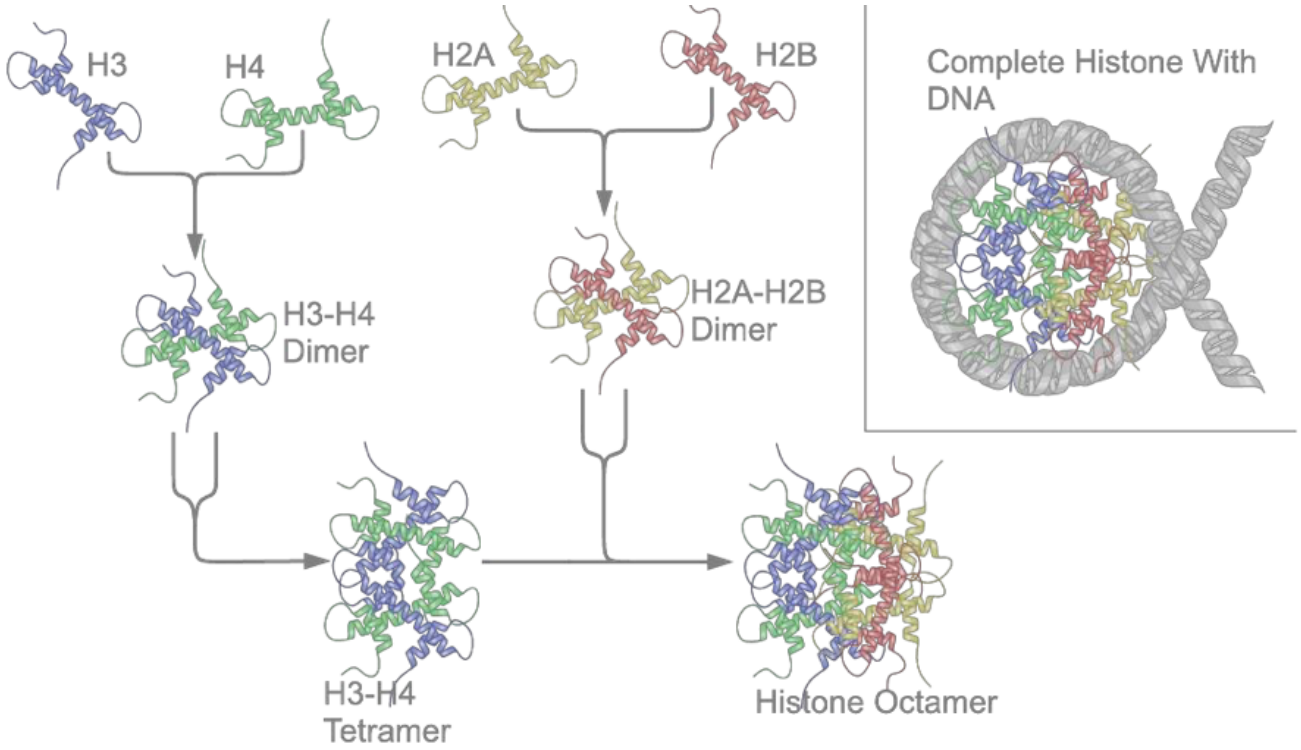


FIGURE 1.4 – Les histones sont formées de 2 hétérotétramères composées chacune de deux dimères de protéines (H3-H4 et H2A-H2B). L'enroulement des molécules d'ADN autour d'un octamère d'histone forme le nucléosome. Richard Wheeler (Zephyris) / English Wikipedia / CC BY-SA 3.0

1.1.1.2 Organisation spatiale de la chromatine

Les nucléosomes successifs s'organisent en une structure de type "collier de perles" de 11nm de diamètre. La dernière classe d'histone, H1, peut venir se fixer sur chaques nucléosomes afin de moduler le niveau de compaction de la chromatine. La chromatine faiblement condensée, appelée euchromatine, constitue un environnement plus favorable à l'expression des gènes. A l'inverse, la chromatine fortement condensée, appelée hétérochromatine, constitue un environnement défavorable car inaccessible. Sous certaines conditions largement débattues, l'hétérochromatine peut être compactée davantage en une fibre ordonnée de 30 nm de diamètre (Figure 1.5). En effet, ces structures sont clairement observées in vitro mais difficilement in vivo [Maeshima et al., 2014].

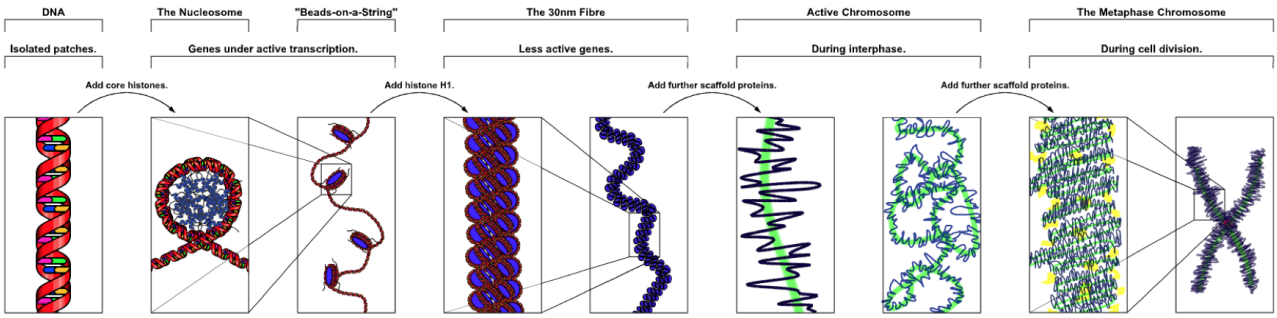


FIGURE 1.5 – Structure de l'ADN en double hélices. Richard Wheeler (Zephyris) / English Wikipedia / CC BY-SA 3.0

Des complexes protéiques non histoniques (CTCF, cohésine) peuvent venir se fixer sur des loci spécifiques afin de structurer la chromatine en territoires appelés "Topologically associating

domains" (TAD). Les différents loci à l'intérieur d'un TAD sont caractérisés par une plus grande fréquence d'interaction entre eux et des interactions minimales avec des loci externes au TAD ([Pombo and Dillon, 2015]).

Certains de ces territoires sont également associés à la lamina nucléaire qui est un réseau de filaments intermédiaire et de protéines associés à la membrane nucléaire structurant la chromatine. A l'échelle la plus large, les chromosomes occupent des positions distinctes dans le noyau, ce qui n'empêche pas de nombreuses interactions entre chromosomes au niveau des territoires frontaliers (Figure 1.6). Au cours de la mitose, la chromatine se restructure avec l'apparition de structures d'ordres supérieures pour former les chromatides de 700nm de diamètres, caractéristiques de la forme chromosomique connue du grand public. L'existence de structures intermédiaires ordonnées présentes dans certains ouvrages est remise en cause par de récents développements en imagerie microscopique. En effet, des chaines désordonnées de chromatine de 5 à 24 nm sont flexibles et peuvent être suffisamment compactées pour atteindre la densité d'un chromosome mitotique ([Ou et al., 2017]).

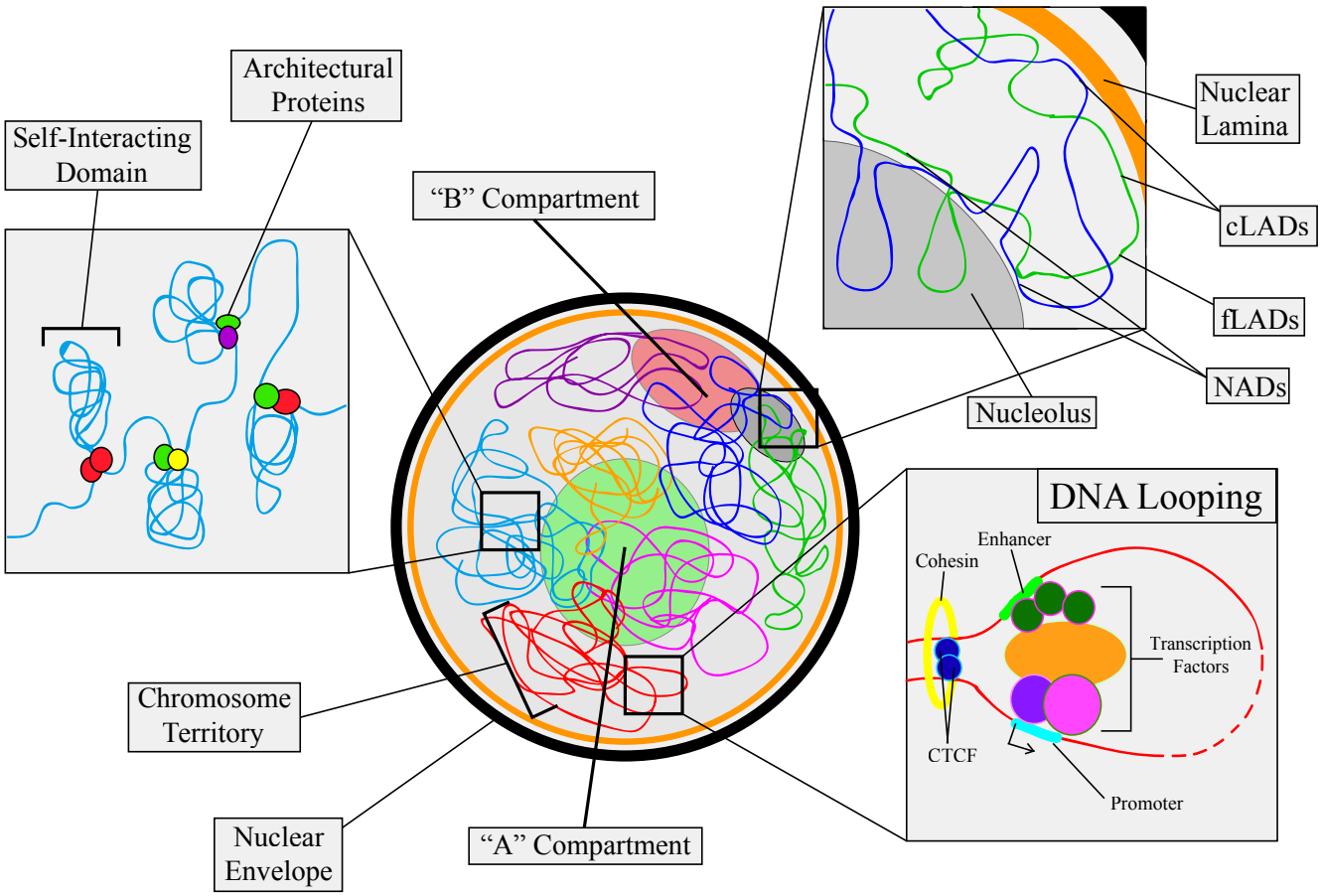


FIGURE 1.6 – Différents niveaux d'organisation spatiale de la chromatine. Evin Wieser / CC BY-SA 4.0

1.1.2 Les éléments génomiques fonctionnels

Dans ce sous chapitre, je présenterais et distinguerais les différents éléments génomiques qu'on peut d'ores et déjà diviser en deux classes :

— Les gènes qui seront transcrit en ARN

— Les éléments cis-regulateurs (CREs). L'épithète « Cis » signifie que la séquence d'ADN régulatrice est capable de moduler l'expression d'un gène présent (en général) sur le même

chromosome. La notion de séquence cis s'oppose à celle de facteurs trans qui désigne quant à elle les facteurs de transcription agissant sur ces séquences cis.

1.1.2.1 Gènes

Dans le sens le plus large possible, un gène est une séquence nucléotidique composé de séquence codante (exon) et non codante (intron) permettant la production d'une molécule ayant une fonction. Cette molécule correspond à une protéine préalablement traduite par depuis ARN messager lui-même issu de la séquence d'ADN du gène après un processus de transcription, épissage et maturation (Figure 1.7). On distingue cependant une grande diversité de gènes dont le transcrit ARN possède directement une fonction. On reconnait aussi l'existence des pseudogènes, des séquences ADN présentant de fortes similarités avec des séquences de gènes codant pour des protéines mais possédant une ou plusieurs mutations perturbant l'un des mécanismes (transcription, épissage, maturation, traduction) permettant d'aboutir à une protéine fonctionnelle.

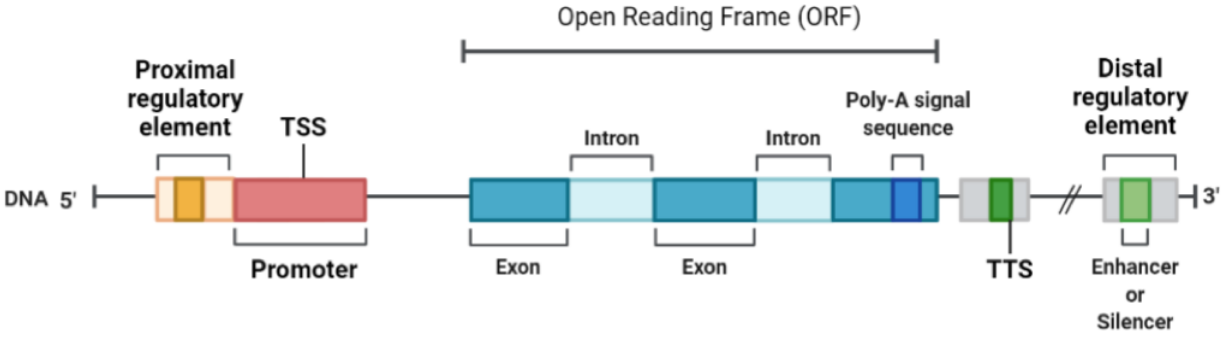
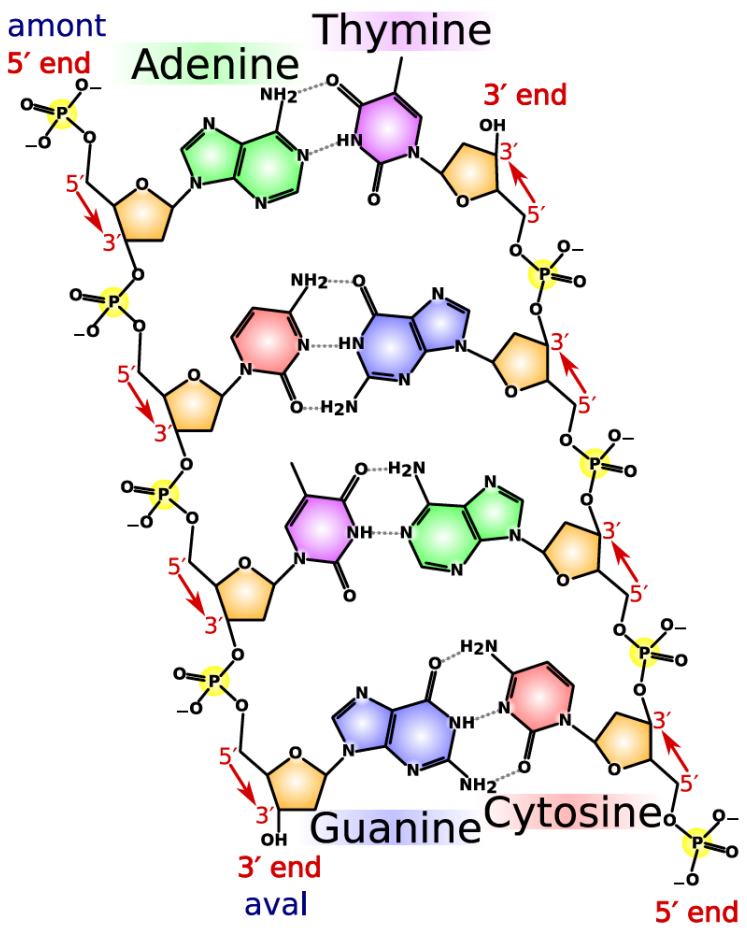
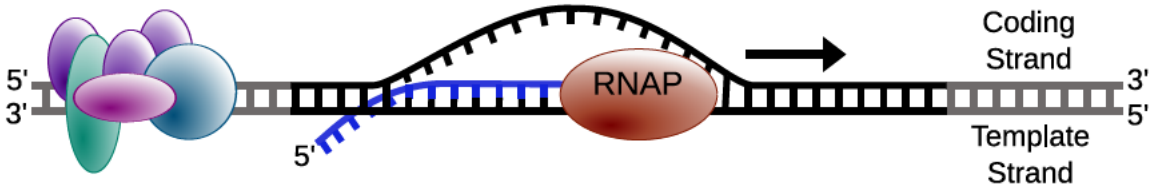


FIGURE 1.7 – Les gènes sont transcrits à partir du TSS jusqu'au site de terminaison de la transcription (Transcription Termination Site, TTS). Le pré-ARNm contient les introns et exons, ainsi que les régions non traduites. Divers éléments peuvent impacter sa régulation.

Chaque carbone du sucre d'un nucléotide est numéroté par convention chimique de 1' à 5', ce qui permet de définir les sens 5'-3' et 3'-5' (Figure 1.8). La terminologie pour la navigation dans le génome se calque sur le sens de synthèse des brins par les différentes polymérases ARN et ADN. Celles-ci se déplacent dans le sens 3'-5' le long du brin matrice et synthétisent donc des brins complémentaires anti-parallèles dans le sens 5'-3' considéré alors comme le sens en-avant ou forward (Figure 1.8). Inversement, le sens 3'-5' devient le sens en-arrière ou reverse. Par rapport à un point de repère sur le génome, l'amont correspond à la région du côté de l'extrémité 5' du brin. Ainsi, la portion d'ARN transcrite mais non traduite dans l'amont d'un gène est appelée 5'-UTR (Untranslated Transcribed Region). On retrouve de la même façon la région 3'-UTR dans l'aval du gène.



(a) Représentation de Cram de nucléotides expliquant la convention de direction



(b) Sens de progression de l'ARN polymérase sur une séquence nucléotidique

FIGURE 1.8 – Représentation du sens de notation de l'ADN

Les gènes sont répartis de façon relativement uniforme sur les deux brins d'ADN de chaque chromosomes [Karki et al., 2018] et environ 10% d'entre eux sont chevauchants avec un autre gène [Sanna et al., 2008]. Afin de pouvoir définir sans ambiguïté l'emplacement d'un gène dans le génome, un brin est appelé "+" et l'autre "-". Chez l'homme, la souris et la plupart des espèces à chromosomes linéaires, le brin "+" correspond par convention au brin dont l'extrémité 5' est la plus proche du centromère. Un gène peut être associé à plusieurs transcrits pouvant différer par leur site d'initiation de la transcription (TSS), les exons retenus après épissage et leur site de terminaison de la transcription (TTS). Différents isoformes d'un transcrit sont en général associées à différents types cellulaires d'un organisme. L'existence de TSS et TTS alternatifs, plus que l'épissage alternatif, sont considérés comme étant les principaux facteurs expliquant cette diversité d'isoformes permettant l'obtention de la majorité des fonctions spécifiques d'un type cellulaire [Reyes and Huber, 2018].

Je vais maintenant m'intéresser aux éléments régulateurs de la transcription des gènes

mentionnés ci-dessus. La régulation des gènes va consister en une série d'activation, répression et modulation de l'expression de chacun des gènes d'une cellule de façon très spécifique. Elle est à la fois qualitative, en déterminant quels gènes doivent être actifs dans quels types cellulaires, et aussi quantitative puisqu'elle définit précisément le niveau d'expression de chacun des gènes.

1.1.2.2 Promoteurs

Les promoteurs sont des éléments cis-régulateurs qui définissent où commence la transcription d'un gène par l'ARN polymérase. Les promoteurs sont situés directement en amont ou à l'extrémité 5' du site d'initiation de la transcription. La région promotrice est composée de deux éléments : la région promotrice centrale (core promoter), qui contient le TSS et la région promotrice proximale, qui contient des sites de liaison spécifiques aux facteurs de transcription. Le promoteur central est défini comme la région minimale d'ADN autour du TSS qui peut induire la transcription (±50 pb). Le promoteur central agit comme une plateforme de liaison pour la RNAPII et facteurs de transcription pour former le PIC (transcription PreInitiation Complex) et initier la transcription. Bien que le core silencer soit suffisant pour démarrer la transcription, son activité est faible. Plusieurs motifs promoteurs centraux ont été identifiés : boîte TATA, îlot CpG, initiateur (Inr), éléments de reconnaissance TFIIB (BRE), MTE (Motif Ten Element) et élément promoteur central en aval (DPE) [Lagrange et al., 1998, Deng and Roberts, 2005, Juven-Gershon and Kadonaga, 2010, Kadonaga, 2012]. Ces motifs, leurs positions relatives par rapport au TSS et les protéines s'y fixant sont résumés dans la table d'une revue écrite par Stark récemment [Haberle and Stark, 2018].

En amont du promoteur, on va retrouver une région qui va permettre la fixation de facteurs de transcription afin d'augmenter le taux de transcription en interagissant avec le PIC. Cette région proximale est associée à une expression tissu-spécifique [Lewis et al., 2005, Smith et al., 2006].

Core-promoter motif	Sequence logo	Consensus sequence[a]	Position relative to TSS	Bound by	Fly	Human
TATA-box		TATAWAWR[49,241]	−31 to −24	TBP[53,242]	+	+
Inr (fly)		TCAGTY[56,243]	−5 to −2	TAF1 and TAF2 (REF.[57])	+	−
Inr (human)		YR[45]	−1 to +1	NA	−	+
		BBCABW[58]	−3 to +3			
DPE		RGWCGTG[59]	+28 to +34	TAF6 and TAF9 (REF.[60]) and possibly TAF1 (REF.[55])	+	Possibly rarely
		RGWYVT[61]	+28 to +33			
		GCGWKCGGTTS[51]	+24 to +32		+	−
MTE		CSARCSSAACGS[63]	+18 to +29	Possibly TAF1 and TAF2 (REF.[55])	+	−
DRE		WATCGATW[245]	−100 to −1	Dref[245]	+	+
TCT		YYCTTTYY[246]	−2 to +6	NA	+	+
BREu		SSRCGCC[64]	−38 to −32	TFIIB[64]	+	+
BREd		RTDKKKK[65]	−23 to −17	TFIIB[65]	+	+
DCEI–DCEIII	NA	CTTC	+6 to +11	TAF1 (REF.[67])	−	+
		CTGT	+16 to +21			
		AGC[66]	+30 to +34			
XCPE1	NA	DSGYGGRASM[247]	−8 to +2	NA	?	+
XCPE2	NA	VCYCRTTRCMY[248]	−9 to +2	NA	?	+
Pause button		KCGRWCG[155]	+25 to +35	NA	+	?

FIGURE 1.9 – Séquences des éléments se fixant sur le core promoter. Adapté de [Haberle and Stark, 2018]

1.1.2.3 Enhancers

Les enhancers sont des séquences cis-régulatrices localisés à distance du TSS et jouent un rôle majeur dans la régulation génique. Le premier enhancers identifié était une séquence répétée en tandem (72pb) dans le virus SV40 ([Banerji et al., 1981]). Plus tard, il a été démontré que cette séquence régulatrice pouvait interagir avec le promoteur du gène de la β-globine, multipliant par 200 la transcription sur une longue distance et peu importe l'orienta-

tion [de Villiers and Schaffner, 1981]. Cette observation a été confirmée chez d'autres virus. Le premier enhancer chez les mammifères a été trouvé dans une région intronique des gènes d'immunoglobulines codant pour les chaînes lourdes et légères [Banerji et al., 1983, Gillies et al., 1983, Queen and Baltimore, 1983]. Depuis, plusieurs enhancers dans différentes cellules, tissus et conditions ont été identifiés et étudiés. Aujourd'hui, le Consortium Encyclopedia of DNA Elements (ENCODE) a identifié des centaines de milliers de loci d'enhancers dans le génome humain via des jeux de données génomiques (Consortium, 2012). Ces derniers suggèrent que plusieurs enhancers peuvent réguler un seul gène et que plusieurs gènes peuvent être régulés par un seul enhancer [Medina-Rivera et al., 2018]. Les enhancers peuvent agir sur leurs promoteurs cibles jusqu'à 1 Mb, quelle que soit l'orientation relative de la séquence et ne régulent pas nécessairement les gènes les plus proches ([Santiago-Algarra et al., 2017]). La séquence de l'enhancer abrite une haute densité de facteurs de transcription de liaison à l'ADN (Figure 1.10) qui varient d'une cellule à l'autre et sont normalement spécifiques à la lignée ; rendant la fonction enhancer cellule-dépendante ([Villiers et al., 1982]).

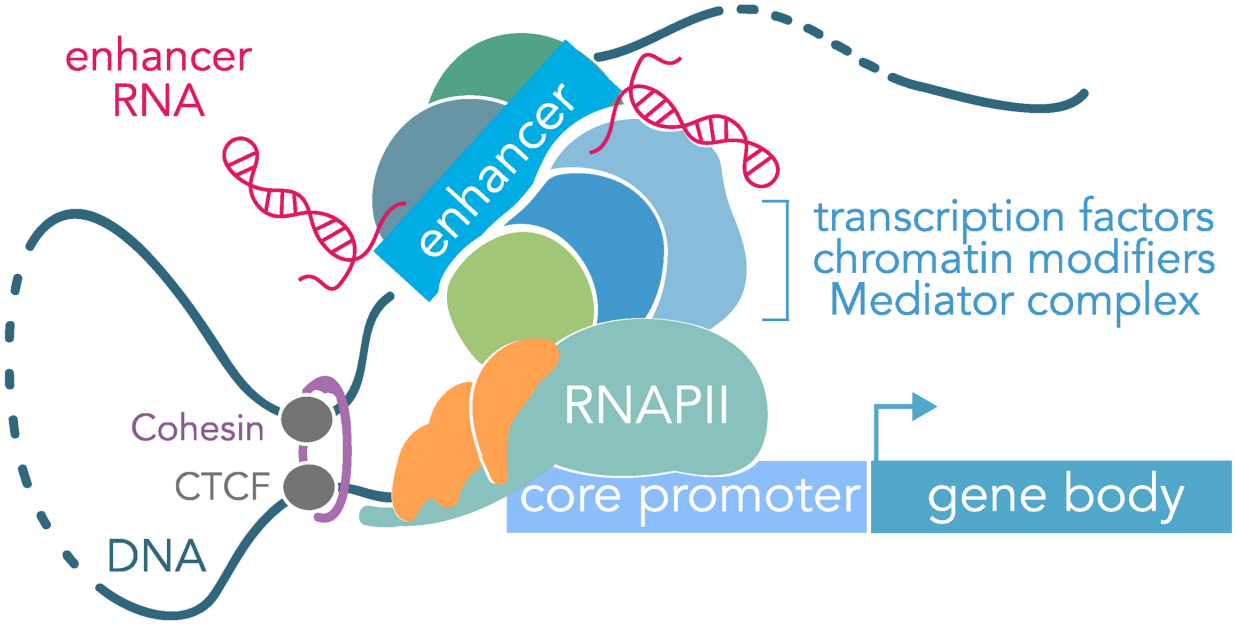


FIGURE 1.10 – Illustration de la boucle chromosomique enhancer-promoteur (médiée par le CTCF et la cohésine) qui permet aux éléments enhancers distaux d'interagir physiquement et d'activer le promoteur du gène cible. Ces interactions augmentent la liaison des facteurs de transcription, des modificateurs de la chromatine et du complexe Mediator au niveau des promoteurs de gènes pour recruter l'ARN polymérase II (RNAPII). Les enhancers sont caractérisés par une chromatine ouverte, des motifs de liaison au facteurs de transcription, des marques d'histones spécifique, une hypométhylation de l'ADN et une transcription bidirectionnelle pour générer des ARN activateurs (ARNe) [Carullo and Day, 2019].

Une autre classe d'enhancer a été décrite comme "super-enhancer". Cet élément régulateur partage de nombreuses caractéristiques typiques des enhancers, mais à plus grande échelle : en moyenne, ils sont 15 fois plus gros (plus de 19 Kb), recrutent plus de protéines médiatrices 1 (protein mediator 1) (18 fois) et BDR4 (16 fois), et ont un niveau plus élevé de H3K27ac (26 fois) et H3K4me1 (10 fois) [Lovén et al., 2013, Whyte et al., 2013]. Les super-enhancers sont densément occupés par des facteurs de transcription maîtres et des complexes médiateurs. Ces facteurs de transcription maîtres établissent des réseaux d'autorégulation qui ne se répartissent pas uniformément mais forment des grappes denses de facteurs de

transcription appelés « Hot points » [Siersbæk et al., 2014]. Les super-enhancers jouent un rôle dans la réorganisation des interactions des promoteurs dans la différenciation cellulaire [Madsen et al., 2020, Novo et al., 2018, Siersbæk et al., 2014], l'expression des gènes circadiens [Kim et al., 2018] et la réponse au stress [Schmidt et al., 2015].

1.1.2.4 Epromoteurs

Les Epromoteurs sont une nouvelle classe d'éléments cis-régulateurs définis comme un élément promoteur qui affiche une activité enhancer dans un cadre expérimental fonctionnel et pourrait réguler l'expression d'un gène distant. Ces éléments ont été validés expérimentalement pour la première fois dans notre laboratoire. Ces interactions promoteur-promoteur ont d'abord été identifiées par des méthodes basées sur le 3C, suggérant que ce réseau de régulation est courant dans les cellules de mammifères [Pancaldi et al., 2016]. L'utilisation de diverses techniques intégrant le séquençage à haut débit dans le test de gènes rapporteurs a permis des mesures quantitatives et simples de l'activité des enhancers. Des Epromoteurs ont été identifiés dans certaines études *in vitro* chez la drosophile [Zabidi et al., 2015], la souris [Nguyen et al., 2016] et l'homme. Parmi ces études, certaines ont validé l'activité fonctionnellement *in vivo* [Dao et al., 2017, Diao et al., 2017, Engreitz et al., 2016] (Figure 1.11).

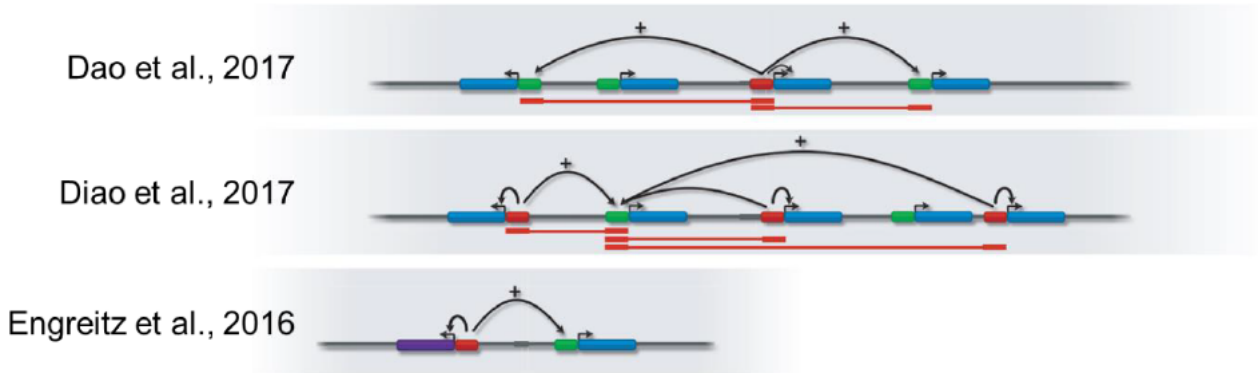


FIGURE 1.11 – Un ou plusieurs Epromoteurs (rouge) peuvent réguler l'expression de un ou plusieurs gènes distaux. Adapté de [Medina-Rivera et al., 2018]

1.1.2.5 Silencers

Les silencers sont des éléments régulateurs ayant un effet inhibiteur sur la transcription des gènes, ce qui signifie qu'ils ont une activité répressive. En tant que régulateurs, ils fonctionnent indépendamment de la distance et/ou de l'orientation par rapport aux gènes cibles. Ils peuvent faire partie des promoteurs proximaux, d'enhancers distaux, ou peuvent être considérés comme des régions régulatrices distales indépendantes. Ils peuvent être localisés loin de leurs gènes cibles, dans une région intronique ou dans la région 3' non traduite [Harris et al., 2005, Sertil et al., 2003]. Les silencers vont servir de point d'ancrage aux facteurs de transcription répresseurs. Ces derniers peuvent également recruter des cofacteurs possédant une activité inhibitrice appelés corépresseurs ([Privalsky, 2004]).

Ayant consacré une grande partie de ma thèse à leurs caractérisations et compréhension, une section leur est dédiée (Section : 1.2.1).

1.1.2.6 Insulator

Un isolateur(insulator) est un élément cis-régulateur qui joue le rôle de barrière séparatrice entre sa région en amont et en aval. Un enhancer actif n'a pas d'influence sur les gènes dont il

est séparé physiquement par un isolateur. Certains isolateurs séparent également des régions condensées et décondensées de la chromatine (Figure 1.12, [Gaszner and Felsenfeld, 2006]. Deux isolateurs successifs sur une séquence ADN, intercalés par un groupe de gènes constituent un TAD dont les propriétés spatiales ont été évoquées dans la section 1.1.1.2.

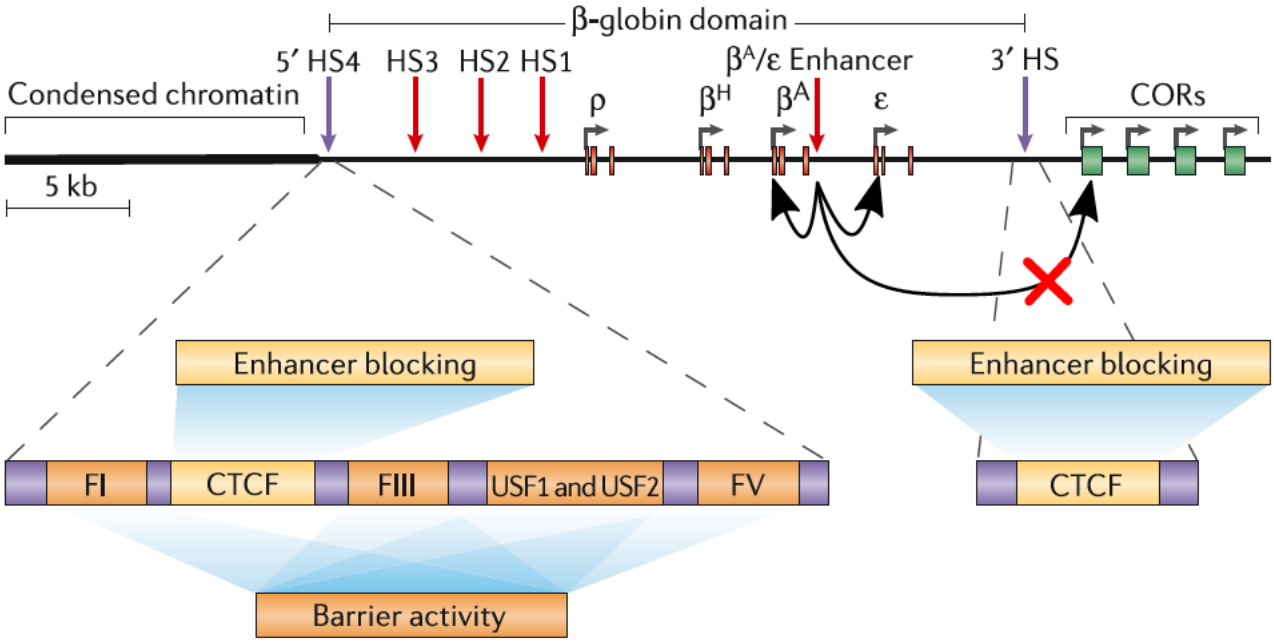


FIGURE 1.12 – Locus de la β-globine du poulet. Chez le poulet (Gallus gallus), les isolateurs 5'HS4 et 3'HS, indiqués par des flèches violettes définissent les limites d'un domaine chromatinien qui contient le cluster de gènes de la β-globine. Les enhancers du cluster sont indiqués par des flèches rouges. L'action de ces derniers est restreinte aux gènes du cluster par un mécanisme médié par la présence de CTCF sur les isolateurs. Des éléments complémentaires (FI, FIII, USF1/USF2 et FV) sur l'élément HS4 protègent le cluster du mécanisme de condensation de la chromatine ayant lieu dans la région amont. Adapté de [Gaszner and Felsenfeld, 2006].

1.1.2.7 Long Non-coding RNA - LncRNA

Moins de 2% du génome humain est codant, tandis qu'environ les deux tiers du génome sont transcrits en ARN non codants (ARNnc). Parmi tous les ARNnc, ceux qui dépassent 200 nucléotides de long sont nommés ARN longs non codants (LncRNA), cela inclut les ARNe et les transcrits antisens (Figure 1.13) ([Derrien et al., 2012, Mathieu et al., 2014]). En moyenne, les lncRNAs sont plus courts que les ARNm (592 bp vs 2453 bp) [Derrien et al., 2012]. Plusieurs mécanismes ont été proposés pour expliquer comment les lncRNAs peuvent réguler la transcription en cis ou en trans [Kopp and Mendell, 2018]. Les mécanismes proposés en cis incluent (i) l'interaction spécifique à la séquence entre lncRNA mature et gène, (ii) le recrutement de la machinerie du splicéosome, ou (iii) le promoteur du lncRNA agit comme un enhancer et interagit avec la région promotrice du gène codant pour la protéine, indépendamment du lncRNA. En plus de la régulation en cis, les mécanismes décrits pour la trans-régulation sont catégorisés en trois groupes : (i) les lncRNA qui peuvent réorganiser l'état de la chromatine dans des régions éloignées, (ii) les lncRNA qui s'associent à l'ARNm ou aux protéines affectant son interaction, et (iii) des lncARN qui modifient la structure nucléaire.

On sait maintenant que les lncRNA peuvent réguler l'expression des gènes au niveau transcriptionnel [Ng et al., 2012, Rinn et al., 2007, Zhao et al., 2006] ou post-transcriptionnel

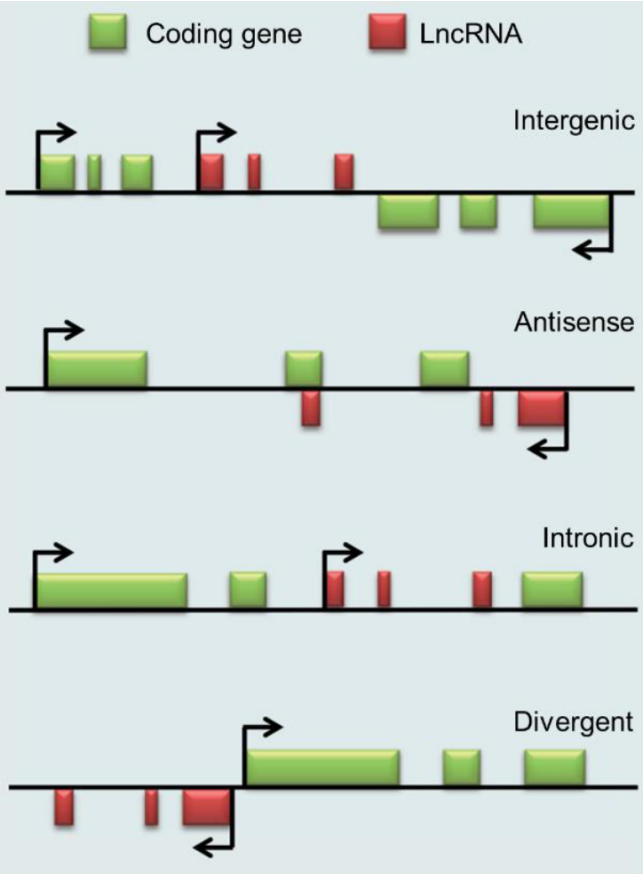


Figure 1.13 – Illustration des différentes classes de lncRNA. Adapté de [Mathieu et al., 2014]

[Cesana et al., 2011, Gong and Maquat, 2011, Yoon et al., 2012], de plus, il a été démontré qu'ils peuvent également contribuer au développement de maladies [DiStefano, 2018, Sparber et al., 2019].

1.1.2.8 Les éléments répétés du génome

Environ 80% du génome humain est non codant. Dans cette sous-section je citerais brièvement quelques éléments non codants. Ils possèdent une fonction indépendante d'une quelconque transcription menant à un produit fonctionnel. Ces régions sont principalement des éléments répétés.

Ces éléments ont été décrits pour la première fois en 1968 par Britten and Kohne [Britten and Kohne, 1968] Pendant longtemps ces structures sont restées incomprise, souvent considérées comme sans importance, voire indésirable faisant référence à de la Junk DNA. Aujourd'hui, grâce aux différents progrès techniques, aux différents génomes séquencés, et aux nombreuses découvertes, on commence à comprendre l'importance de ces séquences répétées dans notre génome. Ils ont ensuite été classés selon des critères tels qu'une répétition successive (en tandem par exemple) ou des répétitions à travers le génome (Figure 1.14). Une base de données reprenant les différents éléments répétés chez les eucaryotes existe et est régulièrement mise à jour [Jurka, 2000, Jurka et al., 2005, Bao et al., 2015] .

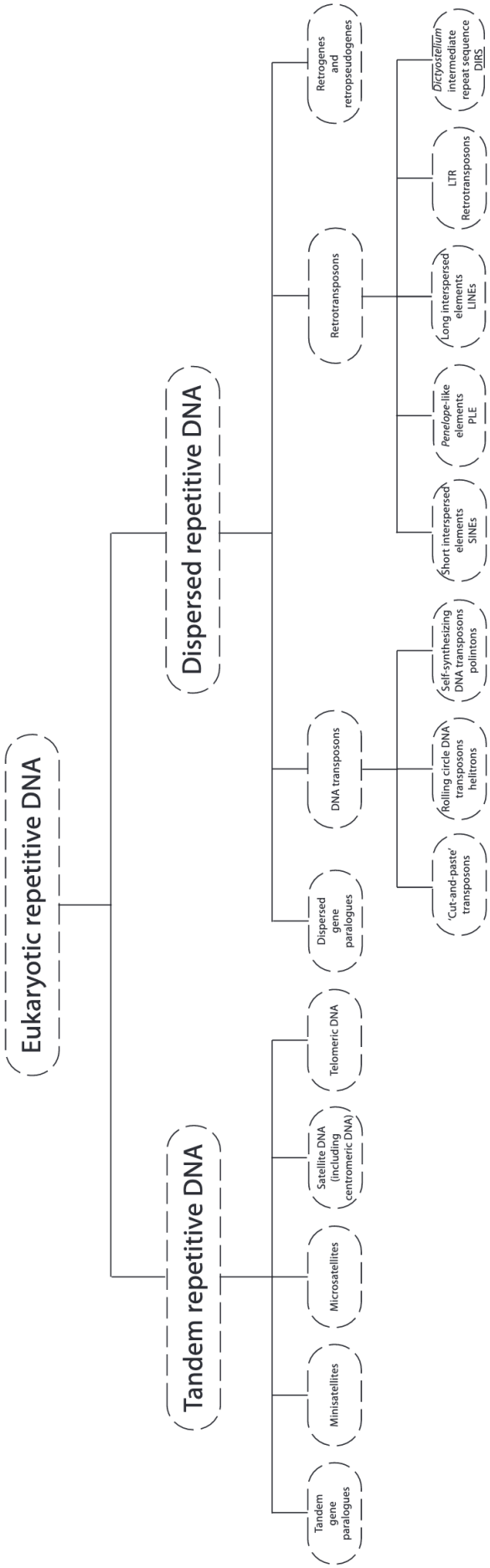


FIGURE 1.14 – Classification des éléments répétés en fonction du type de répétition, successive ou éparse [López-Flores and Garrido-Ramos, 2012].

1.1.2.8.1 Motifs répétés successivement : Ces éléments sont répétés successivement dans des fenêtres qui diffères en fonction de leurs classes. Parmi ces éléments on compte les télomères, centromères ou encore des séquences simples répétées en tandem.

Short Tandem Repeats (STR) : Connu également sous le nom de microsatellite ou séquences simples répétées en tandem – Simple Tandem Repeat. Ce sont des séquences répétées en tandem qui font moins de 9 nucléotides. Le nombre moyen de répétition en tandem de ces STR est de 5 à 40 répétitions dans une fourchette de 1kB [López-Flores and Garrido-Ramos, 2012]. Ces séquences peuvent apparaitre au sein de régions régulatrices. Elles sont hautement instables car très sujettes aux mutations [Gemayel et al., 2010]. Les dinucléotides sont le type le plus important de répétitions microsatellites pour de nombreuses espèces. Chez l'homme, le type de répétition de dinucléotide le plus courant est (CA)n/(GT)n [López-Flores and Garrido-Ramos, 2012]. Les répétitions avec des unités contenant 3 ou 6 nucléotides (répétitions trinucléotidiques et hexanucléotidiques) sont les plus abondants dans les régions codantes, probablement parce qu'ils ne provoquent pas de décalage dans le cadre de lecture. Via des données CAGE (Cap Analysis of Gene Expression) et en utilisant des algorithmes de Deep Learning, il a été démontré qu'il y avait une initiation de la transcription dans ces régions non codante. Ceci a été mis en évidence notamment grâce aux séquences flanquantes des STR révélant l'intérêt des régions entourant les microsatellites [Grapotte et al., 2021]. Il a été démontré qu'elles pouvaient avoir un rôle régulateur, par exemple une répétition en tandem (STR) du motif CAGGTG, qui est la séquence reconnue pour la fixation du facteur de transcription répresseur ZEB1 impliqué dans le maintien des caractéristiques mésenchymateuses, allait être nécessaire au maintien de l'identité épithélial et qu'une délétion du STR allait entrainer la cellule dans un état cancéreux [Balestrieri et al., 2018]. Ce qui démontre que la forte densité de motifs identiques dans les STR peut en faire des sites appropriés pour le recrutement des répresseurs transcriptionnels. De plus, de par leurs hautes instabilités les microsatellites (MSI) vont être à l'origine de processus tumorigénèse. Les tumeurs MSI, principalement observées dans les cancers du côlon, de l'estomac et de l'endomètre, se développent selon une voie moléculaire distincte caractérisée par une accumulation de mutations au niveau des STRs [Hause et al., 2016, Cortes-Ciriano et al., 2017, Lupinacci et al., 2018, Touat et al., 2020].

Les satellites : Ce sont des séquences d'ADN très répétitives et qui constituent une part considérable des génomes eucaryotes. Les monomères (unités répétées) sont répétés en tandem, généralement de plus de 200 nucléotides de longueur et généralement organisées dans des régions de centaines ou de milliers d'exemplaires qui occupent jusqu'à plusieurs mégabases au sein des génomes. Les satellites composent majoritairement l'hétérochromatine.

Origines de réplication : Les origines de réplication correspondent aux régions du génome sur lesquelles se fixent initialement le complexe protéique responsable de la réplication de l'ADN. Entre 30 000 et 100 000 origines de réplications ont été identifiées chez l'homme et la souris ([Leonard and Méchali, 2013]). Leur taille est comprise entre 100 et 1000pb et chez les eucaryotes, la majorité de ces régions contiennent des séquences répétées, riches en guanine capable de former des G-quadruplex. ([Cayrou et al., 2015]).

Centromère : Les centromères correspondent à la région de contact entre les deux chromatides des chromosomes. Chez l'homme et la souris, ces régions sont particulièrement longues (plusieurs MB) et sont constituées de séquences hautement répétées espèce-spécifiques. Une fonction particulière associée aux centromères est l'interaction avec les kinétochores formés lors de la mitose [Barra and Fachinetti, 2018].

Télomère : Les extrémités de chaque chromosome, ou télomères, sont constituées de séquences répétées riches en guanine propices à la formation de G-quadruplex [Wang et al., 2011]. En l'absence de mécanismes compensatoires, l'extrémité des télomères s'érodent à chaque

cycle de réplication à cause d'imperfections dans le mécanisme de réplication par l'ADN polymérase. Les télomères assurent le rôle de tampon vis-à-vis des gènes périphériques et leurs structures particulières permet l'action de la télomèrase, une enzyme ralentissant le processus d'érosion en exercant une action inverse.

Comme nous l'avons dit précédemment, les séquences peuvent être répétées successivement ou à différents endroits dans le génome. Nous nous intéresserons maintenant aux éléments répétés dispersés. Ces éléments sont divisés en deux classes qui diffèrent de par leurs mécanismes. Ils sont constitués d'ADN transposable et de retrotransposons (contenant entre autres les LINEs, SINEs, LTRs).

1.1.2.8.2 Les éléments transposables - TE Ces éléments sont divisés en deux classes qui diffèrent de par leurs mécanismes. Ils sont constitués d'ADN transposable et de retrotransposons. Tous les rétrotransposons eucaryotes actuellement connus peuvent être divisés en 5 types ou selon une classification proposé par Wicker et al. ([Wicker et al., 2007]) que l'on peut voir dans les figures 1.14 et 1.15 : (1) LTR rétrotransposons ; (2) Séquence de répétition intermédiaire de Dictyostelium (DIRS) rétrotransposons ; (3) rétrotransposons non-LTR ou LINEs ; (4) Rétrotransposons de type pénélope (PLE) ; et (5) SINEs. Les rétrotransposons LTR et non-LTR sont les plus répandus et abondants chez les eucaryotes.

Les TEs sont des séquences d'ADN capables de se déplacer et se dupliquer de manière autonome dans un génome, par un mécanisme appelé transposition. Ils représentent une grande proportion du génome humain (environ 45% du génome [Lander et al., 2001]). Leur mobilité est source de mutations, et par conséquent de diversité mais aussi de maladies génétiques. Ils peuvent ainsi être considérés comme des parasites génétiques à l'échelle d'un individu, ou des moteurs de l'évolution à l'échelle d'une espèce [Bourque et al., 2018]. En effet, il a été estimé que la majorité des séquences cis-regulatrices nouvellement acquises au cours de l'évolution des primates sont directement dérivés des éléments transposables [Trizzino et al., 2017]. D'un point de vue de l'évolution des organismes, il a été démontré que les éléments transposables sont en partie contrôlés par des complexes répresseurs comme le complexe HUSH, que nous présenterons plus tard afin de contrôler leurs expansions dans le génome. Il a été démontré que les éléments transposables allaient jouer un rôle dans la régulation de la transcription. Ils vont pouvoir jouer un rôle au niveau de la transcription [Chuong et al., 2017, Sundaram and Wysocka, 2020], des modifications post-transcriptionnels ,[Drongitis et al., 2019], et vont même pouvoir agir sur la configuration de la chromatine [Cournac et al., 2016, Wang et al., 2015]. Ainsi, on comprend aisément qu'ils vont pouvoir influencer la différenciation cellulaire. En effet, il a été démontré qu'une inhibition de séquence rétrovirale endogène (ERV) allait être nécessaire à la détermination des cellules lymphoblastiques de type T [Adoue et al., 2019].

Parmi ces éléments transposables on va brièvement présenter :

LTR : Les LTR contiennent des séquences promotrices et des caractéristiques associées à la transcription. Les rétrotransposons LTR sont transcrits à partir du promoteur dans les 5' LTR. Les rétrotransposons LTR sont similaires aux rétrovirus à l'exception de l'absence du gène *env* dans la plupart des éléments. Ils contiennent un gène *gag*, qui code une protéine de liaison aux acides nucléiques qui pourrait être impliquée dans la transcription inverse, et un gène *pol* qui code pour différents domaines enzymatiques : protéinase, transcriptase inverse, RNase H et intégrase. Il a été démontré que les LTR pouvaient avoir un rôle régulateur, par exemple : dans l'embryon de souris, les éléments MERVL servent de promoteurs alternatifs pour un sous-ensemble de gènes de souris [Macfarlan et al., 2012], tandis que les LTR d'un ERV humain, MER41, peuvent fonctionner comme enhancers inductibles à l'interféron [Chuong et al., 2016].

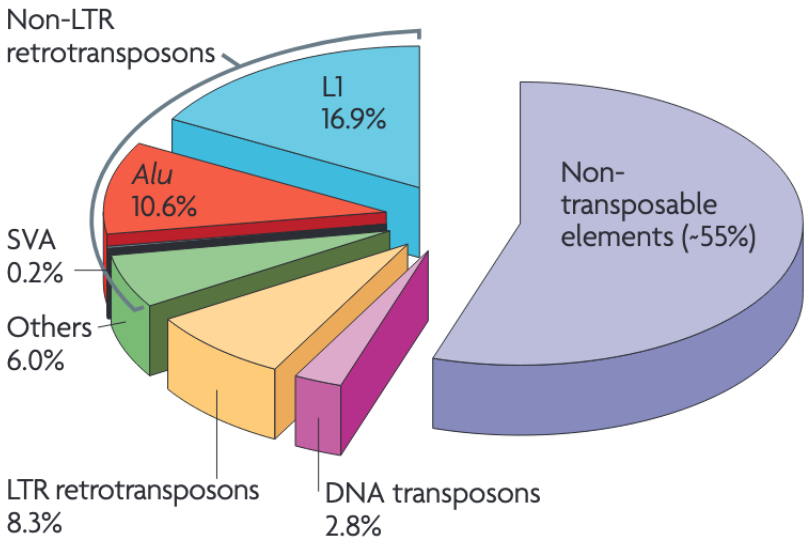


Figure 1.15 – Répartition des éléments transposable au sein des eucaryotes. [Cordaux and Batzer, 2009].

LINE : L'élément transposable le plus représentatif et le plus étudié de ce type de rétrotransposons est le LINE-1 (L1), un élément de 6 à 8 kb, qui chez l'homme atteint jusqu'à 516,000 copies et représente environ 17% de son génome [Cordaux and Batzer, 2009]. La plupart des éléments LINEs présent dans le génome humain sont des séquences de rétrovirus endogènes.

SINE : Les insertions SINEs (ou LINEs) sont utilisées comme marqueurs phylogénétique. En effet, on suppose que tous les organismes porteurs d'une insertion particulière sont dérivés d'un événement irréversible unique qui s'est produit chez leur ancêtre commun [Cordaux and Batzer, 2009]. Les SINEs sont issus de la transcription inverse de la RNA Pol III et pourraient avoir joué un rôle clé dans l'amélioration du potentiel évolutif de leurs hôtes. L'exaptation d'une nouvelle fonction à partir de séquences d'ADN auparavant inutiles, s'est produit de manière récurrente pour les SINEs au cours de l'évolution, générant de nouveaux éléments cis-régulateurs pour des processus tels que l'épissage alternatif, la polyadénylation d'ARNm et l'activité de promoteur [Okada et al., 2010]. Certains enhancers seraient également issus des SINEs et ont été impliqués dans la différenciation des espèces notamment la région cérébrale chez les mammifères [Okada et al., 2010]. Les SINEs vont pouvoir affecter l'expression de diverses manières telle que par l'insertion intronique. En effet, ils vont pouvoir interférer avec les mécanismes d'épissages alternatifs et ainsi induire un saut d'exon. Ils vont également être impliqué dans des processus pathogéniques. Une récente étude GWAS (Genome-Wide Association Study) indique qu'un haplotype CD58 intronique contenant une Alu (famille d'élément SINEs) est associée à un risque accru de développer une sclérose en plaques [Payer et al., 2019, Payer and Burns, 2019].

1.1.3 Expression et régulation génique

Comme nous l'avons vu, l'ADN est le support de l'information génétique. Chaque cellule contient un noyau contenant le même génome. Mais si toutes les cellules contiennent les mêmes séquences d'ADN, comment obtient-t-on tant de tissus cellulaires différents ? Quels processus vont engager une cellule dans une voie de différenciation cellulaire plutôt qu'une autre ?

C'est grâce à la régulation génique et l'expression ponctuelle de gènes que nos cellules vont pouvoir se différencier, s'adapter, et maintenir une homéostasie. C'est grâce aux différents éléments régulateurs mentionnés ci-dessus que la cellule va pouvoir maintenir un niveau de régulation aussi fin. Je commencerais par décrire comment un gène est exprimé et par la suite

quels éléments pourront interagir pour sa régulation.

1.1.3.1 Expression génique

La transcription est le processus par lequel un gène est exprimé, produisant un pré-messager ARN (pré-ARNm). Lors de la transcription, l'ARN polymérase II (ARN Pol II) va se lier sur l'ADN, l'ouvrir et transcrire le brin sens (aussi appelé codant) en pré-ARNm.

La transcription commence par le recrutement du Pol II dans le PIC, lui-même composé de régulateurs transcriptionnels. Ce recrutement peut se faire sur la TATA box. Le résultat du processus de transcription est le pré-ARNm qui sera ensuite « maturé ». C'est-à-dire que les introns ne sont pas retenus, une queue polyA est ajoutée à son extrémité 3 ' et une coiffe méthylé est ajoutée en 5'. L'ARNm est ensuite exporté dans le cytoplasme pour la traduction en protéine.

L'existence du PIC en tant que complexe de plusieurs facteurs est le premier élément qui nous prouve que le processus de transcription, et par extension l'expression des gènes, est régulé par plusieurs acteurs.

En effet, il existe également des régulateurs, dont la liaison est beaucoup moins prévisible et dépend d'autres facteurs discutés ci-dessous.

1.1.3.2 La régulation génique

La régulation génique fait partie intégrante du développement des organismes en permettant aux cellules de se différencier. En effet, une cellule va pouvoir passer d'un état indifférencié à un état différencié en fonction des gènes qu'elle exprime. La succession des changements dans l'expression des gènes au cours de la vie d'une cellule va lui permettre de s'engager dans une voie particulière de différenciation et ainsi acquérir, au fur et à mesure, sa morphologie et sa fonction finale.

Cette différenciation peut également être observée à une échelle beaucoup plus fine pour mettre en évidence des changements beaucoup plus subtils au niveau cellulaire. Par exemple, dans les premiers stades de différenciation des cellules souches hématopoïétiques (HSC, Hematopoietic Stem Cell), des changements dans l'expression des gènes vont permettre aux cellules d'exprimer ou de réprimer des marqueurs spécifiques à leur surface [Chen et al., 2014]. Les cellules HSC vont alors se différencier et acquérir au fur et à mesure des propriétés fonctionnelles spécifiques afin de constituer finalement l'ensemble du système hématopoïétique.

Ainsi, la régulation des gènes est un processus essentiel permettant aux cellules de s'engager dans des voies de différenciation très diverses et complémentaires, et ainsi de participer au développement et au fonctionnement des organismes. Les études sur la reprogrammation cellulaire soulignent également le rôle capital de la régulation des gènes et des facteurs de transcription dans la spécificité des cellules.

1.1.3.3 Les régulateurs transcriptionnels

Les facteurs de transcription (TF) sont des protéines capables de lier l'ADN d'une manière spécifique aux séquences appelées sites de liaison et de réguler la transcription ([Vaquerizas et al., 2009]). Tous les TF contiennent un domaine de liaison à l'ADN (DBD) qui reconnaît des séquences spécifiques. Leurs activités détermine le fonctionnement et la réponse des cellules aux environnements cellulaires.

Il existe des méthodes pour caractériser les liaisons TF-ADN *in vivo* et *in vitro*. Les approches *in vivo* peuvent révéler des événements de liaison au TF dans une condition biologique particulière (par exemple, type de cellule, traitement, cinétique), tandis que les méthodes *in vitro* sont bien adaptées à la caractérisation à grande échelle des mécanismes de fixations des

TF [Inukai et al., 2017]. La combinaison des approches *in vivo* et *in vitro* a mis en lumière les caractéristiques qui influencent l'association TF-ADN (Figure 1.16).

La plupart des régulateurs transcriptionnels (TR) ne régulent pas la transcription génomique à eux seul. Certains d'entre eux possèdent des domaines d'activation sur lesquels d'autres régulateurs, appelés cofacteurs, peuvent se lier une fois eux-mêmes liés au génome. On peut prendre en exemple les complexes régulateurs dont nous avons parlé lors de la présentation des éléments cis-régulateurs. De tels complexes peuvent remplir de nombreuses fonctions, du remodelage de la chromatine à la stimulation de la transcription.

Cette coopération peut être échelonnée avec premièrement des facteurs pionnier qui vont venir se fixer pour ouvrir la chromatine. Cependant, l'aspect de la temporalité est très délicat à étudier et démontrer. Nous nous concentrons surtout sur des complexes obtenus par colocalisation de TF à un instant t, obtenus expérimentalement. En effet, dans la plupart des cas, une observation de la colocalisation des TRs est le résultat d'une coopération entre eux [Biggar and Crabtree, 2001]. L'impact de la coopération des régulateurs transcriptionnels peut être linéaire, avec une proportionnalité entre le nombre de facteurs et la réponse transcriptionnelle [Giorgetti et al., 2010] ou il peut se comporter comme un interrupteur marche-arrêt pour la transcription [Chopra and Levine, 2009].

Dans tous les cas, de telles combinaisons sont responsables du bon fonctionnement des régions de régulation décrites ci-dessus. Plusieurs exemples peuvent être donnés. Un classique est la coopération entre CTCF et RAD21 pour former une boucle de cohésine délimitant les TADs présentés ci-dessus [Stedman et al., 2008]. Dans un exemple qui ne nécessite pas d'interaction protéine-protéine, FOXA1 est un facteur pionnier permettant la fixation ultérieure de ESR1 [Ross-Innes et al., 2012], tandis que FOXA1 est lui-même une cible de GATA3 [Kouros-Mehr et al., 2006].

Certains régulateurs transcriptionnels sont appelés régulateurs maîtres, car ils recrutent et/ou influencent l'activité de nombreux autres régulateurs transcriptionnels sur de nombreux sites différents du génome (The ENCODE Consortium 2012, [Yang et al., 2004]).

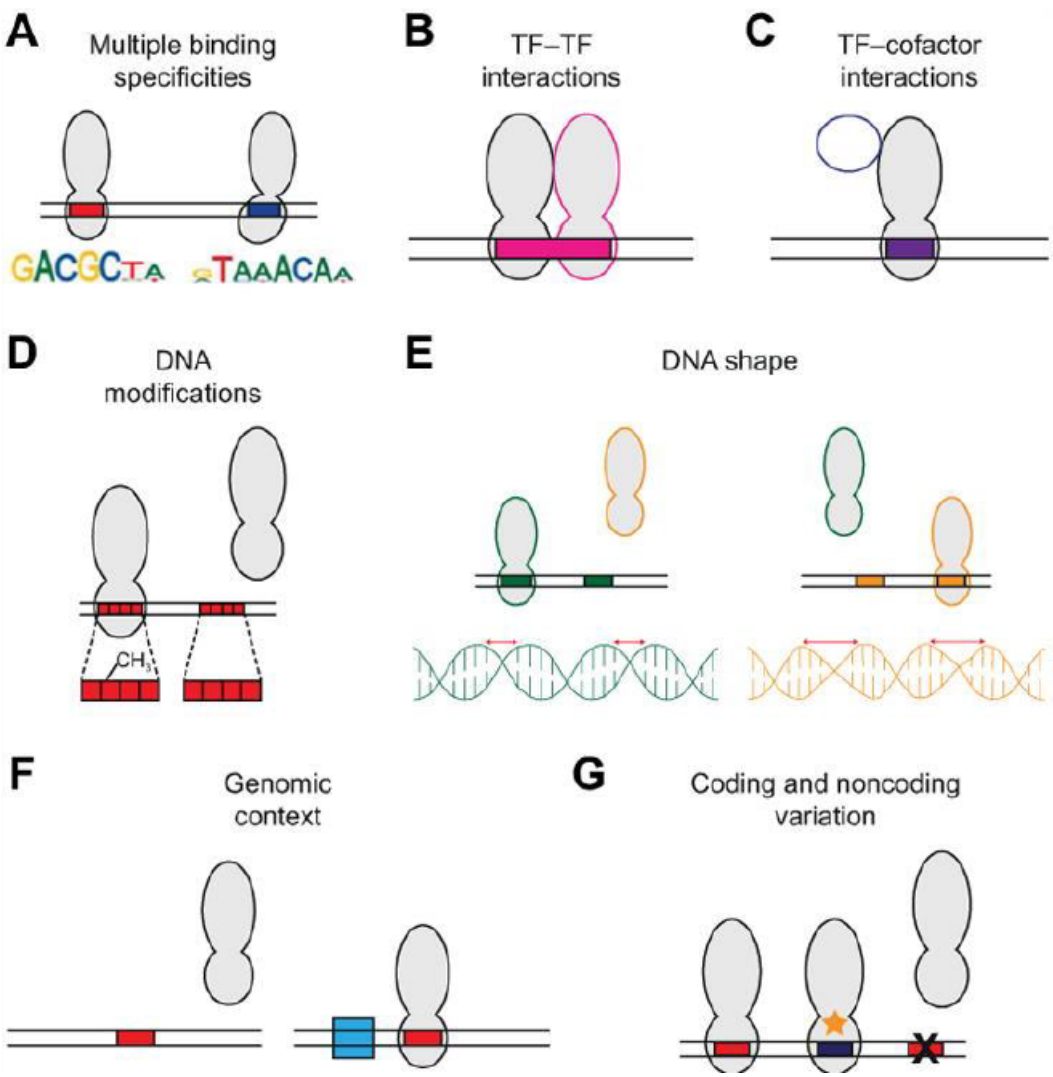


FIGURE 1.16 – (A) Un TF affiche une spécificité de liaison pour une séquence nucléotidique distinct. (B) Des interactions entre TF et (C) des TF liant des cofacteurs peuvent influencer sur la séquence des sites de liaison et présenter un site de fixation différent du motif TF monomère. (D) Les modifications de l'ADN, telles que la méthylation des cytosines (à gauche), peuvent moduler la liaison du TF. (E) De nombreux TF prennent en compte la forme de l'ADN, telle que la distance entre les hélices (représentée par des flèches rouges) et des paramètres tel que la torsion de l'hélice dans le cadre de la reconnaissance TF-ADN. (F) Les caractéristiques en dehors du motif du site de liaison (représenté par la boîte bleue), telle que la richesse en GC peut moduler la liaison TF-ADN. (G) Une mutation de la séquence de la protéine TF (représentée par une étoile orange) ou du site de liaison à l'ADN (représentée par une croix, à droite) peut modifier la liaison TF-ADN. Modifié de [Inukai et al., 2017].

1.1.4 L'épigénétique dans son ensemble

L'épigénétique définit les modifications de la chromatine qui vont influer l'expression des gènes sans altérer la séquence nucléotidique de l'ADN. Ces modifications héritables et réversibles sont largement impliquées dans l'interprétation du génome et sa régulation [Bonasio et al., 2010]. Parmi les différentes modifications épigénétique recensées, nous discuterons des plus populaires tels que : la méthylation de l'ADN, le positionnement des nucléosomes et les modifications post-traductionnelles des histones.

1.1.4.1 Méthylation des cytosines

La méthylation des cytosines (5mC), et plus précisément celles suivies d'une guanine, a capté une bonne partie de l'attention de la recherche en épigénétique. La présence de ces dinucléotides CpG méthylés sur les éléments cis-régulateurs d'un gène est associée généralement à une absence de transcription chez l'humain et la souris. Sur ces régions, la présence de méthylation peut moduler la fixation de facteurs de transcription [Smith and Meissner, 2013],[Baubec and Schübeler, 2014] ou induire la fixation de protéines. Les DNA Methyltransferase, une famille d'enzymes, sont responsables du processus de méthylation de l'ADN.

1.1.4.2 Variant et modification des histones

Comme vu précédemment, les histones vont permettre l'organisation de l'ADN en l'enroulant et en agissant sur son niveau de compaction. Cette régulation de la compaction va se faire par une modification des extrémités amino-terminales des histones qui pointent vers l'extérieur. Ces modifications vont entrainer des changements structuraux de la chromatine et la rendre plus ou moins accessible à différentes protéines qui vont pouvoir se fixer ou non sur l'ADN. Parmi ces modifications d'histone on compte la méthylation, l'acétylation, la phosphorylation ou encore l'ubiquitination [Rothbart and Strahl, 2014]. On représente la nomenclature des modification épigénétique des histones de la façon suivante :

— le nom de l'histone (ex : H3 pour l'histone H3)

— l'acide aminé et sa position (ex : K27 pour la lysine en 27eme position)

— le type de modification (me : methylation, me3 : tri-méthylation ; ac : acétylation ; P : phosphorylation ; ub : ubiquitination)

Comme expliqué dans la définition des enhancers, ces marques vont être de réelles étendards afin d'identifier des régions cis-régulatrice du génome. Une classification des modifications post-traductionnelles des histones peut ainsi être établie en se basant sur leurs effets et leurs localisations dans le génome ; les plus courantes sont présentées dans le tableau 1.1.

Marques d'histones		
Modification	Localisation	Effet associé
H3K9ac		Activateur
H3K9me3		Répresseur
H3K4me1	Enhancers,promoteurs	Activateur
H3K4me2	Enhancers,promoteurs	Activateur
H3K4me3	Enhancers,promoteurs	Activateur
H3Kme3	Enhancers,promoteurs	Activateur
H3K27ac	Enhancers,promoteurs	Activateur
H3K27me3	Enhancers,promoteurs	Répresseur
H3K36me3	Corps des gènes	Activateur
H3K79me3	Corps des gènes	Activateur

Table 1.1 – Liste des différentes modifications post-traductionnelles des histones associées à leurs fonctions et leurs localisations préférentielles dans le génome.

Cette classification permet d'établir un lien entre les marques épigénétiques présentes au niveau des éléments régulateurs et leurs activités. Par exemple, la marque H3K27me3 permet d'identifier des éléments régulateurs inactifs du génome. Il semblerait que l'activité des éléments régulateurs soit liée à la combinaison des différentes marques épigénétiques qu'ils contiennent,

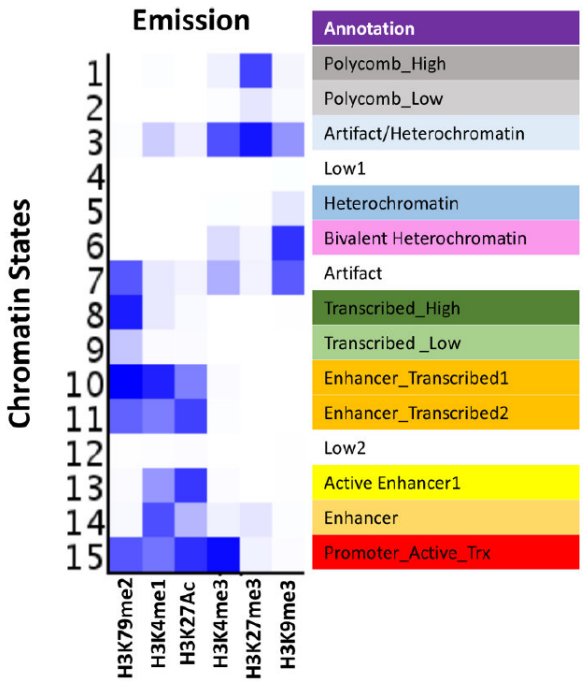


Figure 1.17 – Exemple de modèle de ChromHMM par statut d'histones, sur la base d'échantillons de tumeurs de mélanome.

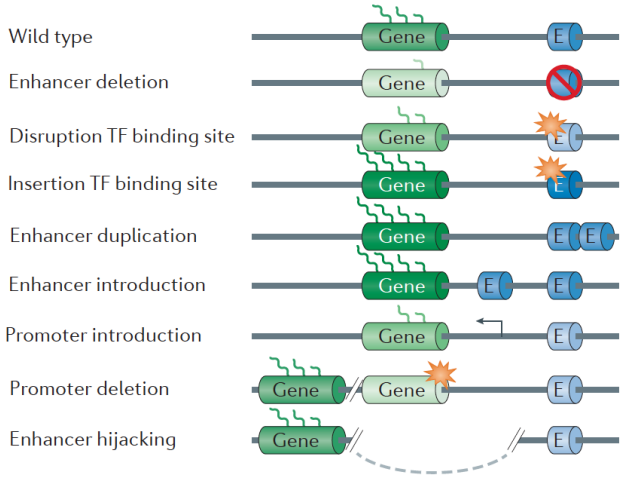
comme l'atteste l'existence de domaines bivalents (poised enhancer/promoter) qui possèdent à la fois des marques activatrices (H3K4me1 ou H3K4me3) et des marques répressives (H3K27me3) [Boland et al., 2014]. Ces combinaisons de marques d'histones vont être un excellent estimateur de l'état de la chromatine afférente (activateur actif, promoteur actif, etc.), meilleurs que les marques individuelles. ChromHMM repose sur ce principe ([Ernst and Kellis, 2012], Figure 1.17) utilisant un modèle de Markov caché (MMC) pour fractionner la chromatine en fonction des combinaisons d'histone marques observées. Chaque état se trouve souvent associé à un élément cis-régulateur actif ou non (promoteur actif, activateur inactif, etc.).

1.1.4.3 Dysfonctionnement des éléments régulateurs dans les processus pathologiques

Nous avons cité divers éléments régulateurs et expliqué certains de leurs mécanismes. Le bon fonctionnement de ces éléments est nécessaire au développement sain d'un organisme. Cependant, un dysfonctionnement de ces éléments va entraîner bien souvent l'apparition d'une pathologie, plus ou moins grave, les plus tristement emblématique étant les cancers, mais également des pathologies rares héréditaires. C'est notamment pour cela qu'il est important de bien comprendre tous les mécanismes de fonctionnement des éléments régulateurs. Afin de proposer des cibles thérapeutiques, ou de développer des moyens de lutte efficace face à ces maladies.

Les dysfonctionnement d'enhancers vont notamment être impliqués dans ce type de processus pathologique. Ils peuvent être supprimés ou être inactivés par des mutations ponctuelles qui perturbent les sites de liaison des facteurs de transcription, ce qui va aboutir à une dérégulation transcriptionnel du gène cible ([Krijger and De Laat, 2016],) Figure 1.18). Par exemple dans la leucémie aiguë lymphoblastique de type T, un enhancer va être crée suite à une mutation qui va permettre la fixation de facteurs de transcription augmentant l'activité de gènes à proximité dont les oncogènes TAL1 [Mansour et al., 2014] menant à la leucémie. Similairement, une mutation dans l'enhancer ZRS va mener à la fixation de facteurs de transcription qui vont mener à une expression ectopique du gène *Shh* qui va conduire au bourgeon de membre (doigt supplémentaire) [Lettice et al., 2012]. Plus généralement, les dysfonctionnements d'enhancers sont fréquents pour réguler à la hausse l'expression d'oncogènes [Zhang et al., 2016].

Les enhancers vont également être sujet à des réarrangements tels que des inversions ou des translocations, menant à des enhancers "hors-contextes", des enhancers placés dans des



Disease	Affected gene	Enhancer
β-Thalassaemia	β-globin genes	LCR
HPE	*SHH*	SBE2
PDD2	*SHH*	ZRS
Lung adenocarcinoma	*MYC*	3′~450 kb SE
T-ALL	*TAL1*	NA
α-Thalassaemia	α-Globin genes	α-Globin enhancers
α-Thalassaemia*	α-Globin genes/*NME4*	α-Globin enhancers
Burkitt lymphoma	*MYC*	IgH enhancer

FIGURE 1.18 – Les dysfonctionnements d'éléments cis-régulateurs tels que les enhancers et les promoteurs vont être impliqués dans des processus pathologiques. Ils vont être sujets aux mutations, peuvent être supprimés, dupliqués, entraînant des pathologies telles que la β-thalassémie, des leucémies, cancer des poumons, etc. [Krijger and De Laat, 2016]

environnements où ils ne devraient pas êtres menants à des activations non voulues. Ce mécanisme est fréquent dans les cancers tels que le lymphome de Burkitt qui juxtapose l'enhancer de la chaîne lourde de l'immunoglobuline à MYC [Taub et al., 1982, Dalla-Favera et al., 1982]. Les enhancers vont également pouvoir être détournés afin de contribuer à des processus de tumorigénèse comme dans les méduloblastomes qui placent le gène *Gfi* sous le contrôle d'un super-enhancer [Northcott et al., 2014].

Ces syndrômes, maladies et cancers ne sont pas que des mots, mais également les maux de familles qui doivent faire y faire face. Grâce à toutes ces découvertes, de nombreuses pistes thérapeutiques ont pu être découvertes, améliorant significativement la prise en charge de ces événements indésirables. Cependant, il reste toujours d'innombrables maladies sans traitements. C'est pour cela qu'il faut continuer, avec rigueur, de tenter de comprendre les mécanismes de régulations de bases et leurs dysfonctionnements afin d'apporter des pistes de traitements.

Au cours de ce chapitre nous nous sommes intéressés à la structure de base de l'ADN. À la conformation qu'elle peut adopter, aux éléments qui la compose, ainsi qu'à divers éléments de régulation de cette dernière. Bien-sûr cette liste est non-exhaustive et de nombreux éléments n'ont pas été couverts. Mais les bases ont pu être posées et on comprend qu'une dérégulation de ces mécanismes vont entrainer des conséquences désastreuses d'où la nécessiter de saisir pleinement la profondeur et la complexité de tous ces éléments. Au cours de ma thèse, je me suis intéressé de près à la découverte et la caractérisation des éléments silencers. Dans la prochaine section, je développerai plus en détail les mécanismes de répression génique.

1.2 Silencers et mécanismes de répressions

1.2.1 Les éléments silencers

La nature aime la stabilité, l'homéostasie. Ainsi on imagine facilement que quand un gène est transcrit ou que son activité augmente à un temps t, la cellule aura besoin à un temps t+1 de réduire l'activité de ce gène ([Chatterjee and Ahituv, 2017, Maston et al., 2006, Ogbourne and Antalis, 1998]). Les éléments de régulation stimulant la transcription tel que les promoteurs ou les enhancers ont été au cœur de la recherche ces nombreuses dernières années. Ils ont été décrits et jouissent de nombreuses références bibliographiques. Nous allons maintenant accorder du temps à leur antagonique : les silencers. Ces éléments sont moins décrits, moins compris, moins connues. Il s'agit de séquence cis-régulatrice qui vont pouvoir agir à distance sur la transcription en réduisant voire en supprimant l'activité génique comme indiqué dans la définition générale dans la section 1.1.2.5.

Ils ont été décrits la première fois dans les années 80 chez la levure ([Brand et al., 1985]). Le rôle des silencers chez les mammifères a été démontré peu de temps après grâce à un élément situé à plusieurs kilobases en amont du gène de l'insuline chez le rat ([Laimins et al., 1986]). Une décennie plus tard, des éléments silencers ont été identifiés au sein des introns des gènes CD4 chez la souris et l'Homme, et ont révélés le rôle majeur des silencers dans la détermination de la différenciation cellulaire, car ce silencer réprime l'expression de CD4 dans les cellules T CD8+. ([Donda et al., 1996, Taniuchi et al., 2004]). D'ailleurs une grande proportion des silencers identifiés l'ont été dans des cellules du système immunitaire notamment les cellules lymphoblastiques de type T. En effet, ce type de cellule permet de générer des mutants KO sans létalité pour le mutant. De plus, la voie de différenciation des cellules T offre un excellent modèle, en effet, après le stade double positif, la cellule va devoir choisir entre deux voies menant au CD4+ ou CD8+ (Figure 1.19). Cette régulation va notamment être menée par l'activation des silencers.

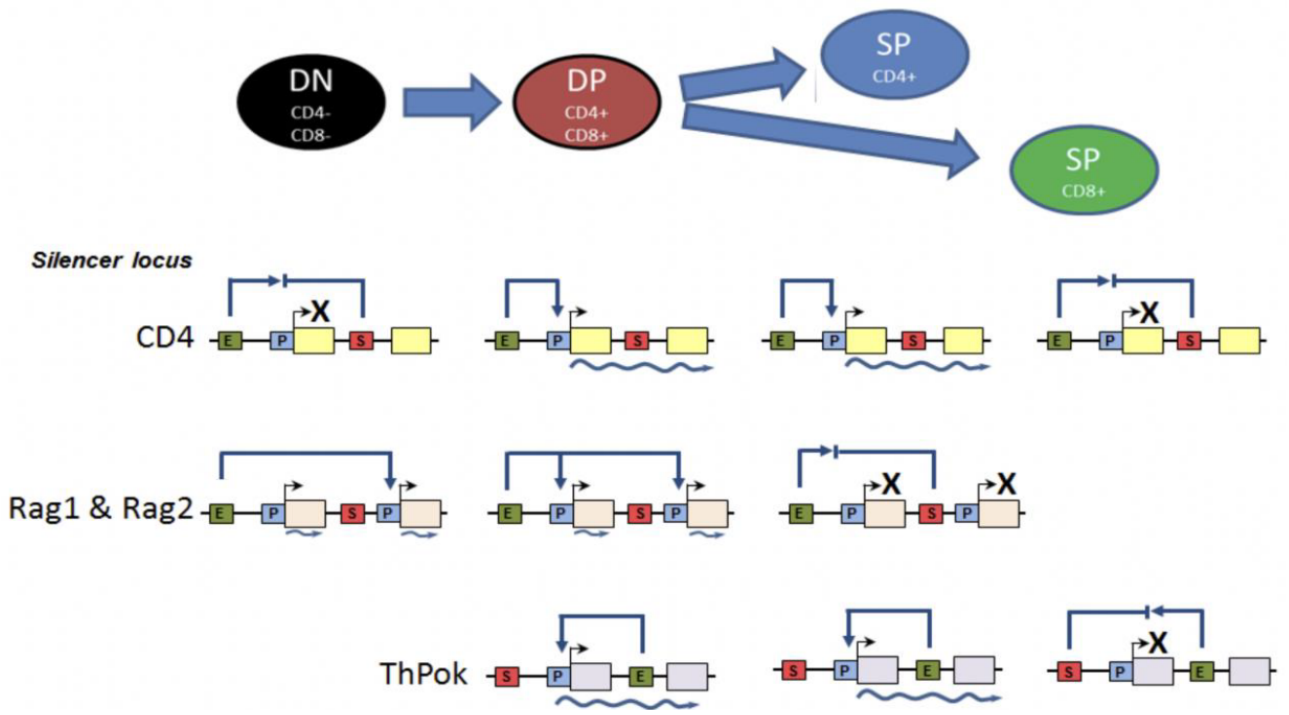


FIGURE 1.19 – Silencers identifiés et caractérisés impliqués dans la régulation des gènes des cellules lymphoblastiques de type T

Parmi les silencers identifiés dans les cellules T, on peut citer le locus *Rag1* et *Rag2 Rag*.

Dans les thymocytes en développements, les gènes *Rag* sont d'abord exprimés au stade CD4-CD8-double négatif (DN) pour favoriser la recombinaison des gènes *Tcrβ*, *Tcrγ* et *Tcrδ*. La recombinaison de Tcrβ provoque une régulation négative du gène *Rag*, une prolifération cellulaire et une différenciation au stade CD4 + CD8 + double positif (DP). Les gènes*Rag* sont ensuite réexprimés dans les thymocytes DP pour favoriser la recombinaison des gènes *Tcrα*. Après l'assemblage du gène *Tcrα* et la différenciation des thymocytes DP exprimant le *TCR*, les gènes *Rag* sont inhibés lors de la différenciation au stade CD4 + CD8- ou CD4-CD8 + simple positif (SP) [Kuo and Schlissel, 2009].

La région intergénique entre *Rag1* et *Rag2* contient un silencer qui réprime fortement l'expression de *Rag2* dans les cellules T DP et atténue l'expression dans les cellules DN T ([Yannoutsos et al., 2004]). À une distance de 86 kb du 5' de *Rag2* un élément ASE (Anti-Silencer Element) va venir supprimer l'effet silencers de la région intergénique dans les cellules DP ([Yannoutsos et al., 2004]). Nous avons un modèle de régulation plus complexe, mais aussi plus réel de ce qui se passe réellement *in vivo* où même un silencer va être inhibé afin de permettre la différenciation cellulaire [Yannoutsos et al., 2004, Hao et al., 2015].

Un autre exemple impliquant un silencer dans les cellules T fait intervenir le silencer ThPOK. Il fonctionne comme un "maitre régulateur" de la différenciation cellulaire des cellules T. En effet, l'absence ou la présence de ThPOK est nécessaire et suffisant afin de mener une cellule lymphocytaire immature au stade CD4+ ou CD8+ [He et al., 2005].

À l'heure actuelle peu de silencers ont été identifiés ([Goodbourn et al., 1985], [Henson et al., 2014], [Petrykowska et al., 2008]) et leurs mécanismes restent peu documentés. Cependant, il en ressort que les silencers fonctionnent indépendamment de l'orientation et à distance du gène régulé. Ils peuvent être situés au sein d'un promoteur proximal, d'un enhancer, au sein d'un intron, dans la région 3' non transcrite ou isolés tel une région de régulation distale indépendante. Enfin, les silencers peuvent coopérer afin d'augmenter leurs affinités à l'ADN ([Harris et al., 2005]), et peuvent agir en synergies afin d'augmenter leurs effets régulateurs ([Sertil et al., 2003]).

On peut classer les silencers dans deux différentes classes (Figure 1.20). Les silencers contexte-spécifique qui vont réprimer l'activité du promoteur (appelés éléments de régulation négatifs (NRE) [Ogbourne and Antalis, 1998] ; et les silencers autonomes, qui vont fonctionner de manière indépendante.

Les silencers fonctionnent en recrutant des protéines qui perturbent ou inactivent la formation du complexe de transcription fonctionnels Pol II au niveau du promoteur. Des protéines répressives sont recrutées, en bloquant la liaison à proximité des protéines activatrices ou en concurrence directe avec les protéines activatrices pour un même site de liaison [Maston et al., 2006].

Les silencers autonomes fonctionnent en établissant un état répressif de la chromatine ([Dean, 2011],Figure1.20). Ceci est généralement accompli en recrutant des protéines capables de modifier l'ADN (par exemple, les méthyltransférases) ou d'agir sur les histones, d'une manière qui favorise la formation de l'hétérochromatine ou des protéines qui aident à stabiliser et à propager l'hétérochromatine (par ex. protéines du groupe polycomb). Cela empêche à son tour les activateurs et les facteurs de transcription d'accéder aux promoteurs des gènes.

Les silencers sont restés pendant longtemps peu documentés dans la littérature. Un des silencers le mieux compris implique le facteur de transcription REST. Il agit sur une séquence connue sous le nom de : Repressor Element 1/Neuron-restrictive Silencer Element (RE1/NRSE). REST est un facteur transcriptionnel du groupe des Zinc Finger, très bien conservé au sein des cordés [Schoenherr and Anderson, 1995, Chong et al., 1995].

Bien que RE1 ait été initialement découvert comme un silencer pour les gènes neuronaux dans les cellules non neuronales, REST s'est également avéré jouer un rôle crucial dans le cerveau et dans la répression des gènes non neuronaux [Lu et al., 2014, Tang et al., 2016]. Ainsi, NRSF/REST peut être impliqué dans le développement d'autres tissus tel que dans le développement cardiaque [Kuwahara et al., 2003], probablement dans les lignées de cellules

TABLE 1.2 – Exemple de silencers validés par reporter assays chez l'Homme et la souris. Une bonne proportion des silencers validés l'ont été dans des cellules du système immunitaire. Adapté de ([Ogbourne and Antalis, 1998, Qi et al., 2015]).

Target gene	Size(bp)	References
Human plasminogen activator inhibitor type-2	302	(Antalis et al., 1996)
Human sperm histone H2b-1	28	(Barberis et al., 1987)
Human apolipoprotein A-II	100	(Bossu et al., 1996)
Human papilloma late mRNA	79?	(Dietrich-Goetz et al., 1997)
Human ϒD promoter	17	(Dirks et al., 1996)
T-cell receptor Vb2.2	39	(Dombret et al., 1996)
Human CD4	15	(Donda et al., 1996)
Human neuronal α1-chimaerin	30	(Dong and Lim, 1996)
Human insulin	21	(Goodman et al., 1996)
Human collagen type 4	21	(Haniel et al., 1995)
Human c-*Fes*	13	(He et al., 1996)
Human thyrotropin β	352	(Kim et al., 1996)
Human platelet-derived growth factor A-chain	30	(Liu et al., 1996)
Human Pi Class glutathione S-transferase	7	(Moffat et al., 1996)
Human T-cell activation gene 3	19	(Oh et al., 1997)
Human hypoxanthine phosphorylase transferase	59	(Rincon-Limas et al., 1995)
Human synapsin I	36	(Schoch et al., 1996)
Human c-*myc*	9	(Takimoto et al., 1989)
Human platelet-derived-growth-factor A-chain	25	(Wang et al., 1994)
Human interleukin-2	9	(Williams et al., 1991)
Human interleukin-8	7	(Wu et al., 1997)
Human interferon-ϒ	25	(Ye et al., 1996)
CCND1	953	(French et al., 2013)
MECP2F3	985	(Liu and Francke, 2006)
TSHB	353	(Kim et al., 1996)
PDGFA	31	(Liu et al., 1996)
PPARD	500	(Yadav et al., 2018)
Mouse bone morphogenetic protein 4	1230	(Feng et al., 1995)
Mouse mammary-tumour-virus long terminal repeat	13	(Giffin et al., 1994)
Mouse thyroid HR β1	14	(Nagasawa et al., 1997)
Mouse thyroid HR β1	8	(Nagasawa et al., 1997)
Mouse major inducible *Hsp*70	1044	(Shimokawa and Fujimoto, 1996)

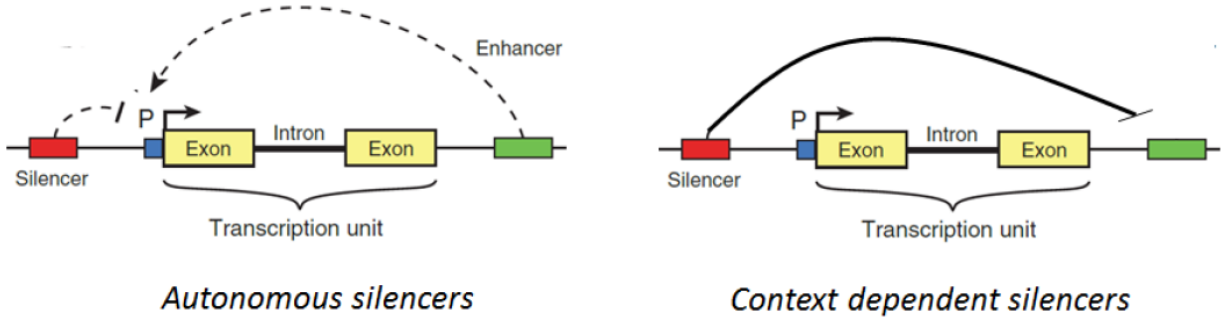


FIGURE 1.20 – Les silencers contexte spécifique vont directement cibler le promoteur du gène cible contrairement aux silencers autonomes qui vont établir un contexte défavorable à la transcription.

lymphocytaires de type B ou T [Scholl et al., 1996] et dans le développement des îlots pancréatiques [Abderrahmani et al., 2001]. En effet il a récemment été démontré qu'une inactivation de REST résulte en une augmentation dans la formation des cellules endocrines pancréatiques

[Rovira et al., 2021].

Ce n'est que récemment que plusieurs études ont amélioré la caractérisation des silencers chez l'homme, la souris et la drosophile par des tests fonctionnels à grande et petite échelle et ont fourni une nouvelle compréhension de leurs mécanismes de régulation. Une base de données a même vu le jour, regroupant des silencers validés expérimentalement et certains prédits informatiquement permettant la mutualisation des données [Zeng et al., 2021].

Parmi ces récentes découvertes on peut principalement citer 3 études majeurs parues récemment [Pang and Snyder, 2020, Ngan et al., 2020, Doni Jayavelu et al., 2020].

Pang et Snyder ([Pang and Snyder, 2020]) ont développé un système d'étude des silencers à l'échelle génomique reposant sur un criblage lentiviral couplé au séquençage à haut débit. Ils ont nommé cette technique le ReSE (Repressive abillity of Silencer Element). Les régions d'ADN candidates sont clonées en amont d'un promoteur contrôlant l'expression d'une protéine pro-apoptotique, Caspase-9 ([Pang and Snyder, 2020], Figure 1.21). Sans pouvoir représseur, le promoteur active l'expression Caspase-9, déclenchant l'apoptose. Cependant, si la région clonée est un silencer, Caspase-9 est réprimé et la cellule survit. Ils ont préalablement sélectionné des régions ouvertes de la chromatine en se basant sur la technologie de FAIRE-seq (section 1.3.1.1.3).

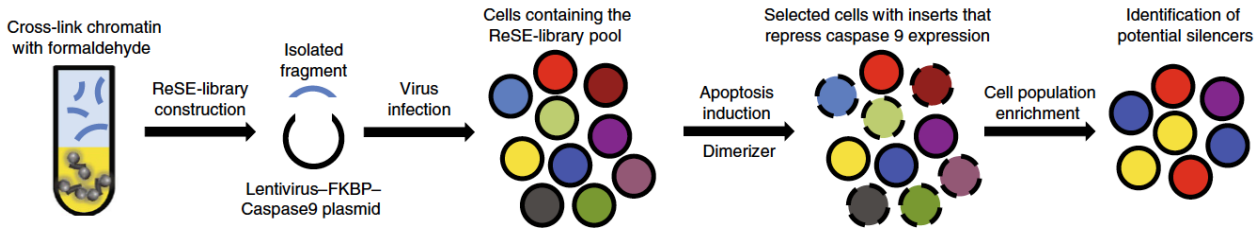


FIGURE 1.21 – Schéma explicatif du ReSe permettant l'identification des silencers à grande échelle. Les régions génomiques préalablement sélectionnées par FAIRE-seq sont transfectées en amont d'un promoteur contrôlant l'expression d'une protéine pro-apoptotique Cas9. Les silencers vont inhiber la Cas9 et éviter l'apoptose ([Pang and Snyder, 2020]).

Grâce à cette technique ils ont identifié plus de 2500 régions silencers tissu-spécifique en se basant sur plusieurs lignées cellulaires tel que K562 ou encore HepG2, dans le génome humain. Ils en ont validé plusieurs utilisant la technologie CRISPR-Cas9 et des tests de gène rapporteur (test luciférase). Ils ont ensuite tenté de caractériser ces éléments silencers identifiés. Leurs régions silencers sont enrichies dans les régions proximales des promoteurs, dans les régions introniques et les régions distales intergéniques (Figure : 1.22)

L'enrichissement des facteurs de transcription répresseurs tel que NCoR et REST a été démontré. Plus surprenant, Pang et Snyder ont constaté que les silencers identifiés sont très

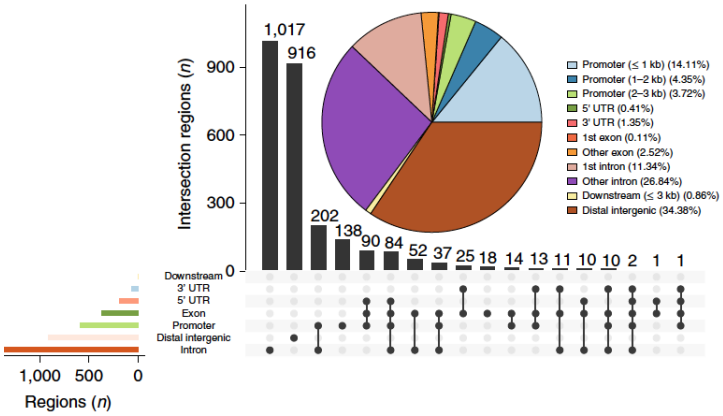


Figure 1.22 – Distribution génomique des silencers identifiées par [Pang and Snyder, 2020]

enrichis en histone méthylé H4 Lysine 20 (H4K20me), une marque de chromatine couramment associé à des gènes activement exprimés. Les silencers identifiés dans la lignée cellulaire K562 eux sont enrichis en H3K9me3, qui est associé à l'hétérochromatine. En revanche, les silencers détectés dans les cellules HepG2 étaient enrichis en H3K27me3, qui est associé à la répression médiée par le complexe répresseur polycomb. Ainsi, ils ont déterminé des enrichissements en marques d'histones différentes en fonction des lignées cellulaires.

Le complexe Polycomb est d'ailleurs au cœur de l'article de **Ngan et al** ([Ngan et al., 2020]). Cette étude met en lumière des mécanismes de régulation par interaction physique. Grâce au ChIA-PET, ils ont pu mettre en évidence différents types d'interactions faisant intervenir le complexe PCR2 (Polycomb Repressive Complexe) et agissant comme des silencers. En mutant les sites d'interactions, ils ont augmenté l'expression des gènes cibles démontrant l'action silencers des zones d'interactions. Plus impressionnant encore, ils ont pu valider l'activité de ces silencers *in vivo* en générant des souris mutantes KO homozygotes pour les silencers. Certaines de ces mutations se sont révélées létales, d'autres ont apporté des modifications phénotypiques significatives (Figure 1.23).

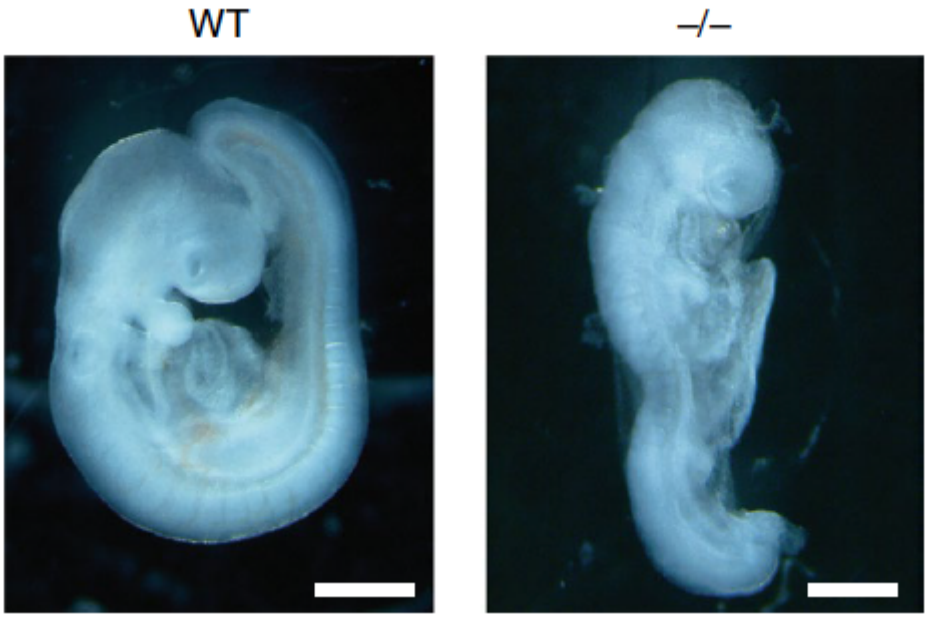


FIGURE 1.23 – Des mutants homozygotes KO pour les silencers ont été générés chez la souris mettant en évidence des différences phénotypiques significatives pointant le rôle et l'importance de ces séquences.

En suivant l'évolution des marques épigénétiques au sein des régions identifiées au cours de la différenciation, ils ont remarqué l'évolution de ces régions silencers en enhancers.

Une autre étude à grande échelle a été menée par **Jayavelu at al.** [Doni Jayavelu et al., 2020]. La technologie de MPRA que l'on décrira ultérieurement a été utilisée afin de tester des milliers de séquences parallèlement. Ces séquences ont préalablement été filtrées afin d'exclure des marques spécifiques des insulateurs ou encore des promoteurs. En résulte un lot de silencers sur lesquels ils ont entrainé un algorithme de Machine Learning afin d'ensuite prédire des silencers dans plusieurs autres lignées cellulaires. Ils ont pu mettre en évidence un enrichissement en facteur de transcription répresseur tel que REST ou YY1. Un comportement remarquable des silencers mis en évidence dernièrement est la capacité de ces derniers à passer de l'état d'enhancer à silencer en fonction du tissu. Un réel changement de fonction est opéré [Gisselbrecht et al., 2020]. Ces éléments enhancers / silencers bi-fonctionnels suggèrent d'autres modèles de répression de la transcription génique, où cet élément peut agir comme un silencer ou comme un enhancer selon le tissu ([Halfon, 2020]) (Figure : 1.24). Jayavelu et al., Pang et Snyder, et Ngang et al. ont

également obtenu des résultats dans ce sens.

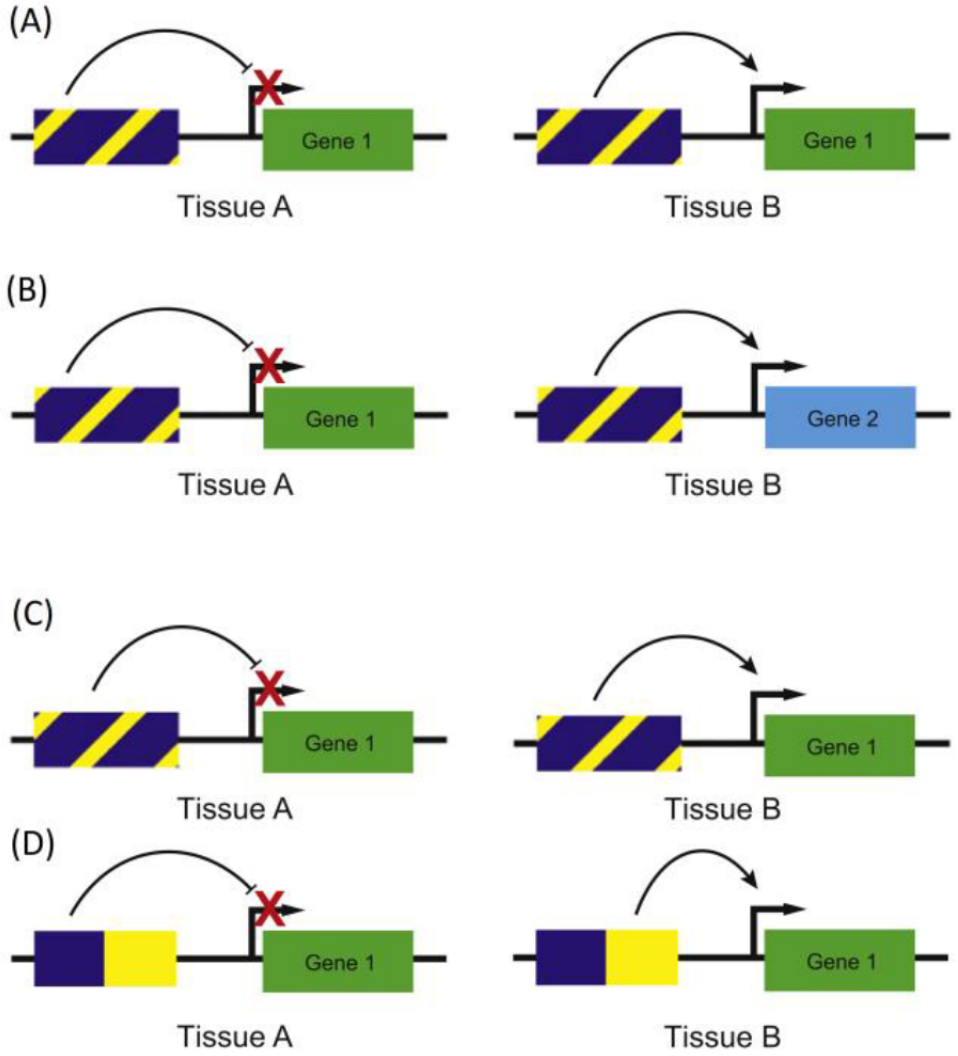


FIGURE 1.24 – (A) l'élément bi-fonctionnel régule un seul gène, réduisant l'activité du 'gène 1' dans le tissu A mais activant le 'gène 1' dans le tissu B. (B), l'élément bi-fonctionnel est un silencer par rapport au 'gène 1' dans le tissu A, mais un enhancer du 'gène 2' dans le tissu B. (C), les fonctions du silencer (bleu) et les fonctions de l'enhancer (jaune) sont mélangées au sein d'une seule séquence bi-fonctionnelle. Une alternative est représentée en (D), où les fonctions de silencer et d'enhancer résident dans des éléments de séquences séparables mais adjacents. Adapté de ([Halfon, 2020])

L'étude de ces éléments bi-fonctionnels nécessite d'aller au-delà des expériences de délétion, qui éliminent inévitablement les deux activités. Les techniques de perturbations épigénétiques ciblées, telles que l'activation ou l'inhibition médiées par CRISPR, sont susceptibles de permettre de mieux comprendre ces questions. Être capable de prédire dans quel type de cellule une région régulatrice donnée fonctionne comme un silencer par opposition à un enhancer serait très utile, par exemple, pour permettre une meilleure compréhension des effets de variations génétiques dans ces régions. Dans certains cas, la distinction entre les deux états régulateurs peut être basée sur la fonction des facteurs de transcription qui se fixent sur les mêmes éléments dans différents types de cellules. Cependant, de nombreux facteurs de transcription peuvent fonctionner à la fois comme activateurs ou comme répresseurs selon le contexte, en recrutant différents ensembles de cofacteurs (Figure : 1.25 [Della Rosa and Spivakov, 2020]).

Ces différentes études ont permis une caractérisation des silencers bien que de nombreuses

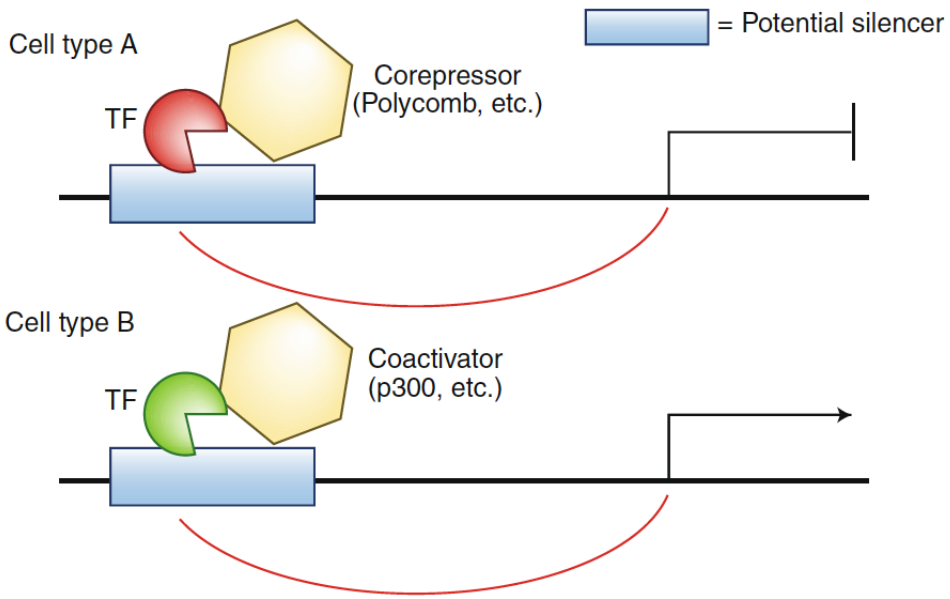


FIGURE 1.25 – Un silencer va pouvoir recruter différents facteurs de transcription en fonction du type cellulaire. Cependant un même facteur de transcription va pouvoir agir différemment en fonction du type de cofacteur associé. Certains silencers vont pouvoir agir comme enhancers en fonction du type cellulaire et de l'environnement chromatinien. [Della Rosa and Spivakov, 2020].

voies reste à explorer. Un mécanisme qui semble commun à ces études et l'utilisation du complexe polycomb et ainsi l'enrichissement en marque H3K27me3. D'ailleurs une autre étude va plus loin dans la caractérisation des régions enrichies en H3K27me3 en mettant en évidence des régions très riches en H3K27me3 (H3K27me3-Rich-Region - MRR). Ce sont des régions possédant un signal très fort en K27me3. Grâce au Hi-C, permettant l'analyse des interactions des régions génomiques, ils ont mis en évidence des interactions entre ces régions MRR et des gènes cibles. Ils ont montré que ces gènes ciblent étaient moins exprimés qu'un lot de gènes pris en comparaison. En étudiant les facteurs de transcription se fixant dans ces zones d'interactions entre les régions MRR et leurs gènes cibles, ils ont mis en évidence la présence de certains facteurs répresseurs tel que REST dont nous avons déjà parlé. Ces résultats ont été validés par des approches fonctionnelles expérimentales tel que le CRISPR, montrant qu'une délétion de ces régions allaient entrainer une augmentation du niveau d'expression des gènes cibles. *In vivo*, ces délétions se caractérisent par des processus métaboliques altérées ainsi qu'une perte de contrôle sur la régulation des processus de tumorigénèse ([Cai et al., 2021]).

Ainsi on comprend qu'en plus des régions régulatrices, les complexes vont avoir une importance cruciale dans l'action de régulation, qu'elle soit activatrice ou inhibitrice.

1.2.2 Les complexes répresseurs

De nombreux éléments sont à prendre en compte dans la régulation génique. Les différents éléments que nous avons présenté sont nécessaires à une compréhension globale des mécanismes de régulation.

1.2.2.1 Complexe répresseur REST

Dans le cas du silencer RE1/NRSF régulé par le facteur de transcription REST. Son mécanisme de répression est plus complexe que ce qui a été présenté dans la section 1.2.1. En effet, afin d'inhiber RE1, REST va recruter d'autres protéines intervenant dans la modification des

histones tels que les complexes d'histones deacetylase (HDAC) et des histones methyltransferase (EHMT2/G9A) [Roopra et al., 2004, Ding et al., 2008], le tout médié par des cofacteurs incluant RCOR1/CoREST et SIN3A ([Huang et al., 1999, Humphrey et al., 2001, Wang et al., 2008, Ooi and Wood, 2007], Figure 1.26).

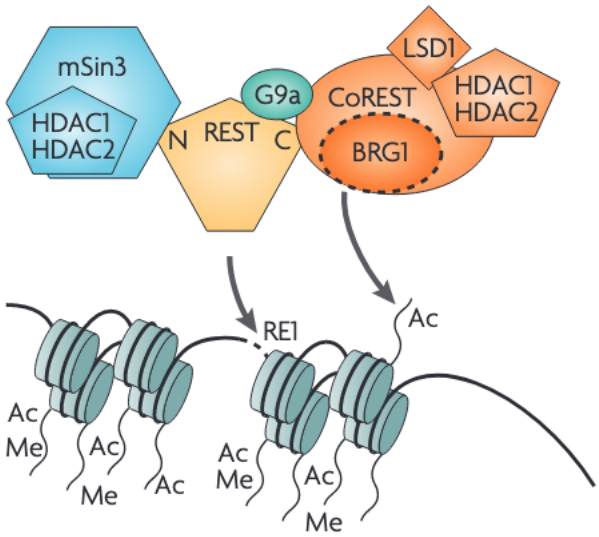


FIGURE 1.26 – Le complexe représseur REST est composé de nombreux cofacteurs. Il reconnait RE1 via ces sous-domaines "Zinc-Finger". Par son recrutement et son association avec des corépresseurs, REST va provoquer plusieurs modifications associées avec une perte des marques de transcriptions actives. [Ooi and Wood, 2007]

Récemment, il a été démontré que Lsd1 (sous unité catalytique de CoREST et du complexe represseur NuRD) qui est une histone démethylase, jouait un rôle crucial dans l'inhibition de certains gènes clés dans la différenciation cellulaire des cellules lymphoblastiques de type T. Sa délétion entrainant la sur-expression des gènes impliqués dans la réponse à l'interféron ainsi que des gènes régulés par les répresseurs transcriptionnels : Gfi1, Bcl6 et, surtout, Bcl11b dans les thymocytes CD4+ CD8+ [Stamos et al., 2021]. Le complexe formé par REST est également composé des éléments Switch/Sucrose Non-Fermentable (SWI/SNF), et le complexe répresseur polycomb 1 et 2 (PRC1, PRC2) ([Dietrich et al., 2012, Ooi and Wood, 2007]).

1.2.2.2 Complexe répresseur polycomb

D'autres études récentes impliquent le complexe répresseur polycomb dans les mécanismes d'actions des silencers. En effet, il a été démontré que la délétion de certaines régions de fixation de PRC2 allait entrainer une dérégulation des gènes cibles et même la létalité embryonnaire. Les cibles étant mises en contact avec les régions silencers via des mécanismes de boucle [Ngan et al., 2020]. Ces boucles sont observées de manière spécifiques afin de connecter le silencer avec sa cible. La présence du gène cible au sein du même domaine semble être une condition nécessaire à la régulation du gène par le silencer [Pang and Snyder, 2020] [Cai et al., 2021]. Certaines régions fixées par ces complexes répresseurs utilisant PRC2 semble également changer de statuts au cours de la différenciation cellulaire pour passer de répresseurs au stade pluripotent de la cellule à activateurs (enhancers) tissu-spécifique [Ngan et al., 2020].

Une autre étude récente va dans le même sens. Elle concerne l'expression du gène Ccl2. Son expression serait contrôlée par des régions cis-régulatrices non redondantes et fonctionnellement distinctes qui partagent des caractéristiques communes aux enhancers "actifs", y compris un niveau de H3K27ac élevé, la liaison de Pol II, la transcription d'ARNe, le recrutement de complexes protéiques (PU.1, AP -1, RUNX1), des coactivateurs (MED1, CBP) et un complexe

de corépresseurs (GPS2, SMRT). Le « super-enhancer » Ccl2 s'est avéré être composé d'un enhancer majeur et d'une région silencer majeure ([Huang et al., 2021]).

1.2.2.3 The Human Silencer HUB : le complexe représseur HUSH

Le complexe répresseur HUSH (Human Silencing Hub) va intervenir au niveau d'une autre marque épigénétique caractéristique de l'hétérochromatine : H3K9me3. Il cible spécifiquement cette marque et une délétion du complexe résulte en un appauvrissement en H3K9me3. Ce complexe est composé des protéines TASOR, MPP8 et PPHLN1 (codant pour la periphiline) (Figure 1.27). MPP8 possède un chromodomaine capable de reconnaitre le H3K9me3 ainsi le complexe HUSH va être recruté grâce à cette protéine, bien que la délétion d'une de ces protéines va résulter en la diminution de la présence des autres sous-unités du complexe. En effet, une perte de la periphiline va entrainer la déstabilisation du complexe sur la chromatine suggérant que la periphiline contribuait au maintien du complexe HUSH sur la chromatine. TASOR lui, est la pièce centrale du domaine, permettant la fixation des autres protéines. Il va être impliqué dans la déposition de marque H3K9me3 surtout au sein des régions génomiques répétées [Douse et al., 2020].

Ce complexe va également utiliser d'autres facteurs tel que SETDB1 afin de faciliter l'enrichissement de la marque H3K9me3. Ainsi le complexe HUSH semble préférentiellement réprimer l'activité des gènes intégrés dans une chromatine comportant un haut niveau de H3K9me3 [Tchasovnikarova et al., 2015]. Nous avons ici une régulation via un mécanisme d'effet de position.

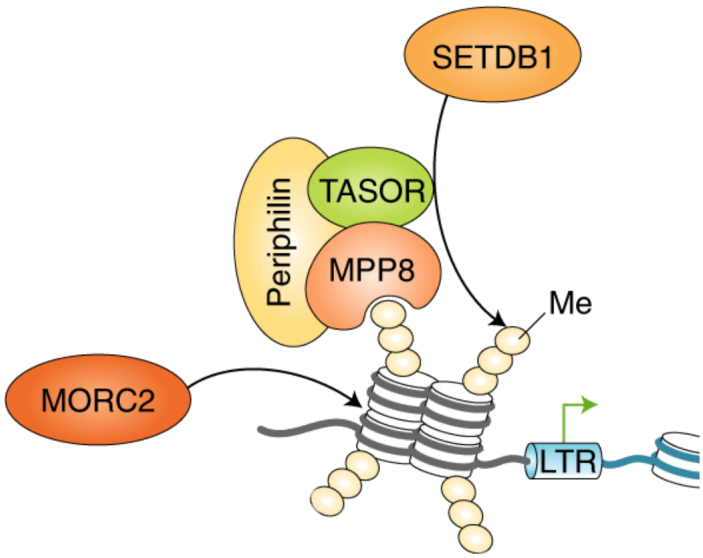


FIGURE 1.27 – Le complexe représseur HUSH est composé des protéines TASOR, MPP8 et de la periphiline. MPP8 va reconnaitre les marques H3K9me3 et la periphiline va maintenir la stabilité du complexe. TASOR agit comme support au sein du complexe et facilite le dépôt des marques H3K9me3, surtout au sein des régions génomiques répétées. MORC2 et SETDB1 ne font pas partis du complexe mais co-agissent avec lui afin d'enrichir ces régions encore plus qu'elles ne le sont en H3K9me3 (SETDB1) et maintiennent la chromatine en configuration fermée (MORC2) . En résulte une répression des éléments répétés notamment les LTR. De TCHASOVNIKAROVA

La marque H3K9me3 va d'ailleurs être responsable de l'inhibition de certains éléments répétés tel que les LTRs (Long Terminal Repeat, aussi connus sous le nom de satellite), des rétrovirus endogènes (ERV), des SINEs (Short Interspersed Nuclear Elements) et des LINEs

(Long Interspersed Nuclear Elements) [Kassiotis and Stoye, 2016]. Comme nous l'avons vu dans la présentation des retrotransposons, ces séquences sont fréquemment associées à diverses pathologies. Par exemple dans les leucémies aiguës myéloïdes (AML) où une mutation critique de SETDB1 mène à la désinhibition de ces éléments transposables et ainsi à la pathologie [Cuellar et al., 2017].

Il a récemment été démontré que le complexe HUSH allait aussi réguler des gènes impliqués dans le mécanisme d'inflammation par exemple en inhibant l'Interféron 1 (IFN1) via une régulation épigénétique des éléments répétés LINEs ([Tunbak et al., 2020]. Ainsi le complexe HUSH est un complexe répresseur et son dysfonctionnement est associé à un grand nombre de pathologies liés à la dérégulation des éléments transposables. Il est associé à une haute densité de la marque H3K9me3 et fonctionne comme représseur des gènes Zinc Finger et des ADN ribosomiques [Tchasovnikarova et al., 2017, Tchasovnikarova et al., 2015].

Une théorie serait que HUSH contribuerait à la protection du génome contre les invasions et insertions dans le génome en réprimant et rendant inaccessible ces séquences à la transcription via la marque H3K9me3. En effet il a été démontré très récemment que ce complexe HUSH fixe et réprime un sous ensemble de gènes endogènes, ne possédant pas d'introns, générés par la rétrotransposition des ARNm cellulaires. Ainsi, l'ADNc sans intron, marque de la transcription inverse, fournit un moyen polyvalent de distinguer les rétroéléments envahissants des gènes hôtes et permet à HUSH de protéger le génome de l'ADN « externe » [Seczynska et al., 2022]. Cette découverte majeure révèle l'existence d'un système de surveillance du génome dépendant de la transcription et explique comment il offre une protection immédiate contre les éléments nouvellements acquis tout en évitant une répression inappropriée des gènes de l'hôte [Seczynska et al., 2022].

1.2.2.4 Le complexe répresseur KRAB-ZFP

Un autre mécanisme de défense de l'hôte au niveau transcriptionnel comprend la liaison de différentes protéines Zinc Finger répressives contenant le domaine KRAB (Krüppel Associated Box). Les complexes KRAB-ZFP constituent la plus grande famille de facteurs de transcription exerçant des fonctions de corépresseurs dans les cellules de mammifères [Ecco et al., 2017, Lupo et al., 2013]. En général, les KRAB-ZFP sont composés de deux unités. Ils peuvent se lier à des séquences d'ADN spécifiques via des motifs Zinc Finger et recruter un complexe répressif via le domaine KRAB. Leurs structures est particulière, avec dans la zone terminal la séquence KRAB qui peut être constituée de deux éléments KRAB A / KRAB B (parfois on peut trouver des domaines SCAN ou des domaines à fonction inconnu DUF). Le plus remarquable, sont ces séquences répétées Zinc Finger en C-terminal ([Sobocińska et al., 2021] Figure 1.28).

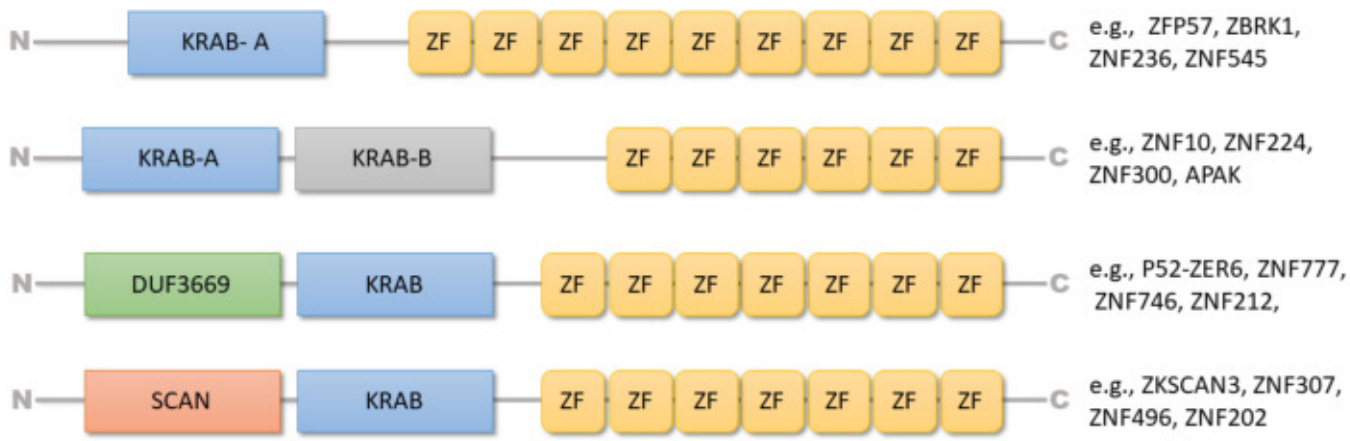


FIGURE 1.28 – Représentation du complexe KRAB ZFP avec en N-terminal les boîtes KRAB ainsi que leurs différentes isoformes et en C-terminal une répétition de zinc finger [Sobocińska et al., 2021].

Ce complexe est connu comme étant un répresseur des éléments transposables en utilisant des cofacteurs tel que KAP1 qui va servir de point d'ancrage à d'autres protéines comme HDAC (aussi impliquée dans le complexe REST) et SETDB1 impliquée dans le complexe HUSH ([Ecco et al., 2017], Figure 1.29)

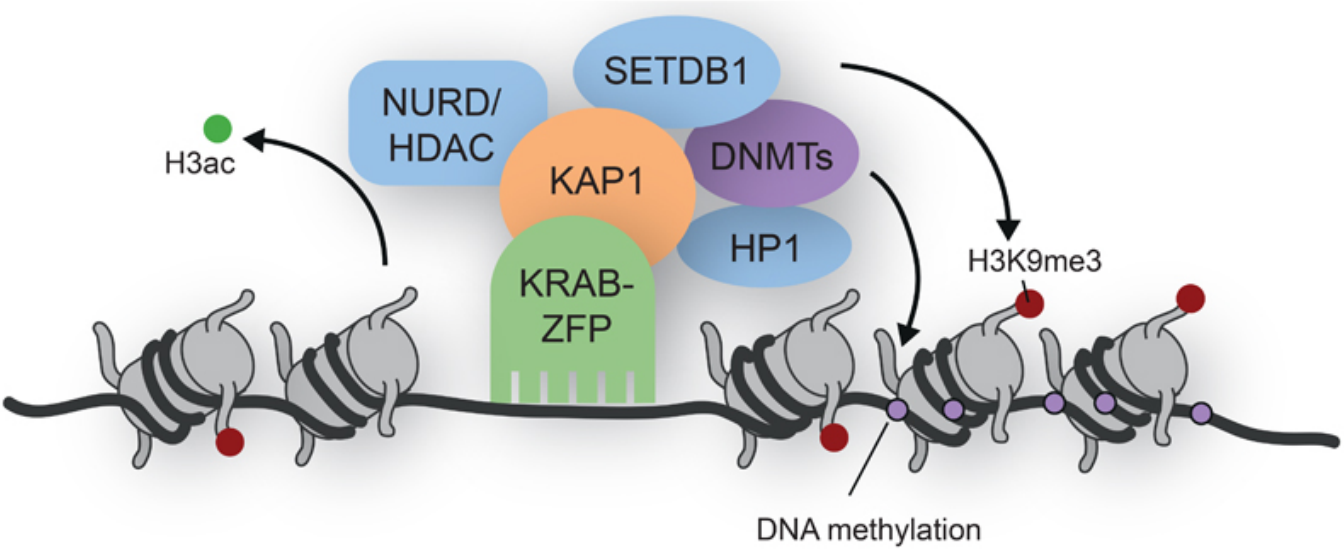


FIGURE 1.29 – Représentation du complexe KRAB ZFP et de ses mécanismes d'actions. KRAB-ZFP va venir recruter KAP1 et d'autres cofacteurs tel que HDAC, SETDB1, HP1, en résulte une structure en hétérochromatine.

La répression va faire intervenir des mécanismes de modifications épigénétiques tels qu'une modification des histones (triméthylation de l'histone 3 au niveau de la lysine 9 - H3K9me3) et de méthylation de l'ADN, le tout résultant en une compaction de l'ADN plus connue sous le nom d'hétérochromatine et donc une inhibition de la transcription.

1.3 Analyse, identification et annotation des éléments régulateurs

En 2001, le Consortium du séquençage du génome humain a publié une première ébauche du génome humain. Il s'agissait déjà d'une révolution dans le domaine de la génomique malgré un génome non totalement séquencé. Le 31 mars 2022 est à marquer d'une pierre blanche, le consortium Telomere-to-Telomere (T2T) a terminé la première version véritablement complète de 3,055 milliards de paires de bases (pb) du génome humain, ce qui représente la plus grande amélioration du génome de référence humain depuis sa première version [Nurk et al., 2022].Parallèlement, une équipe de Stanford a établi le record du monde du temps de séquençage le plus rapide pour un génome entier. La technique développée est capable de séquencer un génome en environ 5h et de fournir un diagnostic génétique en moins de huit heures, un délai jusqu'alors inédit [Gorzynski et al., 2022].

Ces avancées majeures permettront une compréhension plus complète des mécanismes génétiques et épigénétique. En effet, la nouvelle référence T2T-CHM13 comprend des assemblages sans interruption, pour les 22 autosomes plus le chromosome X, corrige de nombreuses erreurs et introduit près de 200 millions de paires de bases de nouvelles séquences contenant 2 226 copies de gènes paralogues, dont 115 devraient être codantes pour des protéines. Ces nouvelles régions comprennent toutes les régions centromériques et les bras courts des cinq chromosomes acrocentriques, ouvrant pour la première fois ces régions complexes du génome à des études fonctionnelles [Nurk et al., 2021].

Malgré les prouesses technologiques et scientifiques récentes, de nombreux mécanismes restent à être explorés afin de percer les secrets de notre génome et de sa régulation. Nous savons que les éléments de régulation transcriptionnelle sont des facteurs clés au cours du développement, des réponses aux stimuli, de l'homéostasie cellulaire et de la maladie. Par conséquent, identifier et comprendre la fonction de ces éléments de régulation est un défi.

1.3.1 Le séquençage du génome

1.3.1.1 Techniques de séquençages basées sur la méthode à haut débit

La réduction du coût de séquençage est un enjeu crucial de ces dernières décennies pour le développement de nouvelles applications en médecine ([Rabbani et al., 2016]), biotechnologie et recherche fondamentale. Ainsi de nouvelles techniques ont émergé, toujours plus efficace. Se sont succédé les approches de séquençage par terminaison de chaîne (Sanger), pyroséquençage (454), ion semi-conducteur (Ion Torrent) et ligation qui n'est quasi plus utilisé (SOLiD). Le séquençage par synthèse (Illumina), qui s'est largement popularisé, pourrait être dans les années à venir, dépassé par, ou au moins coexister avec, de nouvelles technologies aux mécanismes innovants, notamment à l'origine des dernières prouesses citées précédemment. On peut ainsi penser aux technologies permettant de générer des "long reads" (Pacbio SMRT et Oxford Nanopore) bien que des adaptations de protocoles aient été développées pour satisfaire ces applications pour séquençage sur des dispositifs Illumina (10X Genomics [Quail et al., 2012]). Les données exploitées dans ce manuscrit proviennent principalement de séquenceurs Illumina.

Au-delà des méthodes d'exploitations directes des séquences génomiques, de la détection de variant, de réarrangement chromosomique, ainsi que l'étude de la diversité génétique de populations ([Davey and Blaxter, 2010]), se sont développées de nombreuses techniques explorant diverses caractéristiques du génome. Les principes des techniques que j'ai été amené à utiliser sont abordés ci-dessous.

1.3.1.1.1 RNA-seq et analyse du transcriptome Le RNA-seq est une technique de quantification du niveau d'expression des gènes en termes de transcrits ARN ([Morin et al., 2008]. C'est la technique basée sur le séquençage à haut débit la plus répandue, devant la deuxième technique la plus populaire, le ChIP-Seq. Cette technique repose sur des étapes successives : d'extraction de l'ARN des cellules, de sélection pour des familles d'ARNs d'intérêt, de fragmentation et de rétro-transcription en ADN avant séquençage à haut débit (Figure : 1.30). Les ARNs ribosomiques étant très majoritaires, la sélection est nécessaire pour pouvoir quantifier avec efficience les autres types d'ARNs. Si l'intérêt est porté exclusivement sur les transcrits d'ARNs messagers matures, la sélection peut être réalisée par sélection basée sur leur queue polyadénylée. Sinon, un enrichissement relatif pour les ARNs non ribosomiques peut être obtenu par ribo-déplétion. L'étape de rétro-transcription, nécessaire pour un séquençage avec une technologie par synthèse (Illumina), n'est plus nécessaire avec celle par nanopore ([Depledge et al., 2019]). La quantification des transcrits peut se faire après alignement sur le génome ([Dobin et al., 2013]) ou sur des transcrits de référence ([Bray et al., 2016]). Le RNA-seq permet de détecter de nouveaux gènes, des transcrits alternatifs et des gènes de fusion ([Vu et al., 2018] ;[Heyer et al., 2019]).

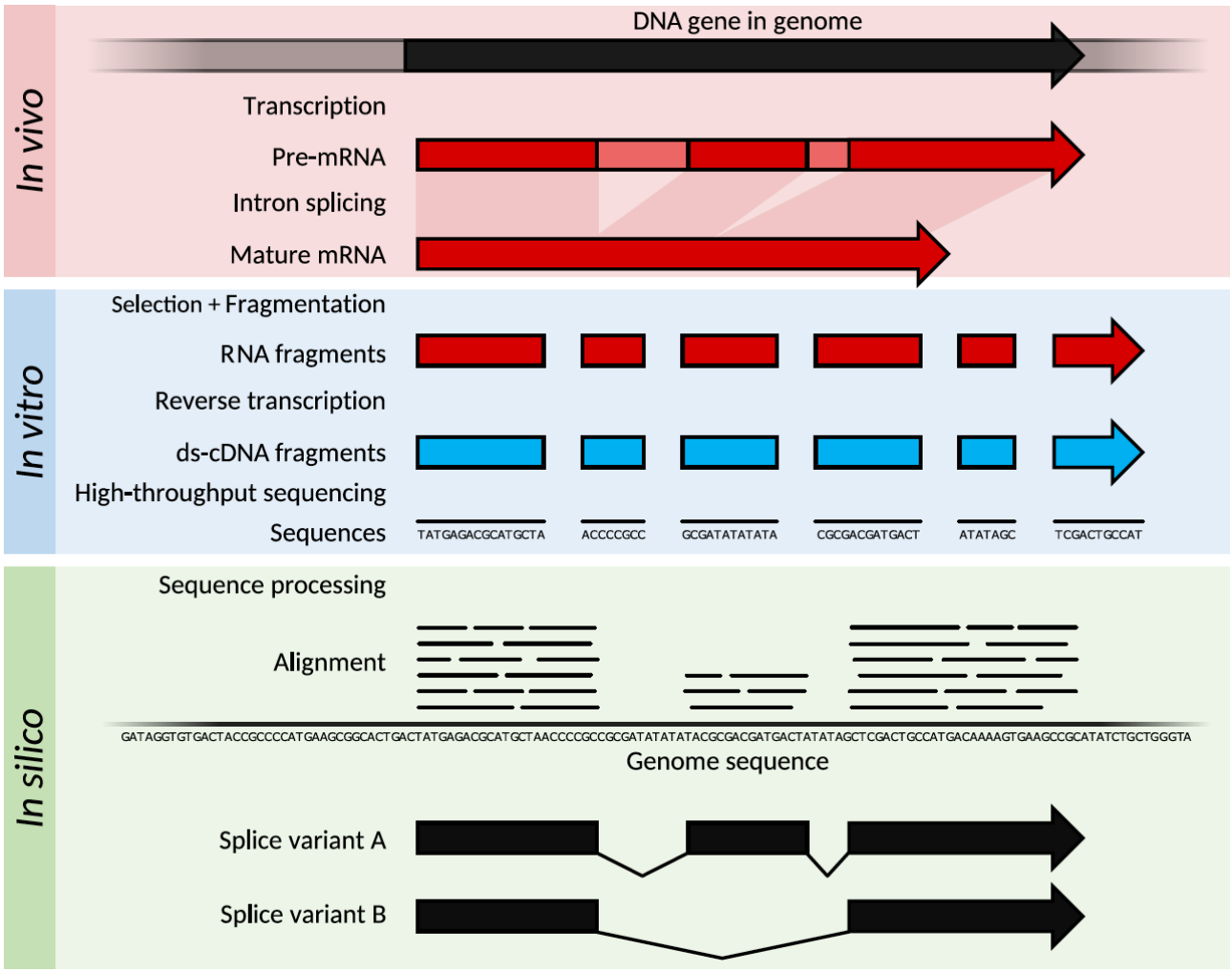


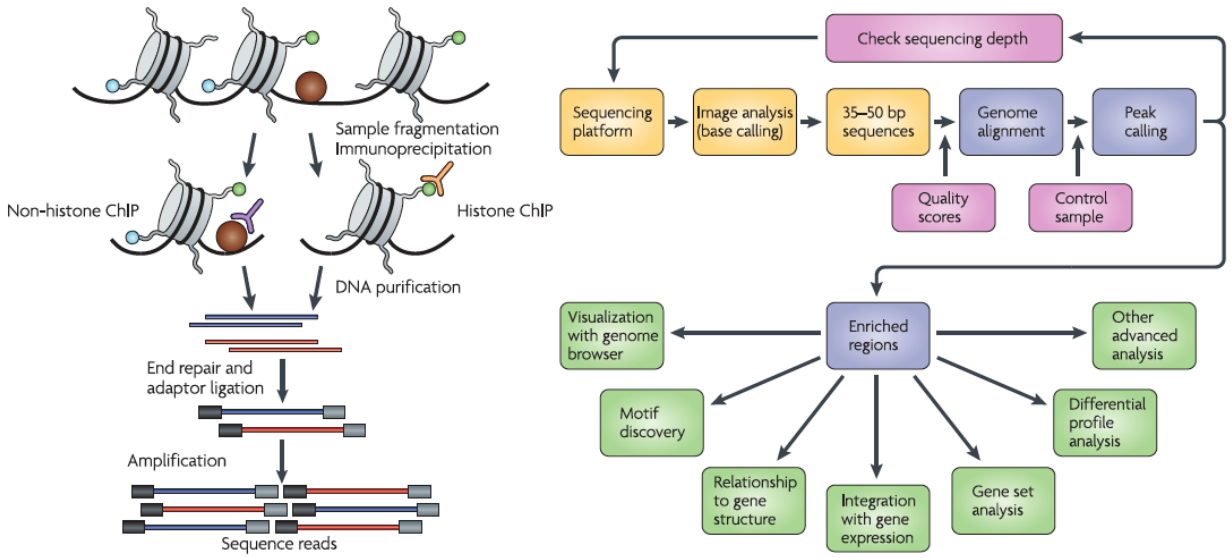
FIGURE 1.30 – Principe du RNA-seq Thomas Shafee Wikipedia / CC BY-SA 4.0

1.3.1.1.2 ChIP-seq et analyse des interactions protéine-ADN Le ChIP-seq (Chromatin ImmunoPrecipitation followed by sequencing) est une technique d'analyse des interactions entre les protéines et l'ADN ([Barski et al., 2007]). Elle permet de détecter et d'identifier les séquences d'ADN liées in vivo par une protéine donnée, de façon directe ou indirecte par l'intermédiaire

d'autres protéines. Cette technique est applicable à l'étude des sites de fixation des facteurs de transcription, des cofacteurs, de l'ARN Polymérase II et des modifications d'histones.

Le ChIP-seq repose sur le principe d'enrichissement des séquences ADN fixées par une protéine d'intérêt grâce à un anticorps spécifique de cette protéine. Elle apporte une couverture du génome plus complète et une plus haute résolution dans le positionnement des sites de fixation identifiés [Barski et al., 2007]. Le formaldéhyde est utilisé pour fixer de façon covalente l'ensemble des protéines à l'ADN. La chromatine est ensuite isolée, fragmentée puis immuno-précipitée par un anticorps spécifique de la protéine d'intérêt (Figure : **??**). Une meilleure résolution peut être atteinte si l'étape de fragmentation, initialement basée sur le principe de sonication de l'ADN, est remplacée ou complétée par une digestion enzymatique : MNase ou lambda exo-nucléase [Rhee and Pugh, 2011].

Les fragments sont séquencés puis alignés sur le génome de l'espèce étudiée. Le signal quantitatif de la couverture du génome par des fragments est converti en signal qualitatif par une étape de recherche des pics (peak-calling). Un échantillon contrôle, nommé input, contenant de la chromatine fragmentée mais non immuno-précipitée sert de référence pour identifier les enrichissements locaux de la chromatine, des artefacts, principalement liés au protocole de fragmentation [Meyer and Liu, 2014], d'amplification [Benjamini and Speed, 2012] et à l'incomplétude des génomes de références [Miga et al., 2015]. Différents types d'analyses peuvent être réalisées ensuite à partir des régions enrichies pour la protéine d'intérêt : analyse de motifs, analyse d'enrichissement fonctionnel en certaines familles de gènes, intégration et analyse différentielle avec d'autres expériences omiques, etc.



(a) Principe expérimental (b) Étapes standards d'analyse

FIGURE 1.31 – (a) Principe du ChIP-seq ainsi que (b) des différentes étapes du traitement bioinformatique des données issues de cette technologie [Park, 2009].

Plus récemment, des alternatives au ChIP-seq ont émergé. On peut notamment citer le **Cut & Run** (Cleavage Under Targets & Release Using Nuclease). Contrairement au ChIP-seq, il n'y a pas besoin de fragmenter physiquement l'ADN des cellules, ce qui permet d'analyser les interactions protéines/ADN dans un état plus naturel. Comme dans le ChIP-seq, un anticorps va reconnaitre spécifiquement une protéine, cependant, dans le Cut & Run, cet anticorps se lie à la protéine cible dans les cellules intactes et découpe l'ADN auquel la protéine est liée via une endonucléase (MNase) couplée à l'anticorps [Skene and Henikoff, 2017], la MNase étant activé par l'ajout de Calcium. Cette nouvelle stratégie permet aux fragments d'ADN d'être séquencés et identifiés plus efficacement qu'il n'est actuellement possible avec ChIP. En plus d'être plus

résolutif et moins long, cette technique est plus économique car elle nécessite un séquençage moins profond.

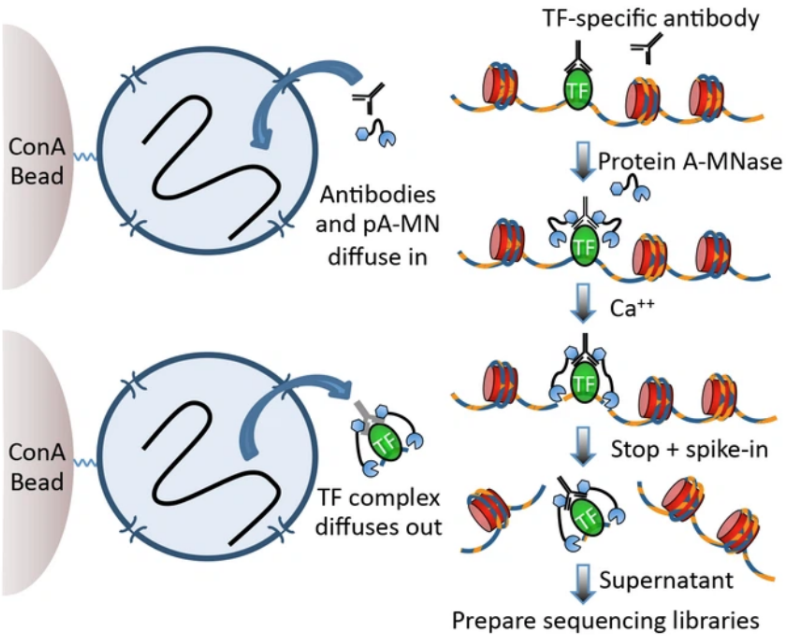


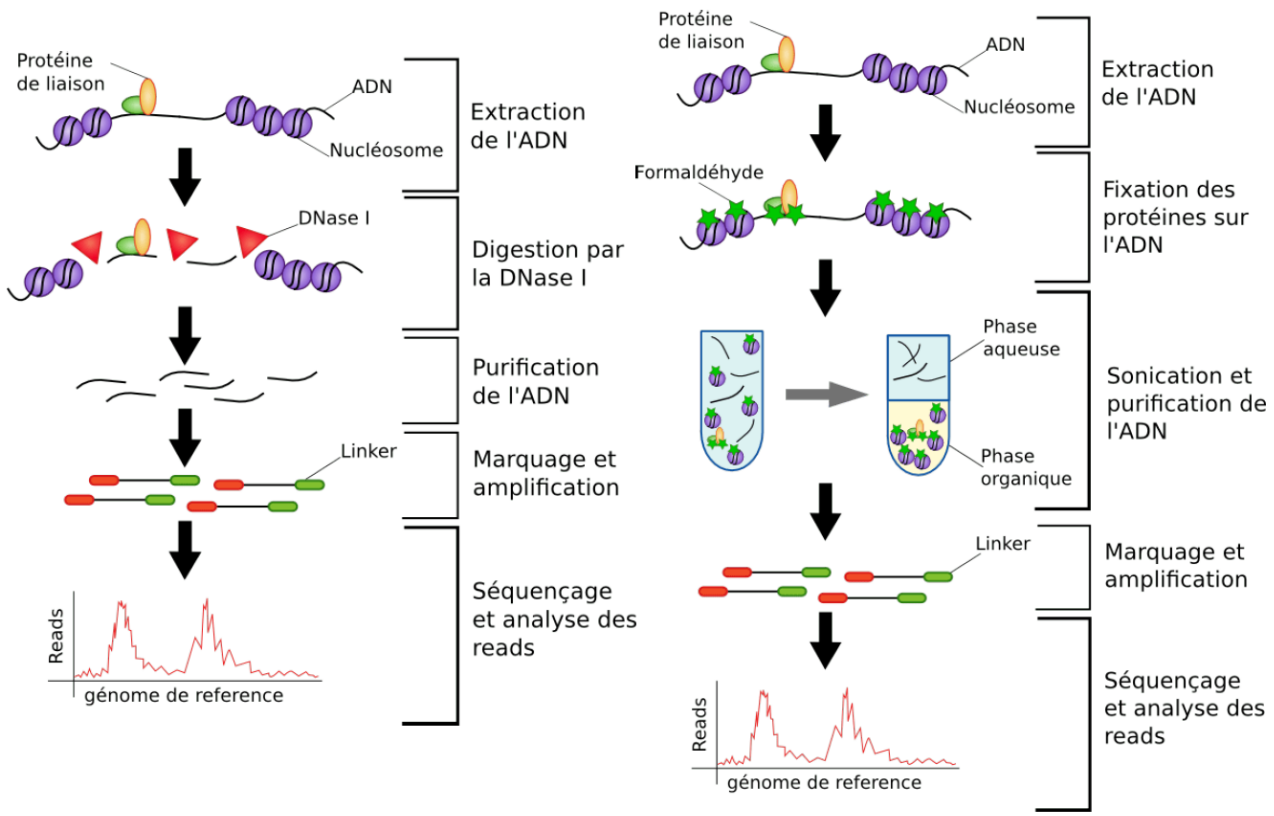
FIGURE 1.32 – Principe du Cut&Run, une nucléase est couplée à l'anticorps de reconnaissance de la protéine cible, permettant de cliver spécifiquement la région d'intérêt permettant une meilleure résolution que le ChIP-seq [Skene and Henikoff, 2017].

Cependant le Cut&Run montre certaines limites du fait que les fragments coupés par la MNase sont relâchés dans le surnageant, ne le rendant pas applicable pour des analyses de Single-Cell.

Cette limitation peut-être dépassé par le **Cut&Tag** (Cleavage Under Targets and Tagmentation) qui repose sur l'utilisation d'une transposase couplée à des adaptateurs générant des fragments prêts pour la PCR et le séquençage [Kaya-Okur et al., 2019]. La région va être cibler par un anticorps spécifique couplé à une transposase. L'ajout de Magnésium va activer la transposase et intégrer les adaptateurs, générant les fragments de la librairie avec une très grande résolution et très peu de bruit de fond (background).

1.3.1.1.3 DNase-seq, FAIRE-seq, ATAC-seq et technologie d'analyse de l'accessibilité chromatinienne Le DNase-seq ([Boyle et al., 2008]), le Formaldehyde-Assisted Isolation of Regulatory Elements followed by sequencing ou FAIRE-seq ([Giresi et al., 2007]) et l'Assay for Transposase-Accessible Chromatin followed by sequencing ou ATAC-seq ([Buenrostro et al., 2013]) sont trois techniques adaptées à l'étude de l'accessibilité de la chromatine. Le DNase-seq repose sur l'utilisation d'une endonucléase, la DNase I, capable de fragmenter le génome uniquement dans les régions ouvertes (Figure 1.33).

Le FAIRE-seq repose sur le principe que la chromatine fixée par du formaldéhyde peut être extraite de la chromatine composée d'ADN après fragmentation (Figure 1.33).

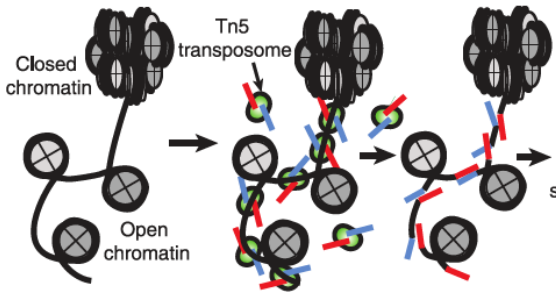


(a) Principe du DNase-seq (b) Principe du FAIRE-seq

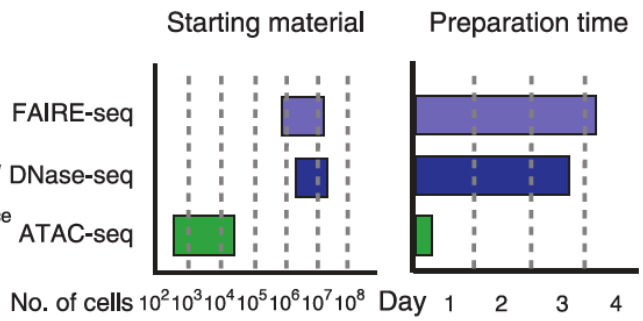
FIGURE 1.33 – (a) Principe du DNase-seq ainsi que du (b) FAIRE-seq.

Enfin, l'ATAC-seq est basé sur l'utilisation d'une transposase mutée, hyperactive, qui va fragmenter les régions ouvertes de la chromatine en insérant directement les séquences adaptateurs nécessaires au séquençage (Figure 1.34).

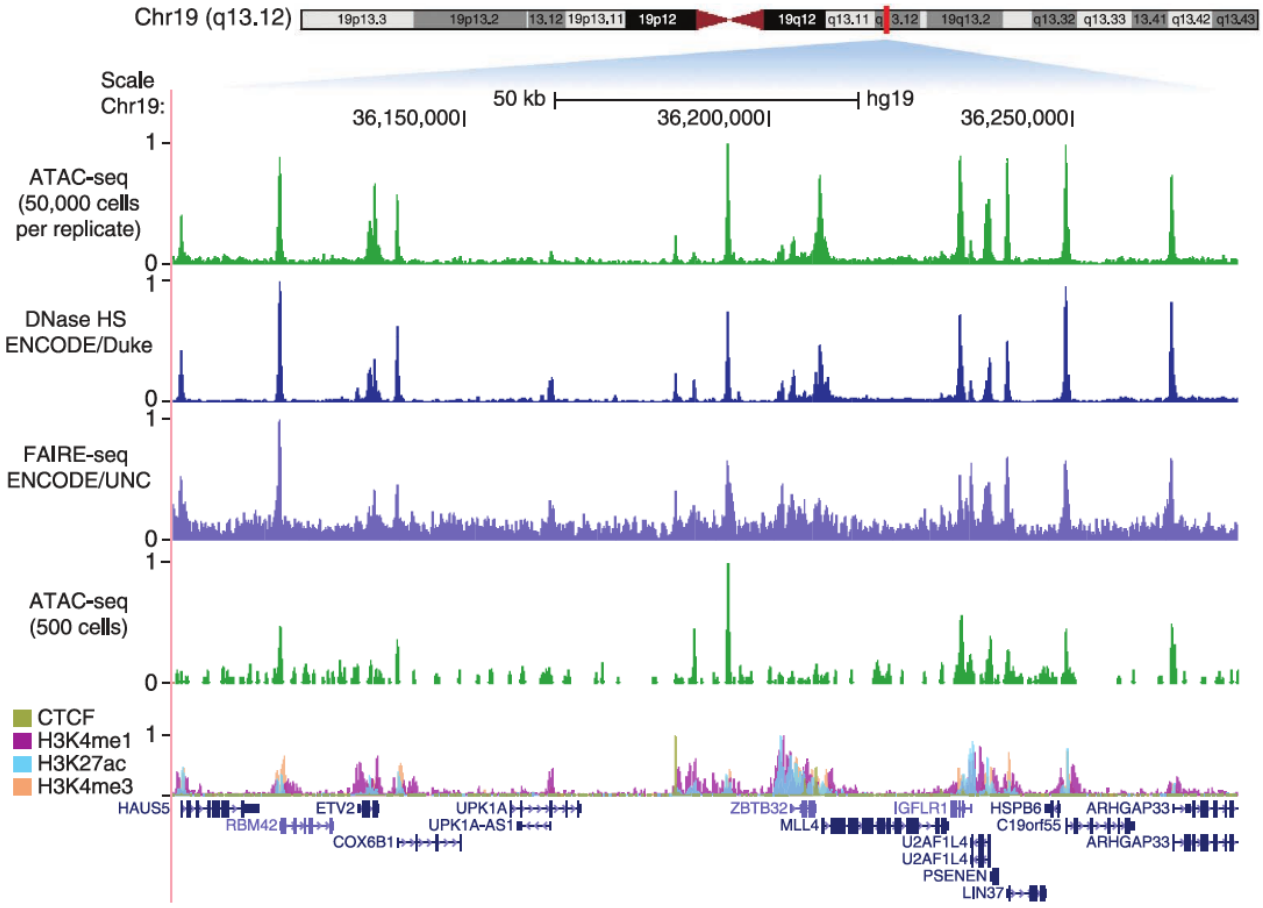
Les protocoles expérimentaux des trois techniques diffèrent. L'ATAC-seq prend l'avantage car il est plus rapide et moins exigeant en quantité de matériel biologique (Figure 1.34). Les étapes d'analyses des données restent similaires à celles du ChIP-seq et les profils obtenus par les trois approches demeurent sensiblement identiques (Figure 1.34). L'ATAC-seq ayant tendance à isoler également les fragments de nucléosomes des régions ouvertes, il peut être utile de filtrer les fragments de tailles voisines à celles d'un nucléosome afin d'obtenir le paysage nucléosomale local en alternative des données MNase-seq (Figure 1.34, [Buenrostro et al., 2013]).



(a) Principe de l'ATAC-seq



(b) Prérequis des protocoles



(c) Comparaison de profils obtenus par DNase-seq, FAIRE-seq et ATAC-seq pour deux quantité de matériel biologique.

FIGURE 1.34 – (a) Principe de l'ATAC-seq et comparaison comparaison du protocole (b) avec le FAIRE-seq et le DNase-seq. L'ATAC-seq est moins gourmand en matériel biologique et en temps.(c) Comparaison des profils obtenus en fonction des différentes technologies ([Buenrostro et al., 2013]).

1.3.2 Les outils et méthodes bio-informatiques pour l'étude des éléments régulateurs

1.3.2.1 Analyse de séquences

Depuis de nombreuses années, l'analyse de séquences est largement utilisée afin d'identifier la présence récurrente de certaines séquences dans le génome. Ces séquences peuvent être utilisées afin de prédire la position de divers éléments génomiques, tel que la recherche du trinucléotide

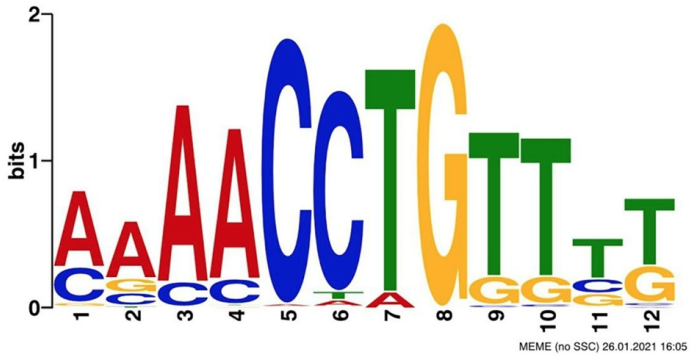
ATG des cadres ouverts de lecture (ORF, Open Reading Frame). Aujourd'hui, l'analyse de séquences est utilisée afin d'identifier et caractériser les éléments cis-régulateurs.

Les promoteurs peuvent par exemple être caractérisés et en partie prédits par la recherche d'éléments particuliers dans le génome, comme les boîtes TATA ou les îlots CpG. La recherche de ces éléments dans le génome est d'ailleurs particulièrement efficace pour identifier les gènes de ménage [Bajic et al., 2004]. Il existe aujourd'hui des bases de données comme EPD (Eukaryotic Promoter Database) [Dreos et al., 2015] et DbTSS (Database of Transcriptional Start Sites) [Suzuki et al., 2015] qui sont constituées de promoteurs dont les éléments du promoteur de base ont été expérimentalement vérifiés. Ces bases de données de haute qualité peuvent être utilisées comme référence et/ou jeu d'entrainement (training set) pour le développement d'outils de prédiction de promoteurs dans les génomes.

Concernant les enhancers et les silencers, aucune séquence nucléotidique n'est réellement conservée, autre que celles des motifs de fixation des facteurs de transcription qui sont souvent dégénérées. Mais là aussi de nombreuses bases de données ont vu le jour certaines basées sur des modèles prédictifs et/ou des données validées expérimentalement comme SilencerDB [Zeng et al., 2021], ENdb [Bai et al., 2020] ou encore VISTA Enhancer [Visel et al., 2007]. Comme indiqué, certaines vont se baser sur de la prédiction. Cette prédiction tend à devenir de plus en plus fiable notamment grâce aux avancées en intelligence artificielle. D'autres vont recueillir des données expérimentales de ChIP-seq par exemple qui vont venir valider une fixation protéique à une séquence. Ces bases de données appliquent un effort de curation rigoureux avant d'être mises à disposition de la communauté publique.

1.3.2.2 La recherche de motif

Elle fait partie intégrante de l'analyse de séquence. Elle consiste en la recherche de séquences de fixation des facteurs de transcription. Les sites de fixations peuvent être représentés par des séquences consensus ou des matrices poids-position. Les séquences consensus utilisent le code IUPAC (International Union of Pure and Applied Chemistry) dont chaque lettre représente plusieurs alternatives possibles d'un nucléotide à une position donnée du motif. Les matrices poids-position (PWM, Position Weight Matrix) permettent quant à elle de décrire de façon plus précise les motifs de fixation (Figure 1.35). Ces matrices contiennent les vraisemblances d'observer un nucléotide particulier à une position spécifique du motif.



	1	2	3	4	5	6	7	8	9	10	11	12
A	211	213	283	272	4	10	23	0	1	2	8	7
C	98	48	40	46	320	300	0	2	6	7	69	11
G	12	54	0	4	1	3	0	324	53	53	62	109
T	5	11	3	4	1	13	303	0	266	264	187	199

FIGURE 1.35 – La séquence consensus représente la probabilité d'un nucléotide à chaque position que l'on peut représenter sous forme de matrice [Leiz et al., 2021].

Il existe actuellement plusieurs bases de données qui répertorient les motifs de fixation des facteurs de transcription tel que TRANSFAC [Wingender et al., 1996, Matys et al., 2006], JASPAR ([Sandelin et al., 2004, Fornes et al., 2020]) ou encore UniPROBE [Newburger and Bulyk, 2009]. Grâce aux avancées technologiques, ces bases de données sont de plus en plus fournies, permettant des analyses de plus en plus complètes. En effet, on observe a une accumulation des connaissances qui ne cessent de s'accroitre.

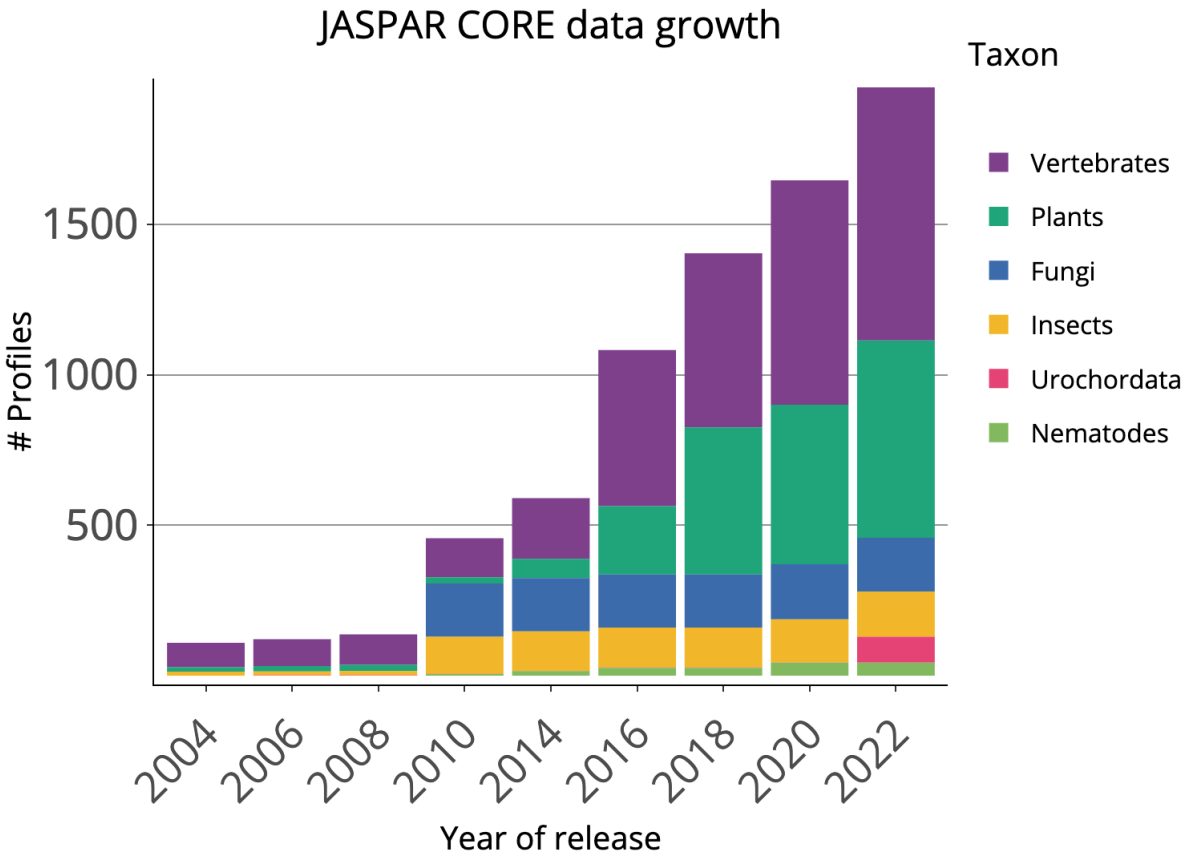


FIGURE 1.36 – Accumulation des données à chaque nouvel version de la base de donnée JASPAR [Castro-Mondragon et al., 2022].

L'existence de bases de données de motifs de fixation donne ensuite la possibilité de rechercher les occurrences de ces motifs dans des séquences d'ADN. De nombreux outils de recherche de motifs ont ainsi été développés et dont les performances varient. La qualité des matrices poids-position utilisées est un élément crucial dans ce type d'analyse. En effet, ce sont elles qui définissent la probabilité d'observer un site de fixation à une position donnée dans une séquence d'ADN. Parmi les outils d'identification de séquences de fixation des facteurs de transcription les plus populaires on compte RSAT [Thomas-Chollier et al., 2008] développé en partie par des membres du TAGC dont Jacques van Helden, la suite MEME [Bailey et al., 2009] ou encore HOMER [Heinz et al., 2010].

JASPAR a également mis en place en 2022 une méthode afin d'analyser l'enrichissement en facteurs de transcriptions dans un jeu de données génomiques avec la possibilité de comparer plusieurs jeux de données [Castro-Mondragon et al., 2022].

Des outils ont également vu le jour afin de permettre une visualisation des facteurs de transcription au sein de région spécifique. On peut par exemple citer TFmotifView, une page web permettant facilement d'interroger le programme sur l'enrichissement en facteurs de transcription au sein de nos régions d'intérêt, renvoyant des figures (scatterplot, visualisation de la distribution des facteurs de transcription au sein de nos régions) ainsi qu'une table de donnée renvoyant l'enrichissement des facteurs de transcription [Leporcq et al., 2020].

1.3.2.3 Analyse intégrative : le multi-omique

Dans ce chapitre, nous avons pu citer différentes techniques de séquençage, différents types d'analyses de séquence. À l'heure actuelle, nous avons une réelle richesse des données. En couplant les données issues de différentes technologies, de différents jeux de données, il sera possible de mener des analyses plus profondes et d'exploiter les relations entre les différents régulateurs.

Ce raisonnement a vu naitre des bases de données à plusieurs dimensions, tel que MethMotif, par exemple, qui va intégrer l'information des sites de fixation des facteurs de transcription (technologie : ChIP-seq) avec le niveau de méthylation de l'ADN (technologie : WGBS : whole genome bisulfite sequencing) (Figure 1.37,[Xuan Lin et al., 2019]).

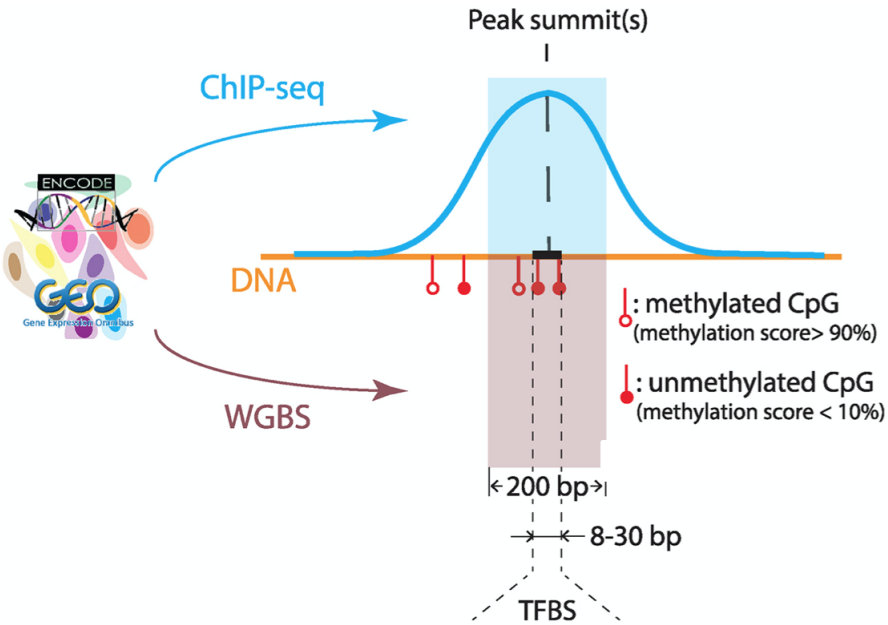


FIGURE 1.37 – Methmotif va coupler les informations de la position génomique (ChIP-seq) avec le niveau de méthylation de cette région [Xuan Lin et al., 2019].

Dans un sens général, le terme multi-omique désigne l'exploitations de différents types de données (transcriptome, épigénome, etc.) pour avoir une image plus complète d'une situation biologique [Subramanian et al., 2020]. Il va s'agir de combiner des informations de différentes natures, telles que la présence d'une protéine avec une marques d'histones, un niveau d'expression génique avec une région d'intérêt, etc. Ce genre d'analyse va nécessiter de croiser différents jeux de données parfois simplement en comparant les chevauchements d'éléments entre deux fichiers.

De plus en plus d'approches multi-omiques utilisent des méthodes d'apprentissage automatisé basé sur l'intelligence artificielle. Parmi ces méthodes, les factorisations matricielles sont très populaires. Cela implique souvent des factorisations personnalisées de matrice représentant chaque jeu de données, avec un objectif d'optimisation visant à la fois à expliquer l'ensemble des données d'origine afin de mettre en évidence des points communs.

1.4 Approche à grande échelle pour l'étude fonctionnel de séquences Cis-régulatrice

Le contrôle de l'expression des gènes nécessite plusieurs niveaux de contrôle tel que l'accessibilité de la chromatine, les mécanismes de modifications épigénétique, la fixation des facteurs de transcriptions et cofacteurs. Tous ces éléments conduisent à la communication entre séquences régulatrices et cible de régulation.

Les techniques précédemment présentées vont pouvoir permettre une identification de certains éléments tel que les enhancers ou les promoteurs ou d'étudier l'accessibilité de la chromatine. Cependant, elles vont vite montrer leur limite pour l'étude des éléments répresseurs tel que les silencers. De plus, ces techniques ne fournissent pas une lecture fonctionnelle des séquences identifiées. Ainsi, de nouvelles méthodes ont récemment émergé surmontant bon nombre des limitations des technologies précédemment citées. Celles-ci incluent des méthodes à cellule unique ("Single Cell") pour mener des recherches spécifique à un type de cellule dans des tissus complexes (exemple : identification de la chromatine ouverte spécifique à un type de cellule [Cusanovich et al., 2015, Buenrostro et al., 2015]), des méthodes de capture de la conformation chromosomique à haute résolution ("3C") qui vont permettre de cartographier plus finement les interactions régulateurs / cibles [Eagen, 2018] et des méthodes permettant le test fonctionnel de séquences à grande échelle tel que le MPRA qui va disséquer ou piéger l'activité de la région ou le STARR-seq que nous présenterons ultérieurement plus en détail. Plus récemment la technologie du CRISPR est utilisée à grande échelle afin de perturber des régions régulatrices dans leur contexte génomique et relier leurs cibles, et l'impact qu'ils auront sur la cellule [Klein et al., 2018]. Ces méthodes peuvent être qualitatives (généralement basées sur le tri cellulaire) ou quantitatives (basées sur le RNA-seq) et conçues pour tester l'activité de l'enhancer ou du promoteur. Ainsi, ce chapitre résumera les tests récents pour l'analyse fonctionnels de l'activité des éléments cis-régulateur.

Couplé ces technologies aux récentes innovations dans le domaine de l'informatique tel que l'apprentissage automatisé reposant sur l'intelligence artificielle et vous obtiendrez de puissants outils de détection, prédiction, et caractérisation des éléments cis-régulateurs.

1.4.1 Les tests fonctionnels basé sur un gène rapporteur

1.4.1.1 Méthodes à bas débit

Une méthode expérimentale répandue permettant d'identifier les éléments régulateurs est la méthode des gènes rapporteurs ([Liu et al., 2009]). Le but de cette méthode est de mesurer l'expression d'un gène en fonction de son contexte génomique direct. Pour cela, un gène, dont l'expression peut être suivie facilement (ex : luciferase, green fluorescent protein : GFP), et une région du génome contenant un élément régulateur sont insérés dans un plasmide.

La comparaison entre l'expression du gène en présence ou en l'absence de la région régulatrice peut être observée de façon directe (Figure :1.38). La méthode du gène rapporteur permet d'étudier la plupart des éléments régulateurs, le promoteur proximal et/ou distal, les enhancers, les silencers et les insulators, mais elle comporte certaines limites. Cette technique est souvent utilisée en validation individuelle. Il est nécessaire d'utiliser d'autres méthodes en parallèle afin d'identifier et d'annoter les éléments régulateurs dans le génome humain.

1.4.1.2 Méthodes à haut débit

Ces stratégies permettent l'analyse simultanée de centaines de milliers de plasmides rapporteurs à la fois (tableau : 1.3) et ont fait l'objet de plusieurs revues les référençant [Santiago-Algarra et al., 2017,

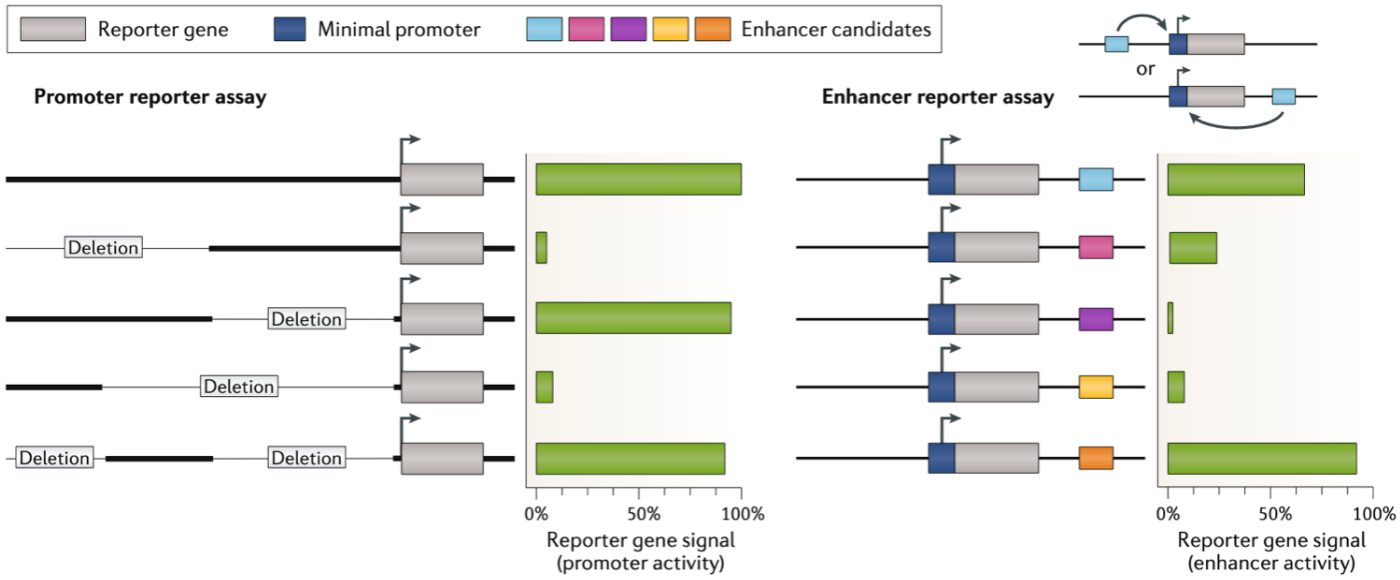


FIGURE 1.38 – Schéma du test de séquence par gène rapporteur. La région testée est placé en amont d'un gène rapporteur. Le promoteur activera le gène et une quantification du signal sera possible (panel de gauche). Il est possible d'effectuer des délétions afin de detecter la sous région responsable du signal. Le test d'enhancer fonctionne de manière similaire (panel de droite). Un enhancer est inséré en amont d'un gène rapporteur avec un promoteur minimal qui a besoin d'une activité enhancer pour être transcrit[Andersson and Sandelin, 2020].

Dailey, 2015, Inoue and Ahituv, 2015, White, 2015, Gasperini et al., 2020]. Il existe deux méthodes majeures à haut débit pour tester l'activité de milliers de séquences régulatrices en une seule expérience. Il s'agit du MPRA (Massively Parallel Reporter Assays) [Patwardhan et al., 2012, Melnikov et al., 2012] et du STARR-seq (Self-Transcribing Active Regulatory Region Sequencing) développé par Alexander Stark et ses collaborateurs [Arnold et al., 2013] (Figure : 1.39).

1.4.1.2.1 MPRA La méthode MPRA a été développée à l'origine pour tester l'activité de promoteur et l'effet de la mutagenèse sur base unique dans trois promoteurs de bactériophages in vitro [Patwardhan et al., 2009]. Par la suite cette méthode a été utilisé afin de tester l'activité activatrice de trois enhancers de mammifères et ses variants in vivo [Patwardhan et al., 2012].

Ce test utilise de l'ADN synthétisée par microarray qui contient l'élément régulateur et une étiquette de séquence unique ou "barcode" / "tag" (Figure :1.40). Ces séquences synthétisées par l'ADN sont clonées dans le plasmide. Ensuite, un promoteur minimal et un gène rapporteur sont insérés entre les régions testées et le tag, laissant ce dernier dans le 3 'UTR. L'activité de n'importe quelle séquence transcrira le tag qui lui est associée. Une fois la librairie MPRA finalisée, elle est transfectée dans des cellules et le séquençage de l'ARN des transcrits/tags est effectué. L'activité de toute séquence est directement proportionnelle à l'enrichissement du tag. Le MPRA peut se révéler extrêmement résolutif. Avec une bonne dose d'ingéniosité, une équipe a développé une approche MPRA haute résolution (également appelée Sharpr-MPRA) qui a permis la cartographie à l'échelle du génome des nucléotides activateurs et répressifs dans les régions régulatrices [Ernst et al., 2016]. Ici, en synthétisant des séquences chevauchantes dense de constructions MPRA, ils ont réussi à déduire l'effet des nucléotides régulateurs fonctionnels avec des propriétés activatrices ou répressives.

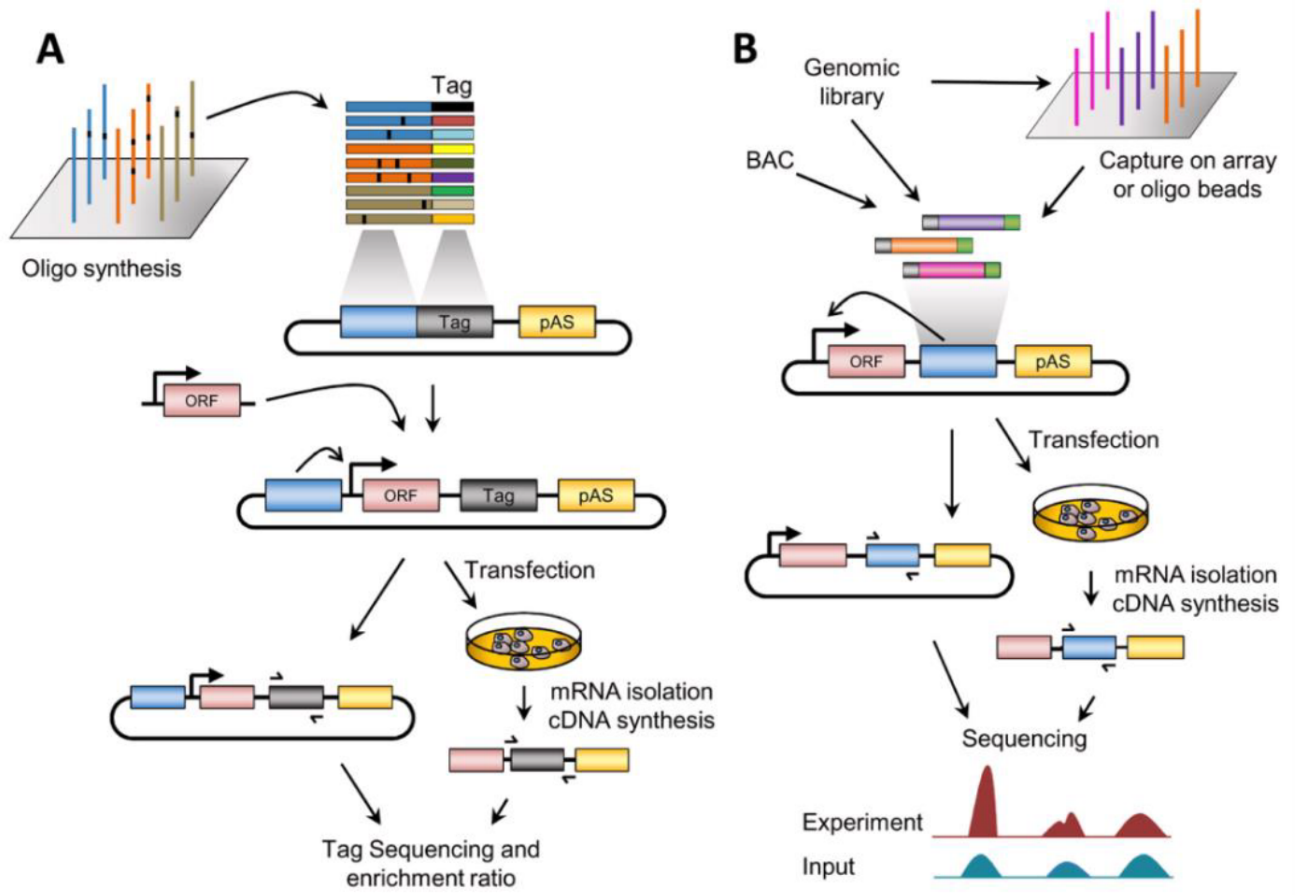


FIGURE 1.39 – Schéma explicatif des deux principales méthodes de test fonctionnel via séquenceur à haut débit avec le MPRA en (A) et le STARR-seq en (B). La différence majeure réside dans le positionnement de la séquence testé. (A) Dans le MPRA, la séquence ne fait pas parti du rapporteur d'activité contrairement au (B) STARR-seq où elle fait partie intégrante du rapporteur d'activité.

Le MPRA pourra être détourné de plusieurs façon en fonction de l'usage souhaité. Par exemple il peut être utiliser pour tester l'impact des polymorphismes mononucléotidiques (SNP) afin d'identifier des variantes de régulation fonctionnelles liées à des traits particuliers ou à des maladies.

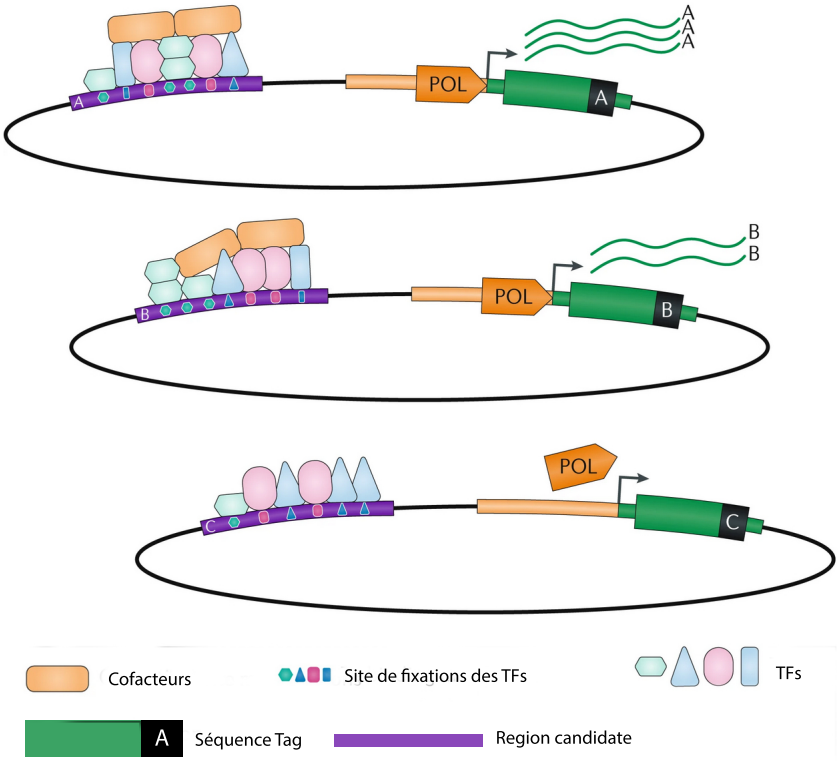


FIGURE 1.40 – Schéma représentant le principe du MPRA avec la région testée, le gène rapporteur situé entre la région testée et le tag [Gasperini et al., 2020]).

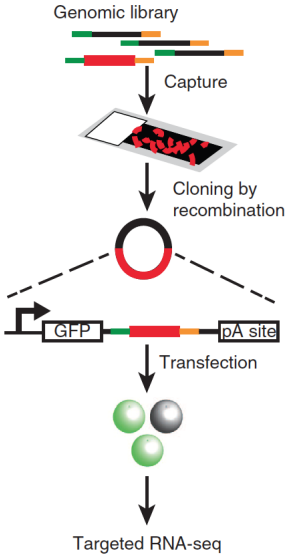
1.4.1.2.2 STARR-seq Une méthode alternative au MPRA est le STARR-seq, un test à haut débit introduit afin d'identifier les activateurs transcriptionnels en fonction de leur activité dans l'ensemble du génome de la drosophile et évaluer quantitativement leur activité [Arnold et al., 2013].

Cette méthode ne nécessite pas de "code barre" synthétisés puisque les séquences d'ADN sont clonées dans le 3 'UTR du gène rapporteur. La séquence active transcrira le gène rapporteur et lui-même, devenant une partie du transcrit rapporteur. Ainsi, l'activité de toute séquence enhancer constitue un enrichissement du transcrit contenant le "code barre" et sa séquence. Ces transcrits peuvent être isolés et détectés par séquençage à haut débit (Figure :1.39).

Initialement utilisé chez la drosophile, cette méthode a été appliquée aux mammifères [Muerdter et al., 2018, Peng et al., 2020] ; cependant, la complexité du génome peut être un facteur limitant et augmenter le coût des expérimentations et du séquençage de façon considerable.

Réduction de la compléxité : L'avantage de cette technique est qu'elle peut être déclinée en plusieurs versions en fonction de l'objectif de recherche. Ceci a également pour avantage de réduire le coût important de la technique de base. En général, ces méthodes surmontent les limitations de l'étape de la synthèse dans les méthodes MPRA et réduisent la complexité et le poids et ainsi le coût de la librairie génomique.

1.4.1.2.3 CapSTARR-seq Dans notre laboratoire une variante du STARR-seq a vu le jour. Il s'agit d'une approche basée sur la capture de fragments (CapSTARR-seq) afin d'évaluer un sous-ensemble de sites. Cette technique a été développé afin d'évaluer l'activité des sites hypersensibles à la DNase I (DHS) de souris [Vanhille et al., 2015] et l'ensemble des promoteurs codant pour les protéines humaines [Dao et al., 2017]. Après enrichissement par capture, les fragments d'ADN sont clonés dans le vecteur du STARR-seq (3'UTR du gène rapporteur) et transfectés dans des cellules (Figure :1.41). Cette méthode fournit une approche rapide et rentable



Principe du CapSTARR-seq

FIGURE 1.41 – Principe du CapSTARR-seq et identification d'enhancers à grande échelle [Vanhille et al., 2015]

pour évaluer l'activité de séquences régulatrices potentielles pour un type de cellule donné et aidera à décrypter les mécanismes de régulation de la transcription [Vanhille et al., 2015].

Cette technique a permis d'identifier avec succès tous les enhancers validés *in vivo* par des souris knock-out tel que Ikzf1, TCRα, TCRβ, ou encore GATA3 (Figure : 1.41). C'est d'ailleurs pour cela que les cellules du système immunitaire représente un bon modèle pour l'étude des séquences cis-régulatrices. Elles permettent de pouvoir générer des souris KO viable afin de valider des résultats [Vanhille et al., 2015] expérimentaux.

L'avantage de cette technique est qu'ici chaque séquence de lecture obtenu en sortie du séquenceur à haut débit (read) va consister en un fragment préalablement capturé. Ainsi on va être à même de caractériser l'activité par fragments en regroupant les reads possédant des coordonnées strictement identiques. D'autant plus qu'on sait qu'on aura une forte densité de fragments dans les régions d'intérêts. Ces fragments se chevaucheront parfois d'un nucléotide (ou plus), provoquant une cascade de fragments en escalier. Si on applique le raisonnement qu'on a pu avoir pour le SharpMPRA [Ernst et al., 2016], cela nous permettra d'être extrêmement résolutif sur l'activité des régions d'intérêts. Cette méthode a principalement été utilisé pour l'analyse de régions activatrices telles que les enhancers ou les promoteurs [Dao et al., 2017], [Santiago-Algarra et al., 2021]. Dans notre cas nous avons réutilisé cette technique afin d'étudier les silencers.

1.4.1.2.4 ATAC-STARR-seq Kellis et Claussnitzer ont développé une stratégie appelée Hi-DRA (High-resolution Dissection of Regulatory Activity) [Wang et al., 2018]. Cette méthode combine ATAC-seq et STARR-seq offrant l'avantage de tester l'activité de fragments en chromatine ouverte. Avec cette méthode, ils ont identifié 65 000 régions présentant une fonction d'enhancer dans les cellules lymphoblastoïdes GM12878. De plus, ils ont pu identifier les sous-unités fonctionnelles des enhancers [Wang et al., 2018].

1.4.1.2.5 FAIRE-STARR-seq Singh et ses collaborateurs ont couplé le FAIRE-seq avec la technique du STARR-seq afin d'évaluer l'activité enhancer sur un ensemble complexe de régions présentant un état décompacté dans les cellules B activées par le LPS [Chaudhri et al., 2020].

Cette technique leur permet d'identifier de nombreuses séquences activatrices localisées dans les régions promotrices proximales, intergéniques et intragéniques.

1.4.1.2.6 BAC Une autre stratégie afin de restreindre les séquences d'ADN cible est l'utilisation de Chromosomes Artificiels Bactériens (BAC). Au sein de ces BAC, les régions ciblées par l'ADN peuvent être clonées dans un vecteur STARR-seq et testées pour la fonction d'enhancer. Cette technique a permis d'identifier la spécificité de certains enhancers pour un core [Zabidi et al., 2015]. Une autre étude a utilisé des BAC pour interroger les locus associés au GWAS à la fibrillation auriculaire chez l'homme et identifier plusieurs éléments régulateurs avec des variants associés à la fibrillation auriculaire [van Ouwerkerk et al., 2020].

1.4.1.2.7 ChIP-STARR-seq La fonction d'enhancer nécessite la liaison du facteur de transcription et une chromatine accessible marquée par des modifications d'histone. Par conséquent, l'utilisation de la stratégie ChIP-STARR-seq, qui intègre l'immunoprécipitation de la chromatine avec le STARR-seq, pourrait potentiellement identifier des enhancers fonctionnels. En utilisant cette méthode, Chambers et ses collaborateurs ont identifié que seule une minorité de régions marquées par NANOG, OCT4, H3K27ac et H3K4me1 fonctionnent comme enhancers dans les cellules souches embryonnaires, indiquant qu'aucun facteur de transcription individuel, marque d'histone ou une combinaison de ceux-ci ne pourrait prédire sans équivoque l'activité de l'enhancer [Barakat et al., 2018].

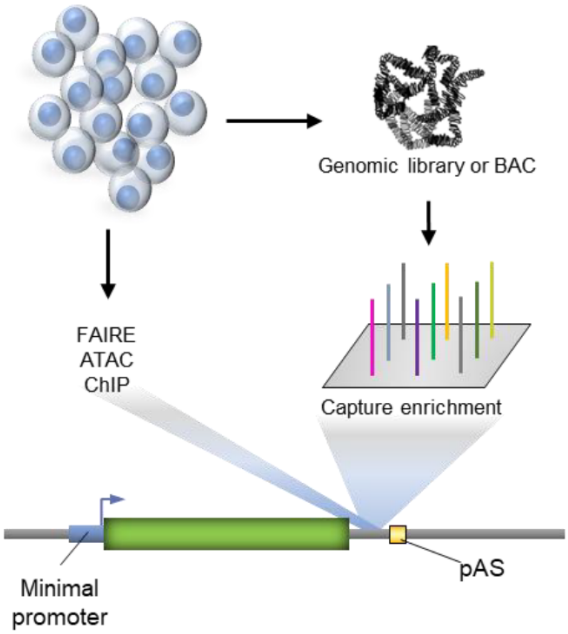


FIGURE 1.42 – Afin dé réduire la complexité et le coût du STARR-seq, différentes variantes ont pu être mise en place afin de cibler des régions particulières du génome

Ces différentes techniques permettent la caractérisation des séquences cis-régulatrices à grande échelle. MPRA et STARR-seq présentent des avantages différents l'un par rapport à l'autre. Par exemple, MPRA a été utilisé pour tester l'impact de milliers de polymorphismes mononucléotidiques (SNP) dans les éléments régulateurs, tandis que STARR-seq peut être utilisé pour tester des variants naturels provenant d'amplifications PCR ou effectuer simplement une analyse de région à grande échelle sans à priori. Dans le MPRA, la taille des séquences cibles est limitée en raison de l'étape de la synthèse qui restreint la taille du fragment à environ 170 pb. L'avantage du STARR-seq est l'élimination des tags puisque les séquences peuvent provenir de fragment du génome par exemple des sites sensibles à la DNase I, des régions

obtenues via l'ATAC, des régions obtenues via FAIRE, des régions capturées par ChIP,etc. Récemment, les laboratoires de Shendure et Ahituv ont fait une évaluation de plusieurs stratégies utilisant le STARR-seq et le MPRA [Klein et al., 2020] Leurs résultats montrent des corrélations élevées entre le test rapporteur classique, le STARR-seq et le MPRA. Ils suggèrent également d'augmenter la taille des fragments tester à 600 pb contre 190 pb actuellement afin d'obtenir un signal plus pertinent sur le plan biologique.

TABLE 1.3 – Exemple récent d'étude ayant fait appel à des techniques de test de gène rapporteur à grande échelle.

Technique	DNA origin	Application	Specificity	No. of regions[1]	Size (bp)	Cell lines	Promoter	Species	Ref.
Lenti-MPRA	Synthetic	Identification of regulatory dynamics during neural differentiation	Centered on ChIP	2,463	171	hPSCs differentiation	Minimal promoter	Human	(Inoue et al., 2019)
STARR-seq	Genomic	Use of STARR plasmid in mammalian cells	Whole-genome	71,968	1000-1500	HeLa-S3	Minimal promoter and ORI	Human	(Muerdter et al., 2018)
MPRA	Synthetic	Analysis of promoter variants in lncRNAs that affect the expression	-80 to +34 bp from TSS of LncRNAs	2,078	114	HeLa, HepG2 and K562	Minimal promoter	Human	(Mattioli et al., 2019)
MPRA	Synthetic	Test of 81 eQTL SNPs associated with CD36	Centered on the SNP	1	150	K562	Minimal promoter	Human	(Madan et al., 2019)
STARR-seq	Genomic	Analysis of variants in cancer risk loci.	Genomic capture centered on the SNP	996	500	HEK293T	SCP1[2]	Human	(Liu et al., 2017a)
ChIP-STARR-seq	Genomic	Identification of enhancers in human embryonic cells	Genomic capture using ChIP	361,737	600	hESCs	Minimal promoter	Human	(Barakat et al., 2018)
STARR-seq	Synthetic	Characterization of transcriptional regulation of cancer risk loci.	Centered on the SNP	374	21	LNCaP	SCP1[2]	Human	(Zhang et al., 2018)
ATAC-STARR-seq	Genomic	Genome-wide testing of putative regulatory regions	ATAC accessible regions	$7x10^6$	150-500	GM12878	SCP1[2]	Human	(Wang et al., 2018b)
BiT-STARR-seq	Synthetic	Characterize variants with allele-specific effects in regulatory regions	Centered on the SNP	43,500	200	GM18507	Minimal promoter	Human	(Kalita et al., 2018)
STARR-seq	Genomic	Genome-wide testing of putative regulatory regions	FAIRE accessible regions	55,133	100-400	Splenic B cells	Minimal promoter	Mouse	(Chaudhri et al., 2020)
STARR-seq	Genomic	Genome-wide characterization of enhancers in two pluripotent states	Genomic DNA sonicated	48,311	700-1200	mESCs	SCP1[2]	Mouse	(Peng et al., 2020)
Methyl STARR-seq	Genomic	Assess the causal effect of mDNA on the regulatory activity of millions of CpG sites.	*MspI*-digested DNA	262,829	300-700	K562	CpG-free EF1	Human	(Lea et al., 2018)
STARR-seq	Genomic	Characterization of variants associated with atrial fibrillation	GWAS loci cloned in BACs	12	450-900	Myocytes	SCP1[2]	Rat	(van Ouwerkerk et al., 2020)

[1] Number of targeted DNA sequences; not necessarily the number of unique fragments that are tested.
[2] Synthetic Super Core Promoter-1

1.4.2 L'intelligence artificielle au secours du traitement des données génomiques

1.4.2.1 Généralité sur le Machine Learning

L'apprentissage automatique ou "Machine Learning (ML)" est défini comme l'utilisation d'algorithmes informatiques qui s'améliorent par l'apprentissage. En général, le machine learning consiste à construire un modèle mathématique avec une structure prédéterminée basée sur des données issues d'échantillons qui vont être la base de l'apprentissage. Il est utilisé afin de réaliser des prédictions sur d'autres données qui étaient "caché" aux yeux du modèle dans le but de limiter les erreurs commises sur ces prédictions. L'apprentissage automatique a un large

éventail d'applications, de l'analyse commerciale prédictive au diagnostic par analyse d'image par ordinateur et filtrage des e-mails.

Le terme machine learning a été inventé en 1959 par Arthur Samuel. Il est parfois considéré comme un sous-ensemble du domaine de l'intelligence artificielle. Le machine learning est également née des statistiques.

De ces domaines est née la science des données ou "Data Science". Il y a souvent une confusion dans l'inconscient commun entre le Big Data et le ML en raison du fait que, en ML, plus il y a de données significatives sur lesquelles apprendre, meilleure est l'apprentissage. Dans cette partie, nous présentons des généralités sur la classification et le but des approches de ML.

1.4.2.1.1 Classification des méthodes de machine learning Les approches d'apprentissage automatisé peuvent être divisées en quatre classes principales. D'abord et le plus commun est l'apprentissage supervisé, où les données de formation contiennent des entrées et des sorties souhaitées. Cela inclut des méthodes comme les classifications et la régression. Parallèlement il y a l'apprentissage semi-supervisé où seulement quelques éléments des données sont manquants. Cependant, dans l'apprentissage non supervisé, les exemples n'ont pas de label. Le but de cette approche est principalement d'essayer de trouver une structure, une tendance, qui se dégage dans les données présentées en entrée, comme avec l'algorithme de clustering. Nous allons présenter ici une liste non exhaustive des méthodes de ML les plus utilisé et les plus populaires.

— Les méthodes de régression, telles que les régressions linéaires ou polynomiales où l'hypothèse est une combinaison linéaire ou polynomiale.

— Les arbres de décisions sont des arborescences utilisant des seuils successifs sur les données d'entrées afin de trier les données dans des 'boites' composées autant que possible de données similaires.

— Support Vector Machines (SVM) recherche un hyperplan de séparation entre deux classes prédéfinies. Cela repose sur le calcul de distance entre les données en entrée à l'aide d'une fonction pour produire une matrice de distances. En conséquence, l'espace de représentation a été transformé en un espace dimensionnel plus grand où un classificateur linéaire peut être utilisé.

— La factorisation matricielle sépare la matrice de données en un produit d'autres matrices dont les composants sont significatifs d'une certaine manière.

— Les méthodes ensemblistes consistent à regrouper plusieurs modèles parmi ceux présentés ici. Par exemple, la méthode Random Forest consistent à utiliser plusieurs arbres de décision entraînés sur différents sous-ensembles aléatoires de données, suivis d'un vote à la majorité.

1.4.2.1.2 Deep learning Basé sur un principe différent mais faisant partie du machine learning, je présenterais très brièvement le deep learning. Souvent dans le Machine Learning, l'extraction et la structuration des données sont manuelles et nécessite l'intervention d'un individu. Le deep learning lui va pouvoir être appliquer sur des données non structurées et appliquer un modèle en " couches de neurones " permettant d'identifier les informations "cachées" (Figure : 1.43)

Les réseaux de neurones (Deep Neural Networks - DNNs) sont de puissants systèmes paramétriques non linéaire. Globalement, on peut les schématiser comme un assemblage de régressions logistique, dont la sortie est transmise à d'autres régressions ultérieures. Parmi les algorithmes de deep learning on peut citer les **CNN (Convolutional Neural Networks)**. Brièvement, l'idée ici est d'avoir des filtres à chaque couches (dimensions) de neurones où chaque filtre individuel apprendra une petite combinaison locale d'éléments. Le but de ces réseaux est d'apprendre des combinaisons locales d'éléments à travers les dimensions couvertes par le filtre.

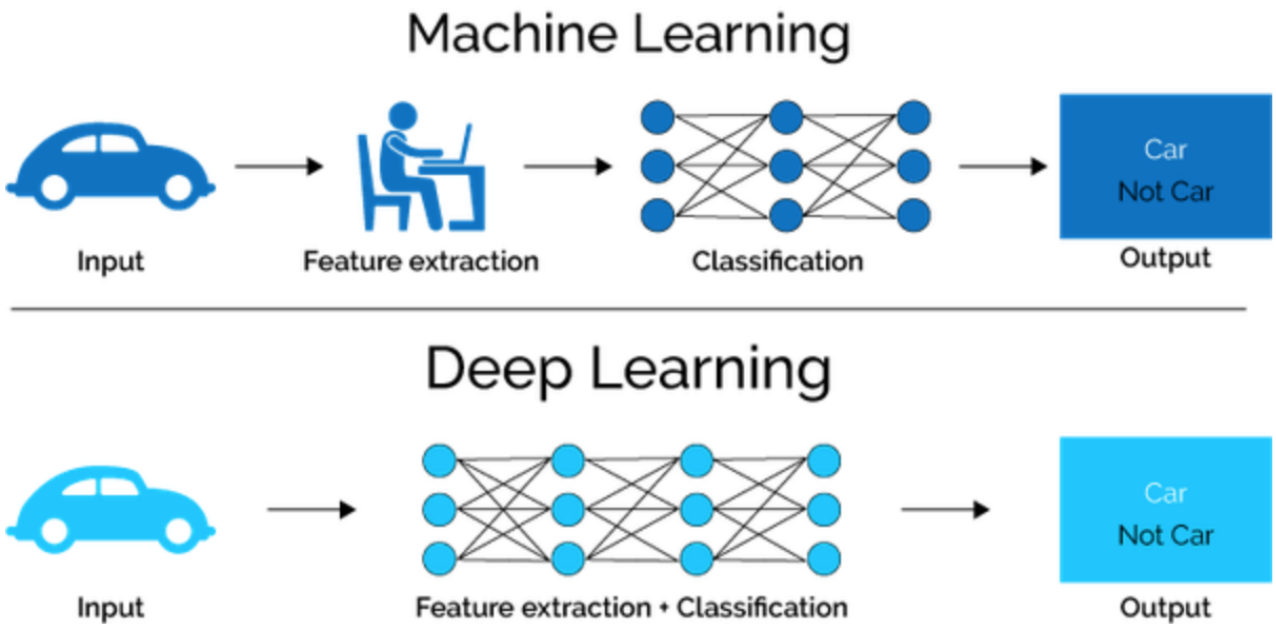


FIGURE 1.43 – Exemple récent d'étude ayant fait appel à des techniques de test de gène rapporteur à grande échelle de Jay Shah.

Un autre type d'architecture spécifique pour les réseaux de neurones sont les **auto-encodeurs**. Leur but est d'apprendre une représentation compressée (c'est-à-dire un encodage) des données d'entrée. Suite à cette compression, le modèle apprend à reconstruire les données d'entrée d'origine sur la base de la dimension codée. Il cherche à minimiser la différence entre l'entrée et la reconstruction, d'où le nom. Cette compression implique une diminution du bruit du signal.

La génomique n'a pas échappé à la révolution qu'est l'intelligence artificielle regroupant le Machine learning et le Deep learning. Une revue fait l'état des lieux des différentes méthodes qui peuvent être appliqué à de gros jeux de données génomiques [Eraslan et al., 2019].

Dans la partie qui suivra, je parlerai d'outils récents basés sur des méthodes d'apprentissage automatisé pour l'analyse de données génomiques.

1.4.2.2 Outils bio-informatiques pour l'analyse de séquences régulatrices basé sur le Machine Learning

Grâce à l'intelligence artificielle l'accumulation et le partage de données, nous sommes à l'aube d'une nouvelle ère. Dans les prochaines années, le rythme des découvertes devrait s'accélérer grandement. Même d'énormes groupes tel que **Google** à travers sa plateforme de recherche sur l'intelligence artificielle **DeepMind** et son outil **Alphafold** [Jumper et al., 2021, Tunyasuvunakool et al., 2021] participent à la recherche santé en utilisant des algorithmes basés sur le machine learning. En partenariat avec l'EMBL-EBI, ils ont développé une base de données AlphaFold sur la structure des protéines, qui offre une image complète et à haute résolution du protéome humain, doublant ainsi le volume de connaissances accumulées sur les structures protéiques humaines de haute précision. Il s'agit de la contribution la plus importante que l'intelligence artificielle ait apportée à l'avancement des connaissances scientifiques à ce jour, et c'est un excellent exemple des avantages que l'IA peut apporter.

Plus en rapport avec mon projet de thèse et l'étude des séquences cis-régulatrices, on peut notamment citer **Enformer** [Avsec et al., 2021a]. Cet outil basé sur le Deep learning (CNN) propose de prédire l'expression génique en se basant sur les interactions génomiques. Enformer est également capable de prédire les interactions enhancers-promoteurs en se basant directement sur des données expérimentales. En comparant leurs prédictions avec des données issues de

CAGE, ils ont pu observer une très bonne fiabilité des prédictions. Un autre usage intéressant d'Enformer est sa capacité à prédire l'effet d'une mutation au sein de la séquence régulatrice, ouvrant des voies pour l'étudie des variantes génétiques liées à des pathologies. Enformer est entrainé pour prédire les données génomiques fonctionnelles, y compris l'expression génique à partir de séquence génomique de 100kB en entrée.

La même année, en 2021, le même auteur a publié un second outil, **BPNet** [Avsec et al., 2021b], basé sur le même type d'algorithme (CNN), afin de prédire des séquences de fixation de facteur de transcription au sein de séquence régulatrice. Ainsi BPNet propose de modéliser la relation entre la séquence cis-régulatrice et les sites de liaison aux facteurs de transcription proposant un nouveau modèle pour la représentation des motifs au sein des séquences non pas basé sur une surreprésentation statistique mais qui va représenter directement la séquence responsable de la fixation du facteur de transcription. Il identifie les sous régions des motifs ("core motif") importantes pour la fixation des facteurs de transcription. Un fait intéressant soulevé par BPNet est l'existence de motif anormalement long dont le "core motif" correspondant au site clé de la fixation du FT semble étendu sur toute la longueur de la séquence. Cela implique que les instances génomiques de ces motifs partagent une composition de base presque identique sur toute la longueur du motif. Il s'avère que la majorité de ces éléments sont des séquences répétées, notamment des répétitions de virus à rétrotransposons endogènes (ERV). Plus généralement, il propose leur outil en alternative à des outils tel que Homer [Heinz et al., 2010] ou MEME [Bailey et al., 2009] pour la recherche de site de fixation de facteur de transcription soulevant le fait que les motifs issus de BPNet ont surpassé celles obtenues par des approches traditionnelles telles que le balayage des séquences basé sur les matrices poids-positions. Cet outil s'intéresse également à la combinaison de facteur de transcription au sein de la même région et s'intéresse à l'organisation de plusieurs sites de fixation des facteurs de transcription au sein d'une même région. Il propose d'y répondre en prédisant comment la distance entre les facteurs de transcription influent sur leurs façons de coopérer.

Cela rappel d'ailleurs le principe de **OLOGRAM MODL**, développé au sein de notre laboratoire par Quentin Ferre [Ferré et al., 2021]. OLOGRAM MODL s'intéresse aux chevauchement multiples (supérieur à 2) de régions génomiques et calcule leur enrichissement mutuel statistique par ajustement de Monte Carlo d'une distribution binomiale négative, résultant en des p-valeurs plus résolutives. Cet outil propose une méthode optionnelle basé sur du machine learning pour trouver des complexes d'intérêt. En effet, on sait que la plupart des composants de la chromatine tels que les régulateurs transcriptionnels ou les histones sont connus pour fonctionner en combinaisons et former des complexes lorsqu'ils se lient au génome. Ainsi OLOGRAM-MODL propose de retrouver au sein d'une multitude de jeux de données possédant des régions chevauchantes, des combinaisons d'éléments préférentiellement colocalisés, c'est-à-dire s'ils sont rencontrés plus que prévu par hasard.

Un autre outil innovant basé sur du Deep learning est **DeepSTARR** [de Almeida et al., 2021]. DeepSTARR propose une prédiction de l'activité enhancer à l'échelle du génome en se basant sur des données de STARR-seq. L'aspect innovant de cet outil réside dans le fait qu'ils ont été capable pour la première fois de prédire des séquences enhancers synthétique fonctionnelle ouvrant la voie à toutes sortes de nouvelles études. En effet, grâce à DeepSTARR, il va être possible de "créer" des séquences enhancers et de les tester fonctionnellement. DeepSTARR est basé sur un modèle CNN (Convolutional Neural Networks), les différentes couches vont identifier tout d'abord les caractéristiques de séquence locales (par exemple, les motifs de facteurs de transcription). À travers les différentes couches (Figure 1.44), le modèle va se complexifier afin de mettre en évidence des sous régions dirigeant l'activité des facteurs de transcription tandis que la totalité des couches entièrement connectées combinent ces caractéristiques et pour prédire l'activité de l'enhancer.

DeepSTARR a également mis en évidence l'importance des séquences flanquantes des facteurs

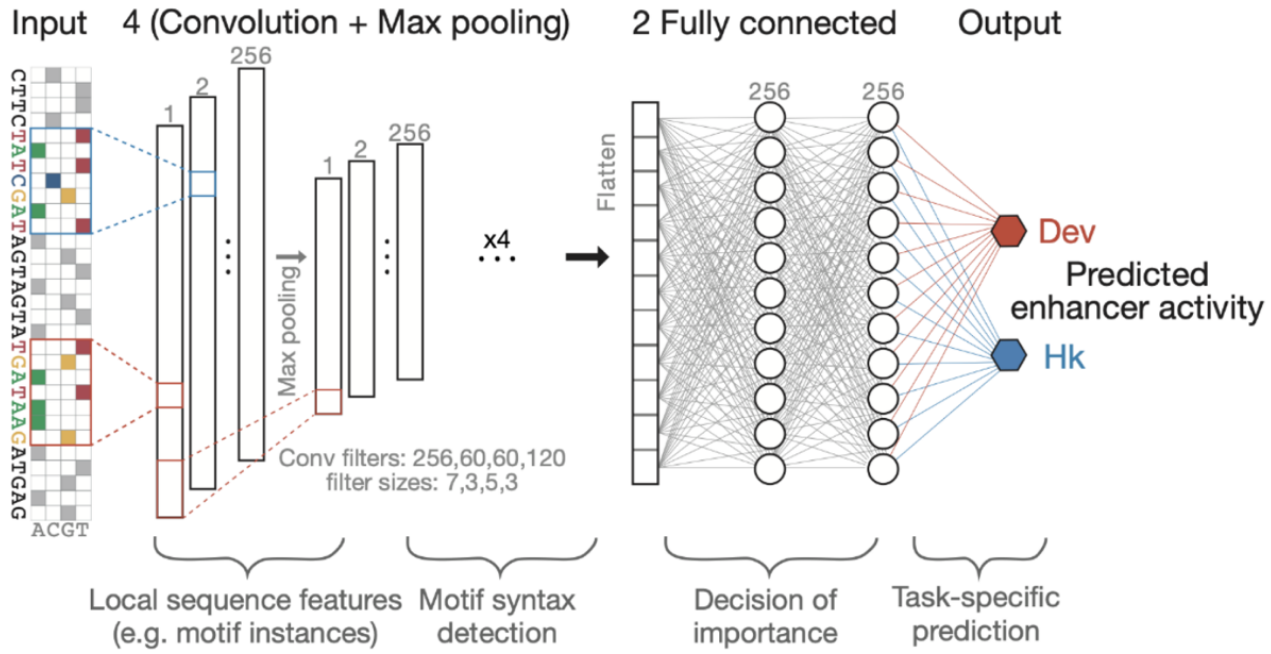


FIGURE 1.44 – Schéma du fonctionnement de DeepSTARR.

de transcription montrant que l'activité d'un même facteur de transcription pourra varier en fonction de ses séquences flanquantes. Ce n'est pas la première fois qu'un outil basé sur le Machine Learning pointe l'importance des séquences flanquantes des séquences régulatrices. En effet, il a été mis en évidence que les séquences flanquantes des STR (Short Tandem Repeats) allaient être décisives dans la distinction des différentes classes de STR mais également et surtout pour prédire l'initiation de leur transcription [Grapotte et al., 2021], ceci a été permis grâce à l'outil **DeepSTR**.

1.4.3 Le criblage CRISPR

1.4.3.1 La technologie du CRISPR

La séquence du CRISPR (Clustered Regularly Interspaced Short Palindromic Repeat) a été détecté pour la première fois chez *Escherichia coli* en 1987 [Ishino et al., 1987] et plus tard en 1993 chez l'archaea *Haloferax mediterranei* [Mojica et al., 1993]. Ce complexe consiste en une série d'unité palindromique répétée séparée par des séquences dérivées de virus [Mojica et al., 2005]. L'identification des régions CRISPR a conduit à la découverte de quatre gènes conservés présents à côté des régions CRISPR (CRISPR-associated system, Cas) [Jansen et al., 2002].

Le système CRISPR a ensuite été classé en six types. Les trois premiers sont les plus utilisés : Cas3 pour le type I, Cas9 pour le type II et Cas10 pour le type III. À ce jour, la plus utilisée et la plus étudiée est Cas9, une protéine effectrice multi-domaine qui peut se lier et cliver l'ADN cible. Le système Cas9 forme une structure à bases appariées entre un ARNc trans-activant (tracrARN) et l'ARNc de ciblage pour cliver l'ADN double brin (ADNdb) cible. Le clivage spécifique au site se produit à des emplacements déterminés à la fois par la complémentarité d'appariement des bases entre l'ARNc et l'ADNdb cible et un court motif, appelé motif adjacent protospacer (PAM), juxtaposé à la région complémentaire de l'ADN cible [Jinek et al., 2012]. La séquence PAM cible de Cas9 est 5'-NGG-3', tandis que les autres systèmes Cas vont reconnaitre des PAM différents et leurs mécanismes varient légèrement également.

De nos jours, un système CRISPR-Cas9 basique se compose d'un seul guide d'ARN (sgRNA) et de la nucléase SpCas9. Le sgRNA est un ARN non codant chimériques formé par le crRNA et le

tracrRNA. Dans ce complexe, le tracrRNA fonctionnera comme une séquence de recrutement de nucléase, tandis que le crRNA guidera le complexe vers la cible. L'une des applications classiques de la technologie CRISPR-Cas9 est l'inactivation des gènes. Le résultat du Cas9-sgRNA est une cassure double brin dans l'ADN cible généralement réparée par la voie de jonction d'extrémité non homologue (NHEJ), entraînant une "cicatrice" dans la séquence d'ADN (petites délétions, insertions ou indels) ; ou par voie de réparation dirigée par homologie (HDR), permettant des modifications précises des bases à l'aide d'une matrice d'ADN.

Des délétions génomiques peuvent être obtenues par double ciblage CRISPR-Cas9 en utilisant deux sgRNA. La NHEJ réparera les deux cassures double brin par suppression du segment intermédiaire. Cette technique a été utilisée pour créer une suppression allant d'un kilobase [Horii et al., 2013] à plus de 30 mégabases [Essletzbichler et al., 2014]. L'outil CRISPR se révèle être une révolution dans le domaine de la biologie. Sa facilité d'utilisation et sa versatilité font de lui un réel couteau suisse (Figure 1.45).

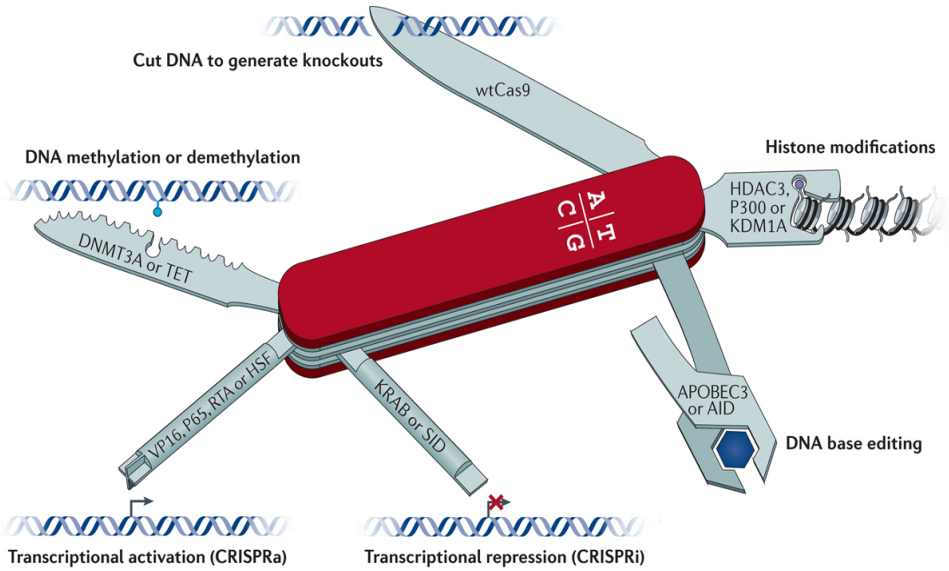


FIGURE 1.45 – La technologie du CRISPR offre une multitude d'application tel que d'appliquer une délétion, une baisse ou une augmentation d'activité sur une région [Doench, 2018].

1.4.3.2 Criblage CRISPR à haute densité

Au cours de ces dernières années, les criblages à grande échelle, c'est-à-dire le test simultané de milliers de sgRNA qui vont venir créer des perturbations individuelles dans une seule expérience, sont devenus un moyen populaire de réaliser des criblages génétiques dans les cellules de mammifères [Westbrook et al., 2005, Shalem et al., 2014, Wang et al., 2014]. Les criblages à grande échelle constituent une approche rentable pour étudier les phénotypes à l'échelle du génome et s'appuient sur des innovations technologiques avec les lentivirus, la synthèse d'oligonucléotides et le séquençage à haut débit. Le criblage à haute densité va débuter avec l'introduction de sgRNA qui constituent une librairie dans une population de cellule. Les sgRNA vont pouvoir s'intégrer dans le génome des cellules cibles grâce aux lentivirus perturbant le génome de la cellule hôte. Par la suite une pression de sélection est appliquée à la cellule comme une exposition à une drogue, une infection virale ou simplement à la prolifération cellulaire. Les sgARNs sont ensuite comptés dans le flot de cellules retenu après la sélection. Cela se fait

généralement par séquençage à haut débit. Dans les données obtenues, la déplétion par sgARN spécifiques identifie les gènes dont la perturbation sensibilise les cellules à la sélection, tandis que leur enrichissement identifie les gènes dont la perturbation confère un avantage sélectif (Figure : 1.46).

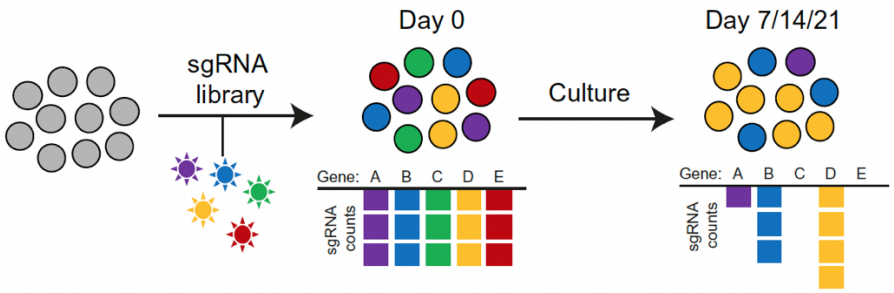


FIGURE 1.46 – Les cellules sont transfectés par des lentivirus contenant la librairie de sgRNA. En comptant le nombre de guide obtenues après une pression de sélection, on sera mesure d'identifier les gènes impliquer dans la réponse à cette pression [Lopes et al., 2021].

À ce jour, la plus populaire utilisation du CRISPR est le CRISPRko, où les gènes cibles sont dénaturés de façon critique par le guide d'ARN. Au-delà du CRISPRko, qui est basé sur les nucléases Cas, le système CRISPR-Cas peut également être utilisé pour d'autres fonctions. Une variante de la protéine Cas9, désactivité (dCas9) va pouvoir être utiliser afin de mener d'autres investigations biologiques. Le terme désactivé fait référence au fait qu'elle ne pourra pas couper l'ADN et ainsi ne pas mettre KO le gène cible. Parmi ces utilisations : le CRISPR interférence (CRISPRi) ou encore le CRISPR activation (CRISPRa) (Figure : 1.47). Dans le CRISPRi, la dCas9 est associé à un complexe répresseur qu'on a présenté préalablement le complexe KRAB. La répression par CRISPR peut donner différents phénotypes comparativement au CRISPRKo, car elle est moins susceptible d'activer des voies qui vont pouvoir pallier à l'inactivation du gène [Rossi et al., 2015]. Une limitation de CRISPRi par rapport au CRISPRko est la nécessité d'une expression continue de la protéine dCas9 et du guide ARN pour maintenir l'inactivation. Dans le CRISPRa, la dCas9 et de l'activateur transcriptionnel VP64 sont fusionné et vont prodiguer une modeste activation des gènes cibles [Gilbert et al., 2013]. L'activation par CRISPR a récemment été amélioré afin d'y incorporé le système dCas9-SunTag, qui augmente l'activation des gènes en recrutant de nombreuses copies de VP64 [Tanenbaum et al., 2014], et le système dCas9 – VPR, qui combine trois domaines activateurs (VP64 et domaines activateurs des facteurs de transcription p65 et Rta) en une seule protéine de fusion [Konermann et al., 2015]. Un autre système d'activation, le système SAM (Synergistic Activation Mediator) largement utilisé, recrute deux facteurs de transcription HSF1 et p65 pour augmenter l'activation du gène cible [Chavez et al., 2015].

Parmi les autres applications de la dCas9 on peut également compter :

Des modifications épigénétiques : En effet la technologie CRISPR va pouvoir être utiliser afin de perturber l'épigénome [Amabile et al., 2016], [Braun et al., 2017], [Hilton et al., 2015].

Remodelage de la chromatine : Le CRISPR a également été utilisé pour manipuler la structure tridimensionnelle de la chromatine, notamment en induisant une boucle de chromatine entre deux loci génomiques [Kim et al., 2019], [Morgan et al., 2017].

L'édition de bases : Les éditions de bases CRISPR sont des fusions de dCas9 avec des domaines protéiques qui modifient chimiquement les bases de l'ADN ; cela permet l'introduction de modifications génétiques sans induire de cassures double brin. L'édition de base a été utilisée pour modifier les variants génétiques associées à certaines pathologies, où elle permet un meilleur contrôle des changements induits par rapport à CRISPRko [Rees and Liu, 2018].

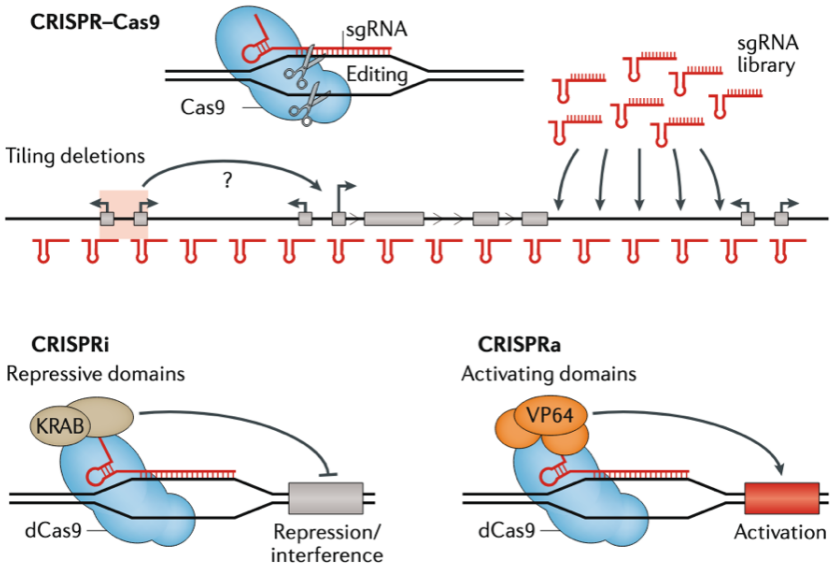


FIGURE 1.47 – Principe des différentes stratégies CRISPR [Andersson and Sandelin, 2020].

Les techniques de CRISPR ont souvent été utilisé afin d'étudier les gènes impliqués dans certains processus. D'ailleurs, bien souvent, les guides ciblent les promoteurs des gènes. Des études récentes ont fait part d'une utilisation du CRISPR sur des régions distales.

Par exemple Fulco et ses collaborateurs ont utilisé le CRISPRi afin d'identifier des connexions enhancer-promoteur et de mettre en lumière des régions affectant la régulation des gènes dans le loci GATA1 et MYC [Fulco et al., 2016]. Une librairie de plus de 98.000 guides a été généré afin d'étudie des séquences d'une longueur de plus d'une megabase chacune à proximité de ces deux facteurs de transcription majeurs. Cette approche non biaisé car sans à priori à permis d'identifié au moins 9 enhancers contrôlant l'expression des gènes et la prolifération des cellules K562.

Une autre étude plus récente a utilisé le CRISPRi afin de cibler les super-enhancers à l'échelle du génome afin d'identifier les super-enhancers essentiels à la prolifération des cellules leucémiques. Pour cela ils ont utilisé des cellules exprimant ETO2-GLIS2 dérivée d'un patient AMKL (Acute MegaKaryoblastic Leukemia) qui est une forme de leucémie aiguë myéloïde (AML) souvent caractérisé par des fusions de gènes, comme le CBFA2T3-GLIS2 récemment identifié (également appelé ETO2-GLIS2) chez 20 à 30% des patients [Thiollier et al., 2012]. En déployant un criblage CRISPRi à grande échelle dans cette lignée cellulaire en ciblant les super-enhancers, ils ont été capable de mettre en évidence des régions super-enhancers fonctionnellement liées au maintien de la leucémie [Benbarche et al., 2022] (Figure 1.48).

Ainsi on comprend mieux pourquoi on peut qualifier la technique du CRISPR de couteau suisse. Cette technique est terriblement efficace, et n'a pour limite que l'ingéniosité des équipes de recherche. Avec le bon modèle cellulaire, les bonnes cibles, les bonnes méthodes, les bons protocoles, on est capable de mettre en évidence des mécanismes de régulations impliqués dans la prolifération cellulaire, la survie de la cellule, des gènes sensibles à des drogues, etc.

Afin de tirer profit de ces expériences, des méthodes bio-informatiques ont vu le jour afin de traiter et interpréter les données de crible CRISPR. Ces méthodes inclus généralement le contrôle de la qualité des séquences, l'identification des gènes impliqués dans un phénotype sous pression sélective, les voies métaboliques des gènes essentielles identifiés, etc. Ainsi de nombreux outils ont vu le jour, des revues ont suivi permettant la comparaison des outils les plus populaires [Bodapati et al., 2020]. De par la popularité de cette technologie, l'accumulation de données CRISPR dans des bases de données publiques offre des opportunités intéressantes pour la méta-analyse des relations génotype-phénotype. Des bases de données CRISPR ont été

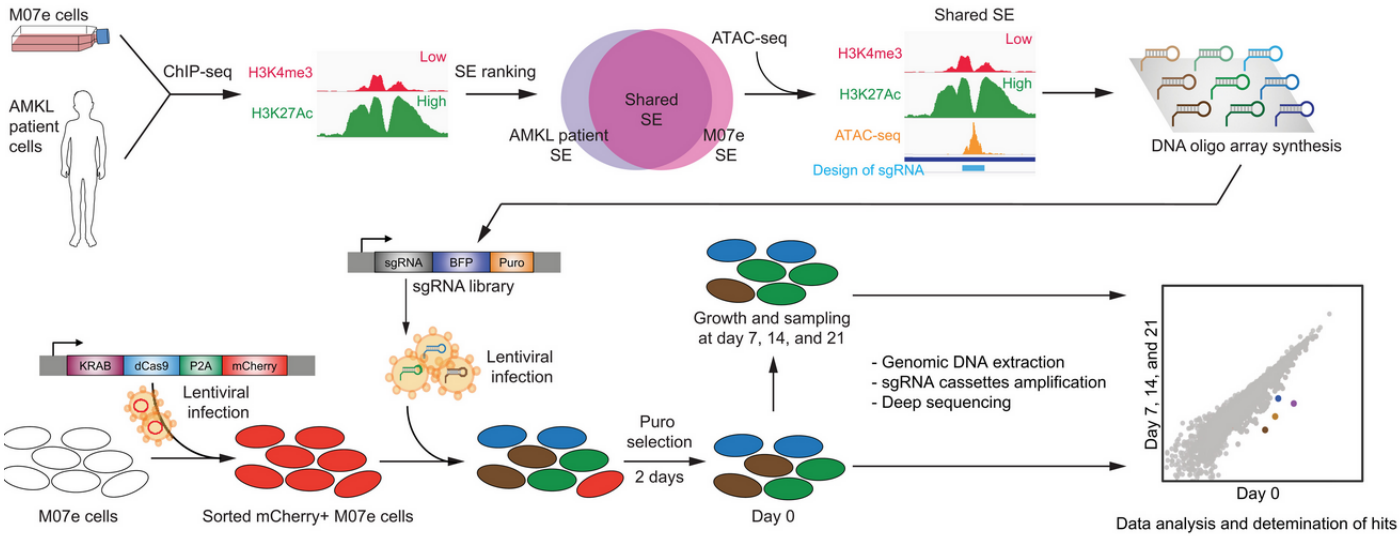


FIGURE 1.48 – Exemple d'une stratégie CRISPRi pour l'identification de super-enhancers impliqués dans un type de leucémie ai myéloide. La selection de la lignée cellulaire et des régions ciblées par les guides sont des étapes cruciales pour l'efficacité de l'étude. [Benbarche et al., 2022].

développées afin d'identifier, récupérer et analyser ces données. DepMap fait figure de référence fournissant des données de CRISPR et les profils génomiques associés pour des centaines de lignées cellulaires cancéreuses ; et la base de données BioGRID ORCS qui recueille et conserve les données CRISPR de la littérature scientifique [Oughtred et al., 2019].

1.5 Reproductibilité, inter-opérabilité, accessibilité des données

Lors des chapitres précédents, nous avons pu exposer différentes techniques expérimentales, différents moyens d'analyses de ces données. À l'heure actuelle il n'y a jamais eu autant de données disponibles dans la communauté scientifique, ce qui se traduit par une explosion du nombre de base de données. L'émergence du calcul haute performance ou HPC avec l'ouverture de grands centres de calculs, l'augmentation de la puissance de calcul, la parallélisassions sur GPU, etc. a, elle aussi, contribué à l'augmentation de cette production de données. Face à l'ère du Big Data, l'initiative FAIR (Findable, Accessible, Interoperable, Reusable) vise à normaliser l'ensemble des données et métadonnées afin de faciliter le partage de données et leur utilisation. Par données on entend un ensemble de faits, de mots, d'observations, de mesures ou la description d'un objet/événement/etc. Pour la biologie, ces informations peuvent être un nom de gène, sa position dans le génome, la protéine qu'il code, etc. Les données peuvent être exploitées pour en retirer des informations complètes. Elles peuvent être renvoyées ou analysées pour prendre certaines décisions. Les métadonnées sont décrites comme des données sur les données. Cela signifie que les métadonnées contiennent la description informative et pertinente des données originales. Elles aident l'utilisateur à connaître la nature des données et à déterminer leurs utilités.

1.5.1 FAIR Consortium

C'est en 2016 qu'est publié le premier article sur les principes FAIR (The FAIR Guiding Principles for scientific data management and stewardship) [Wilkinson et al., 2016]. L'objectif est de créer un "guide" de bonnes conduites à suivre afin de faciliter l'accès aux données, faciliter leur utilisation et uniformiser le partage de ces données. Ces principes reposent sur 4 critères :

F indable : l'information doit facilement être trouvable,

A ccessibility : accessible,

I nteroperable : doit être compatible avec plusieurs systèmes,

R eusable : les données doivent être reproductible.

1.5.1.1 Identifier & Accéder aux données

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

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

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

1.5.1.2 Interoperable

De nos jours, le partage de données est au cœur de la recherche. Des données générées dans le laboratoire vont souvent être couplées à d'autres jeux de données extérieurs. Elles seront comparées et/ou intégrées afin d'ajouter de l'information dans l'analyse. Cela implique que les données doivent être interopérables (utilisables) avec toutes sortes d'applications ou pipeline pour l'analyse, le stockage et le traitement.

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

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

1.5.1.3 Reusable

L'objectif de l'initiative FAIR est d'optimiser la reproductibilité des données. En effet, un grand nombre de publication ne permettent pas de reproduire à l'identique les résultats obtenus. En pratique, une telle reproductibilité des méthodes est beaucoup plus facile à mettre en œuvre avec des procédures strictement déterministes. C'est le cas en informatique, mais les expériences biologiques sont beaucoup plus sujettes à l'incertitudes en raison de la nature du matériel biologique qui peut-être hautement variable et de l'impossibilité de contrôler tous les paramètres. Alors que la science expérimentale doit s'efforcer de réduire ces incertitudes afin que les résultats puissent être reproduit, la bio-informatique n'a pas cette excuse. En mettant à disposition des données d'entrée identique, n'importe qui doit pouvoir atteindre exactement les mêmes résultats que ceux présentés. Pour ce faire, les métadonnées et les données doivent être décrites avec précision afin d'être reproduites et/ou combinées dans différents contextes.

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

Des méthodes permettent de faciliter la mise en place et le respect d'un tel protocole qui a pour but une science plus claire, une science ouverte, une science fiable.

1.5.2 Outils informatiques et reproductibilité

L'explosion des données biologiques a aussi conduit à un challenge dans l'analyse et le traitement des données. Ce nombre de données toujours croissant suscite des questions quant à leur traitement. L'analyse et l'intégration de ces données repose, de plus en plus, sur l'automatisation du traitement de l'information. On ne peut pas traiter l'analyse de plusieurs milliers de données de la même façon qu'une analyse isolée sur un ordinateur personnel. Il faut mettre en place des stratégies de traitement des données. Ces stratégies peuvent concerner la puissance de

calcul ou les outils bio-informatiques mis en œuvre. Il faut mettre en place des solutions qui soient reproductibles et transposables afin de s'assurer de la pérennité des données produites. Un pipeline est généralement constitué de composants linéaires, où l'un des outils logiciel en alimente un autre.

1.5.2.1 Modularité des pipelines

Au vu de la taille des données biologiques produites, il devient nécessaire d'utiliser des supercalculateurs capables d'une grande puissance de calcul, d'une capacité de stockage conséquente, ainsi que de la possibilité de paralléliser des tâches. Ces supercalculateurs ou HPC sont des clusters de "machines" ou nœuds reliés entre eux par un noeud maître capable de répartir les tâches et d'accéder aux données stockées. Ces noeuds sont eux-mêmes composés de cœurs (node). Les cœurs peuvent, eux-mêmes, être divisés en fils d'exécution ou "thread", mais cette division est purement virtuelle. Les cœurs d'un même nœud peuvent communiquer entre eux. Ce n'est pas le cas pour les noeuds d'un même cluster. Il est donc nécessaire d'avoir un noeud maître qui peut envoyer et récupérer des tâches aux noeuds du cluster. Les logiciels capables de distribuer les tâches à travers les noeuds sont appelés Portable Batch System (PBS) tel que TORQUE et SLURM (Simple Linux Utility for Resource Management).

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

Les gestionnaires de workflow se chargent de mettre en relation toutes les étapes du pipeline, de suivre leur exécution et de configurer la parallèlisation sur différentes architectures. Ils utilisent leur propre langage et syntaxe. En bioinformatique, les gestionnaires de workflow les plus courants sont Snakemake [Köster and Rahmann, 2012] et Nextflow [Di Tommaso et al., 2017]. Au cours de ma thèse, j'ai utilisé Snakemake car il est basé sur le langage Python qui est très intuitif et est largement utilisé au sein du laboratoire TAGC me permettant un soutien de mes confrères bioinformaticien. Snakemake permet la composition de workflows basés sur un graphique de règles dont l'exécution est déclenchée par la présence, l'absence ou la modification de fichiers et répertoires attendus. Il génère lui-même le graphique de dépendance des règles (Direct Acyclic Graph, DAG) et peut relancer uniquement les parties pertinentes du workflow. La construction du DAG se fait à partir de la sortie finale et remonte jusqu'aux fichiers d'entrées. Il est possible d'écrire des règles utilisant du Bash, du Python et du R nativement. Snakemake permet de configurer, pour chaque règle, les ressources demandées, ainsi que de définir l'environnement à utiliser (voir partie suivante). Nextflow est une alternative à Snakemake. À la différence de Snakemake le DAG est construit à partir des fichiers d'entrées et "descend" jusqu'à la sortie. En conséquence, les différentes étapes peuvent être générées dynamiquement au cours du workflow et permettre les embranchements dans le pipeline. Le gestionnaire de workflow Nextflow supporte l'utilisation du langage Groovy basé sur Java. Grâce à l'utilisation des HPC et des gestionnaires de workflow tels que Snakemake il est donc possible de moduler relativement facilement les pipelines d'analyse et d'intégration de données.

1.5.2.2 Reproductibilité des analyses

Comme nous l'avons expliqué plus haut, la reproductibilité des données est un enjeu majeur de la biologie. Concernant la bioinformatique, une grande question se pose : Comment reproduire et vérifier les résultats obtenus par des pipelines ? En science, les publications dans les revues scientifiques sont basées sur "l'évaluation par les pairs" (peer review). Une expérience impossible à reproduire a peu de valeur dans le milieu scientifique. En bioinformatique, il n'est pas rare d'observer une section « matérielle et méthode » peu dense et informative ne permettant pas la reproductibilité des données (outils ou script bioinformatique utilisés, paramètres des outils, version, etc.). Une meilleure reproductibilité des workflows passe, tout d'abord, par une meilleure traçabilité des outils et scripts utilisés. En plus d'une documentation du code et du projet il faut également fournir les informations minimales permettant la reproductibilité des expériences :

— Système (OS) et version où les scripts ont tourné (Ubuntu, MAC, etc.)

— Schéma du pipeline avec étapes + entrée et sortie

— Scripts utilisés et commentés

— Outils bioinformatiques utilisés : numéro de version, paramètres utilisés

— Description des fichiers d'entrée et de sortie (origine, format, information contenue, optionnalité des fichiers, etc.)

Pour toutes ces raisons, il est conseillé d'utiliser des contrôleurs de version décentralisés, tels que GIT afin de s'assurer de la traçabilité de son pipeline. GIT permet de créer des "points de sauvegarde" des dossiers et fichiers contenant le pipeline sur un serveur distant au cours du développement (Figure :1.49). Cela permet de revenir facilement à une version précédente, mais aussi de créer des branches de développement sans toucher à la version principale du workflow (Figure : 1.50). GIT permet donc de partager et de coopérer plus facilement sur la création et le développement de pipeline et permet de s'assurer d'une meilleure traçabilité et reproductibilité du workflow.

Une meilleure reproductibilité des pipelines passe aussi par la portabilité des pipelines. L'accès à un pipeline et à ses scripts n'assure pas forcément la reproductibilité des données. En effet, un pipeline développé pour une machine ou un serveur particulier n'est pas forcément adaptable pour d'autres architectures (PBS, OS différent, dépendance logicielle manquante, etc.).

Des outils informatiques ont donc été développés afin d'isoler l'exécution de chaque outil dans un environnement contenant toutes les dépendances nécessaires à son exécution. Afin de pallier à ce problème dans mon cas, je me suis énormément servi de Conda. Conda est écrit en python. C'est un gestionnaire de package permettant de créer facilement des environnements isolés. De très nombreux outils d'analyses de données sont installable via Conda et plus particulièrement Bioconda [Grüning et al., 2018]. Lorsqu'un outil est installé, Conda va gérer automatiquement tous les packages nécessaire, toutes les dépendances, automatiquement, le tout dans un environnement isolé, n'impactant pas le système de l'utilisateur. Un environnement Conda peut facilement être créé à partir d'un fichier de configuration listant les outils à installer. Il est intéressant de noter que l'on peut coupler Conda à Snakemake. Ainsi lors de la création d'un pipeline, on peut renseigner les différents outils ainsi que leurs versions, renseigner ce fichier de configuration à Snakemake qui se chargera d'installer le tout proprement dans un environnement unique à l'analyse. De plus, comme indiqué plus haut, Snakemake va garder une trace des analyses déjà effectué, ainsi si le pipeline est relancé, les programmes ne seront pas installés à nouveau. Un des avantages de Conda est de ne pas nécessiter de droit spécifique pour son installation. En effet, sur les serveurs de calcul, il est courant de ne pas détenir les droits administrateurs indispensables à l'installation des logiciels. Conda permet de contourner cette limite car il installe les outils dans un dossier local.

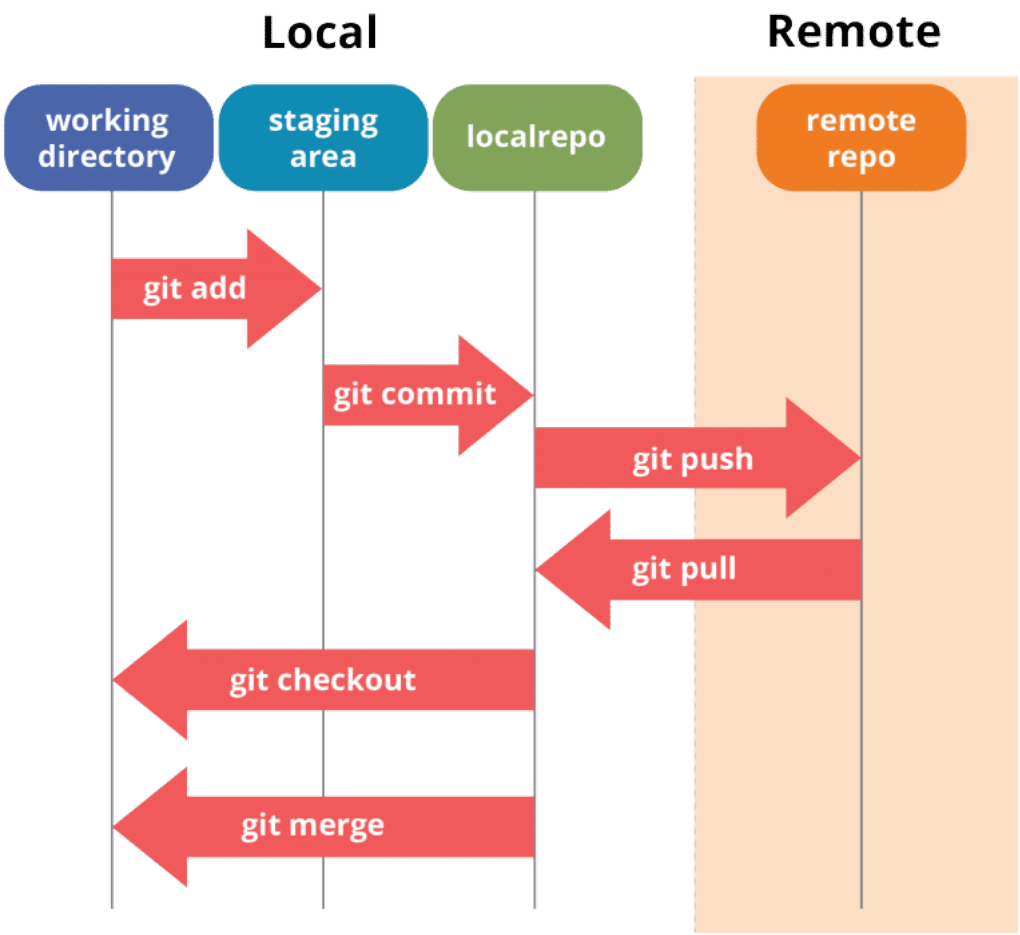


FIGURE 1.49 – Git permet de sauvegarder le travail en local, le commenter, puis l'envoyer sur un serveur distant afin de sécuriser la sauvegarde mais également la rendre accessible aux différents collaborateurs.

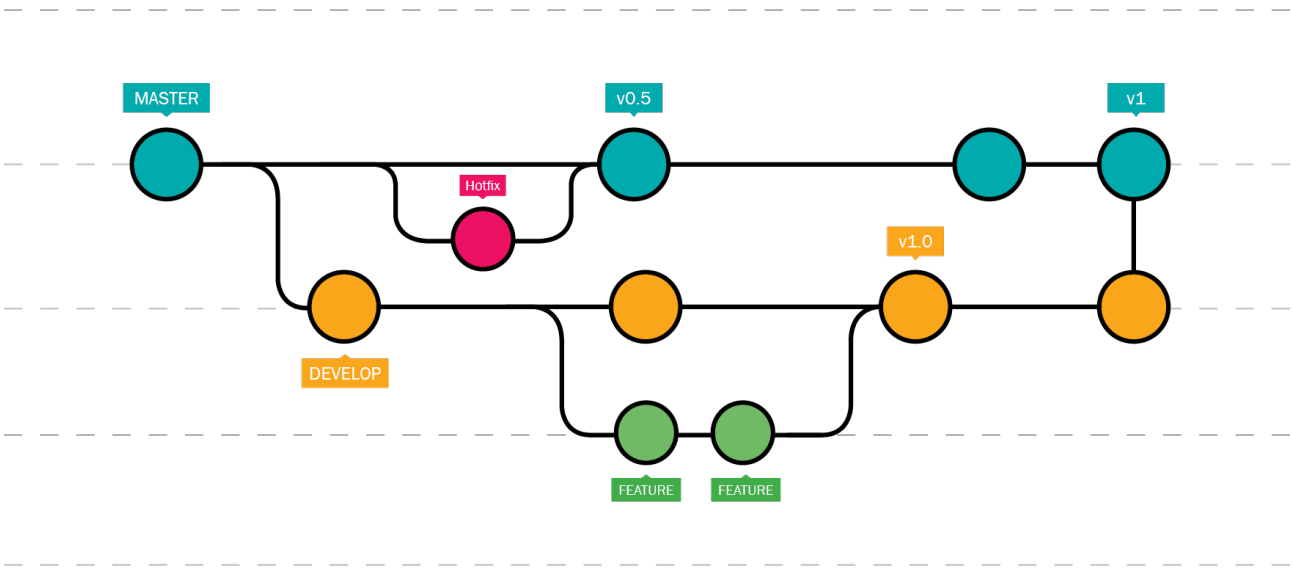


FIGURE 1.50 – Git permet de créer plusieurs sauvegarde son projet avec de revenir à une version précédente. Il permet également la création de branche en parallèle pour un développement stable du pipeline

Docker et Singularity sont des alternatives à Conda. Personnellement je n'ai pas utilisé ces outils car je trouvais la collaboration entre Snakemake et Conda optimale, mais ils sont largement utilisés ainsi je les présenterais brièvement. Ils se basent sur la création de conteneurs reproduisant n'importe quel OS dans lesquels sont installés les outils. Il est facile d'importer une image Docker dans Singularity. Docker et Singularity ont l'avantage de permettre d'isoler parfaitement les processus exécutés au sein des conteneurs au contraire de Conda. Disposant d'une grande communauté, il est possible d'échanger et partager ses conteneurs via un Hub de partage. Par contre Docker/Singluarity exige un accès Admin. Ces droits administrateurs sont indispensables pour composer des conteneurs. Docker et Singularity sont également gérés par Snakemake.

Ainsi nous avons présenté différents méthodes et outils afin de générer des programmes réutilisables par le plus grand nombre afin de permettre et de faciliter l'accès à une science ouverte et reproductible. Il faut que le pipeline et les scripts soient documentés. Il est indispensable que le pipeline et ses dépendances soient facilement partageables, réutilisables et portables. Ces contraintes passent par une combinaison de l'utilisation de logiciel de contrôle de version à distance (Git), de gestionnaire de workflow (Snakemake) et d'environnement (Conda). Initié le plus tôt possible les bioinformaticiens à ces outils est nécessaire à un maintien des bonnes pratiques et nécessaire pour une science ouverte [Wilkinson et al., 2016].

2 Résultats

2.1 Objectifs du projet de thèse :

Aux différents stades de vie et de développement d'un organisme, les cellules le composant vont devoir se différencier, s'organiser et s'adapter. En effet toutes les cellules d'un organisme eucaryote contiennent le même support génétique et en fonction du contexte cellulaire, ce n'est qu'une fraction des éléments qui le compose qui vont devoir s'exprimer. C'est notamment grâce à la régulation génique par les éléments cis-régulateurs que la cellule va être dirigée dans une voie de différenciation plutôt qu'une autre. Cette régulation précise va impliquer des activations et des répressions finement orchestrées par les différents acteurs de la régulation génique que nous avons exposés au cours de l'introduction. Les éléments activateurs tels que les promoteurs et enhancer ont été largement étudiés, permettant une caractérisation de ces éléments. Cependant, malgré l'importance des silencers dans le développement et les réponses au signal, ils ont généralement été beaucoup moins étudiés que les enhancers, avec plus de 1,5 million d'articles PubMed pour les "enhancers" contre environ 100 000 pour les "silencers".

Quelques études sporadiques ont mis en évidence des silencers. Ces éléments ont été caractérisés de façon individuelle et non pas par des approches à grande échelle. En effet au commencement de ma thèse, il n'existait pas d'approches à haut débit afin d'étudier ces éléments répresseurs. L'activité des silencers peut être évaluée par des tests de gènes rapporteur où un fragment d'ADN est testé pour évaluer sa capacité à influencer l'activité d'un élément promoteur associé à un gène rapporteur, comme c'est le cas pour d'autres éléments cis-régulateurs. Alors que plusieurs de ces tests ont été développés pour évaluer systématiquement l'activité enhancer, aucune stratégie à grande échelle n'était décrite pour les silencers. Au cours de ma thèse, quelques études sont apparues, proposant une identification et une caractérisation des éléments silencers à grande échelle en utilisant des techniques de gènes rapporteurs.

Dans notre laboratoire, notre équipe s'intéresse aux mécanismes de régulation des lymphocytes T. Notre groupe a pris part dans la découverte et la caractérisation d'éléments cités plus hauts comme les promoteurs et les enhancers, mais a également mis en lumière un nouveau genre d'enhancer : les Epromoteurs. Au sein de notre équipe une variante du STARR-seq, le CapSTARR-seq a été très largement utilisée et a permis d'identifier avec succès les enhancers à l'échelle génomique. C'est exactement sur cette technique que nous nous sommes basés pour l'identification des silencers. Nous avons réadapté la technique CapSTARR-seq afin d'identifier ces éléments à l'échelle du génome. Ainsi les objectifs de ce projet étaient :

1. La conception et le test de nouveaux vecteurs de CapSTARR-seq. En effet afin d'optimiser l'identification des silencers, différents vecteurs de STARR-seq ont été testés en changeant systématiquement le promoteur, ceci afin de comparer les résultats obtenus et éventuellement déterminer la meilleure approche permettant l'identification des silencers ;
2. Mettre en place une méthode de traitement automatisée des données de CapSTARR-seq permettant notamment l'identification des silencers ;
3. L'évaluation de la cohérence et de la reproductibilité des différentes stratégies ;
4. Une analyse bioinformatique approfondie du contexte (épi)génomique des silencers ;
5. La validation des régions silencers potentielles par une approche de gène rapporteur classique et également une validation fonctionnelle par une approche génétique utilisant la

technologie CRISPR-Cas9.

C'est dans ce contexte d'identification et de caractérisation des silencers que mon projet de thèse s'est inséré. En effet le CapSTARR-seq est une stratégie relativement nouvelle et peu répandue, ainsi il fallait dans un premier temps créer une méthode de traitement informatique automatisée en tenant compte des spécificités de cette technologie.

1. Mon premier objectif était de créer un pipeline d'analyse de données CapSTARR-seq. Il a tout d'abord été construit dans le but de permettre l'identification des silencers car il s'agit de ma thématique de recherche. Par la suite, je l'ai adapté afin de permettre l'identification de séquence possédant une activité répressive mais également activatrice. Il s'agissait également d'identifier si le signal répresseur était continu au sein de la région ou si l'activité inhibitrice allait être localisée au sein d'une sous-unité.
2. Dans un second temps, il s'agissait d'effectuer une caractérisation génomique et épigénétique la plus complète possible des éléments silencers identifiés. Comme par exemple caractériser leur distribution génomique, mettre en évidence les facteurs de transcription qui allaient se fixer au sein des régions identifiées, mener des analyses multi-omiques en intégrant des données de RNA-seq par exemple afin d'évaluer l'expression de gènes à proximité des silencers identifiés, caractériser les différentes marques d'histones spécifiques aux silencers. En intégrant toutes ces données et analyses, il s'agissait de proposer des mécanismes de régulation des régions silencers.
3. Dans un dernier temps, il s'agissait de mettre en perspectives les différents résultats obtenus en fonction des différentes approches expérimentales et ainsi comparer les différentes stratégies mise en place pour l'identification des silencers.

Mon projet de thèse s'insère dans une démarche de compréhension globale des silencers pour lesquels de nombreuses questions restent sans réponses, tel que proposé dans une revue récente sur les silencers [Segert et al., 2021] :

— Existe-t-il une signature chromatinienne spécifique aux silencers ? Les études actuelles n'ont pas identifié de signature chromatinienne caractéristique des silencers.

— Quels cofacteurs sont nécessaires pour la fonction silencer ? Quelles protéines non caractérisées sont impliquées dans l'activité silencer ?

— Comment se comportent les silencers en 3D ? Comment les boucles silencers-promoteur interagissent-elles avec la structure des domaines topologiquement associés (TAD) qui structurent le génome à plus grande échelle, et sont-elles similaires aux boucles enhancer-promoteur ?

— Comment les séquences fonctionnelles sont-elles structurées dans les silencers, existe-t-il des sous domaines responsables de l'activité inhibitrice ? Peut-on observer une différence de taille significative entre enhancers et silencers ? Les silencers présentent-ils des nombres, une complexité et des principes d'organisation des sites de liaison aux facteurs de transcription similaires à ceux des enhancers ?

— Comment les silencers sont-ils conservés au sein de l'évolution ? Dans quelle mesure le changement évolutif des silencers entraîne-t-il l'évolution des réseaux de régulation et des phénotypes ? Existe-t-il une signature de conservation qui distingue les silencers des autres éléments génomiques ?

— Comment les variations nucléotidiques (SNP) affectent-elles l'activité des silencers et sont-elles sources de pathologies ? Comment prédire l'impact des variants non codants dans les éléments bi-fonctionnels sur l'expression des gènes et les phénotypes ?

2.2 Résumé des résultats :

Afin de caractériser les éléments silencers, il me fallait dans un premier temps les identifier. Dans ce projet, nous nous sommes basés sur la technique du CapSTARR-seq afin de tester systématiquement les régions sensibles à la DNase I (DNase I Hypensitive Site – DHS) dans la lignée cellulaire P5424 qui est une lignée cellulaire des lymphocytes T chez la souris.

Ainsi j'ai créé un pipeline que l'on peut diviser en deux parties : une première utilisant des outils tel que bedtools, ainsi que des commandes en bash et awk afin d'optimiser le traitement des données et une seconde en R. Ces outils, commandes et codes sont chaînés entre eux via Snakemake que j'ai pu présenter dans l'introduction. Snakemake me permet d'assurer la reproductibilité des résultats en utilisant une succession de tâches codées simplement et efficacement. Ce pipeline m'a permis de calculer l'activité des régions testées et de mettre en évidence des régions présentant une activité répressive. Afin d'être considéré comme silencers, les régions devaient présenter un signal cDNA au moins deux fois moins élevé que le signal de la librairie, soit un fold change de 0.5 qui correspond dans l'échelle logarithmique de base 2 à -1 ($Log_2 \leq -1$) (Section : 2.3). Ce pipeline a été utilisé dans plusieurs projets de l'équipe menant à des publications dans lesquels je suis co-auteur (Section : 2.5.1, Section : 2.5.2).

Par la suite, j'ai effectué la caractérisation la plus complète possible de ces régions et ceci systématiquement dans toutes les conditions. Par conditions j'entends les différentes constructions des vecteurs STARR-seq que sont SCP1, pPGK et pR-Eα. Ainsi j'ai effectué de nombreuses analyses bioinformatiques faisant intervenir des données générées dans notre laboratoire tel que des données de RNA-seq ainsi que des données issues de bases de données publiques telles que JASPAR. Les résultats sont mis en forme dans un article en cours de soumission dont j'ai mené la totalité des analyses bioinformatiques et créé les figures sous ggplot afin d'en présenter les résultats. Saadat HUSSAIN a quant à lui été à l'origine des manipulations expérimentales dont le fruit m'a servi de support tout au long de mon projet de thèse (Section : 2.4).

M'intéressant plus largement aux mécanismes de régulation par les éléments distaux et à leur identification par des techniques de crible fonctionnel à grande échelle, je me suis également intéressé à la technique de criblage par CRISPR/Cas9. Quelques librairies existent déjà permettant de générer des guides ARN qui vont cibler les gènes. Cependant, afin d'étudier les éléments distaux, ces librairies ne sont pas suffisantes. Ainsi, j'ai créé un pipeline permettant de générer des librairies personnalisées. Pour cela je me sers de l'outil CHOPCHOP [Montague et al., 2014, Labun et al., 2019] que j'ai parallélisé via Snakemake afin de pouvoir générer une liste de guides qui cibleront les régions renseignées en entrée. Pour le moment, ce pipeline a été utilisé afin de générer une librairie permettant de cibler des enhancers distaux impliqué dans les leucémies lymphoblastiques aigüe de type T. Par la suite, je me suis intéressé aux différents outils permettant l'analyse de données CRISPR. Parmi ces outils, j'ai retenu MAGeCK dont le modèle mathématique reposant sur une loi binomiale négative me semblait le plus pertinent (Section : 2.5.4). J'ai également contribué au développement d'OLOGRAM, un outil permettant de comparer la colocalisation significative d'éléments génomiques. Ceci m'a permis de participer à la phase de test d'un outil en développement au sein du toolkit PyGTFtk (Section : 2.5.3).

2.3 STARR Track un outil pour l'analyse de données CapSTARR-seq

Afin d'analyser les données de CapSTARR-seq, mon premier objectif était de créer un pipeline que j'ai nommé STARR Track. Le développement de ce pipeline était motivé par la nécessité d'avoir un outil permettant de tenir compte des spécificités du CapSTARR-seq, et de permettre une visualisation simple et efficace des données. En effet, la visualisation des données me semble toujours être un bon point de départ lors d'une analyse. C'est à mon sens un réflexe nécessaire avant de débuter une analyse qui permet de détecter d'éventuelles anomalies et d'orienter la suite du processus de traitement des données.

Concernant les spécificités du CapSTARR-seq, un des avantages est que chaque ensemble de reads présentant des coordonnées strictement identiques peuvent être considérés comme un "fragment capturé", un clone testé dans le vecteur CapSTARR-seq. Bien sûr, la taille du read ne reflète pas la taille initiale du fragment capturé. Initialement, j'utilisais MACS2 pour estimer la taille du fragment capturé pour les données séquencées en single-end. Désormais, la plupart des séquençages sont effectués en paired-end ce qui facilite la tâche de l'élongation. Ainsi, on peut directement estimer l'activité par fragment et ceci de façon individuelle. Pour citer un désavantage, le CapSTARR-seq n'est pas efficace à 100% et on ne capture pas uniquement les régions d'intérêt mais également des séquences dispersées et aléatoires à travers le génome, ce qui va ajouter du bruit aux données. Mais ce que j'estimais être un inconvénient s'est révélé être un avantage. En effet, j'ai pu me servir de ces régions afin de les comparer avec les régions capturées et m'en servir comme background pour la détection d'éventuels faux positifs. Cette partie étant toujours en développement, je ne la présenterai pas dans la suite de l'explication du pipeline, mais la commenterai dans la section discussion (Section : 3.2.1).

Concernant la structure du pipeline, il prend en entrée le fichier BAM de la librairie, les fichiers BAM des différents réplicats cDNA ainsi que la liste des coordonnées des régions génomiques capturées. Les réplicats sont fusionnés afin de créer un fichier "merged". Le fichier merged et les différents réplicats sont traités de façon indépendante. Par la suite, tous les reads sont concaténés en une liste unique de reads qui servira de référence afin de comptabiliser les occurrences de chaque reads partageant les coordonnées strictement identiques. Les occurrences sont comptabilisées via BedTools intersect en utilisant l'option -c afin de rapporter le nombre de reads qui chevauchent avec le même read dans le fichier de références. Cette étape me permet d'avoir un fichier BED, où chaque ligne représente un fragment capturé avec le nombre de reads dans la condition cDNA ainsi que dans la librairie (Figure 2.2).

L'activité par fragment est calculée en divisant le signal cDNA par le signal dans la librairie, en prenant soin de normaliser par le nombre de séquences. Un code couleur allant de vert (signal positif = enhancer) à rouge (signal négatif = silencer) est calculé proportionnellement à l'activité du fragment. Ce code RGB est par la suite incorporé à la 9ème colonne du fichier BED permettant la visualisation du signal dans un explorateur tel que IGV ou UCSC Genome Browser Track. Cette étape de visualisation s'est avérée extrêmement importante dans le développement de STARR Track. Elle m'a permis de remarquer que de nombreux fragments étaient capturés aux abords des régions d'intérêts (DHS) et qu'ils possédaient une forte activité. Une hypothèse serait que ces séquences soient répétées, induisant un décalage lors de l'hybridation de la sonde, ce qui expliquerait la capture de ces séquences flanquantes.

Afin de tenir compte de ce facteur, les régions d'intérêt ont été étendues (Figure 2.3). Les fragments chevauchants ont été annotés par l'identifiant unique que possède chaque région. Pour identifier les sous-régions possédant l'activité la plus prononcée, ce que nous avons appelé les régions "cores", la région est divisée en un nombre de paires de bases choisi par l'utilisateur en fonction de la résolution souhaitée. Ainsi, l'activité de la région et de la sous-région correspond à la moyenne de l'activité des clones qui appartient à cette région / sous-région. Ceci nous permet

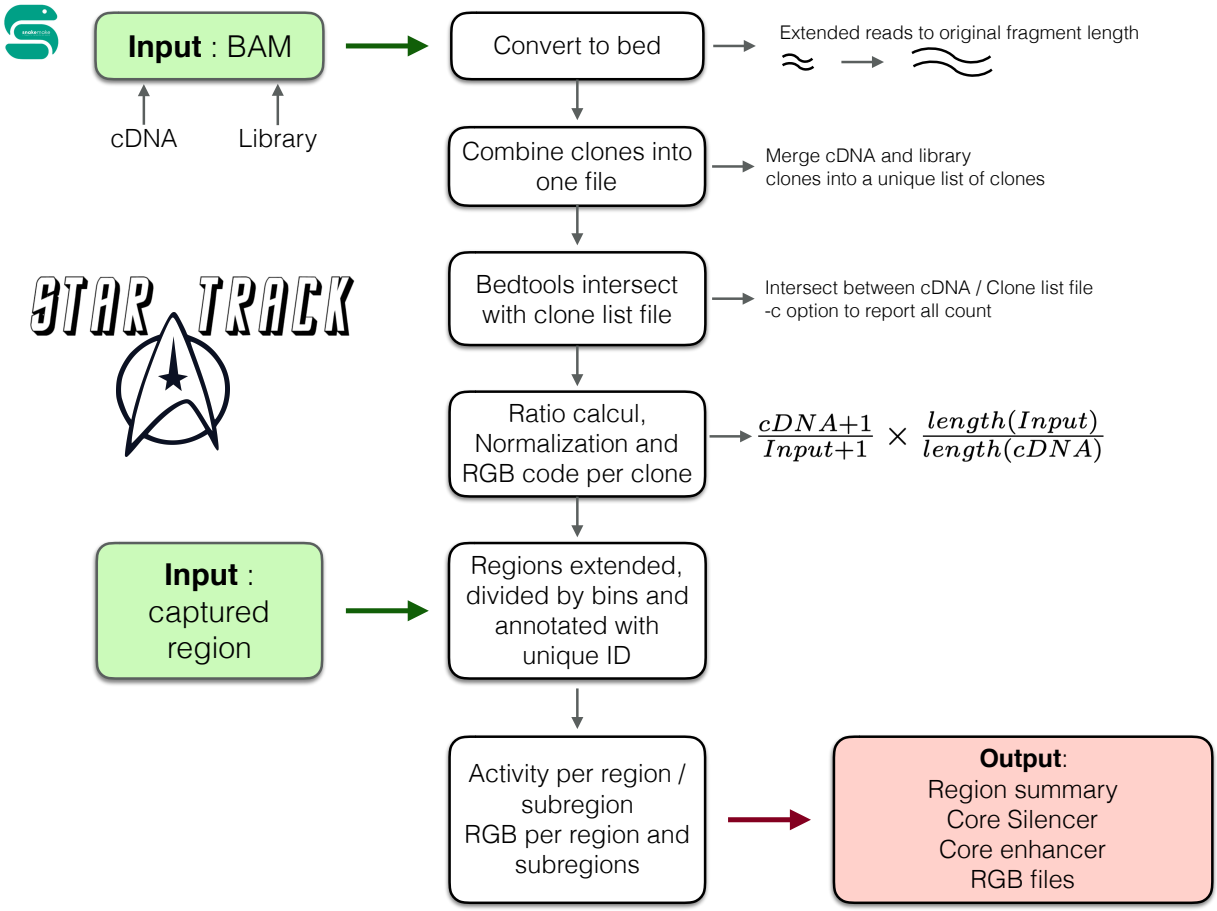


FIGURE 2.1 – Les étapes du pipeline STARR Track sont chainées via Snakemake qui assure la reproductibilité du pipeline et gère la parallèlisation du traitement des réplicats séparément ainsi que du fichier merged.

chr	start	end	.	sc	strand	cDNA	Lib
chr1	3026317	3026632	.	24	+	19	24
chr1	3026118	3026433	.	38	-	19	15
chr1	3026178	3026493	.	42	-	72	32

FIGURE 2.2 – Coordonnée d'un clone avec le signal du cDNA (colonne 7) et le signal de la librairie (colonne 8). L'activité du clone correspond à la formule cDNA / input.

d'identifier les silencers, que nous considérons comme les régions possédant un $Log_2(FC) \leq -1$ et les enhancers qui ont une activité supérieure au point d'inflexion, ce dernier étant calculé dynamiquement en fonction de l'activité des régions. Les core-silencers correspondent aux sous-régions successives possédant un $Log_2(FC) \leq -1$. Un silencer peut contenir plusieurs core-silencers, si des sous unités possédant un $Log_2(FC) > -1$ s'intercalent entre des sous régions possédant un $Log_2(FC) \leq -1$

Les résultats provenant de ce pipeline m'ont permis d'identifier une liste de silencers basée sur les différentes stratégies CapSTARR-seq testé. De plus, j'ai également mis en évidence des sous-régions pilotes, me permettant en fonction de l'analyse effectuée, d'être plus résolutif ce qui se traduira par une caractérisation plus fine des silencers.

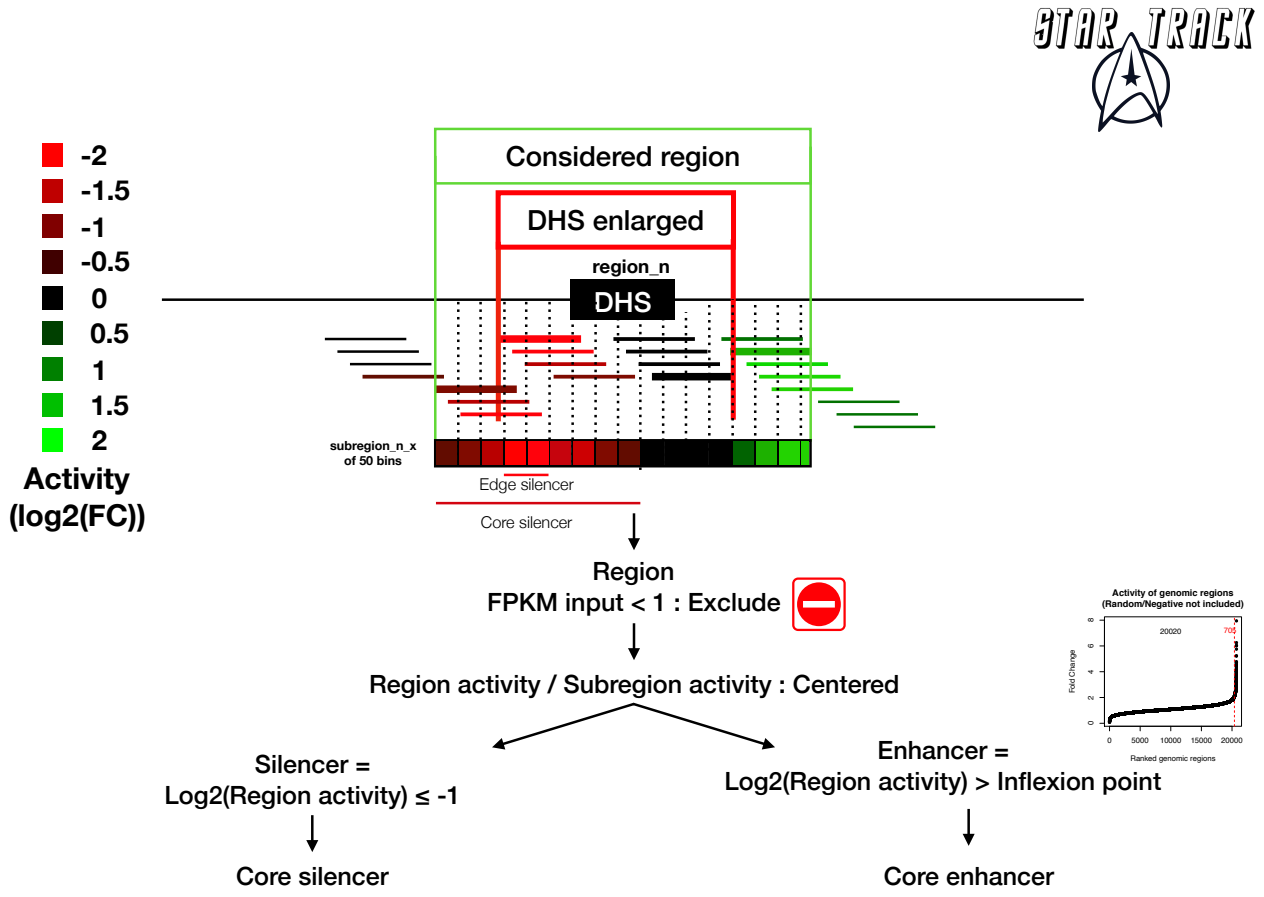


FIGURE 2.3 – Les régions sont étendues en prenant le 5' du premier clone chevauchant une région et le 3' du dernier clone de la même région. L'opération est renouvelée afin d'obtenir les coordonnées de la région étendue. Les régions possédant un FPKM inférieur à 1 dans la librairie ont été exclues afin de ne pas tenir compte des régions possédant une couverture trop faible, L'activité des régions et sous régions est calculée permettant l'identification de silencers et d'enhancer ainsi que de leur sous unité pilote "core".

2.4 Identification des éléments Silencers par une approche de test de gène rapporteur à grande échelle

La compréhension du fonctionnement des mécanismes de régulation est un élément clé dans le déchiffrage de l'organisation de nos cellules et dans la compréhension de plusieurs processus pathologiques. Alors que les enhancers et les promoteurs ont été largement étudiés et sont bien référencés, l'identification et la compréhension des silencers émergent à peine [Della Rosa and Spivakov, 2020].

Mon projet de thèse était axé sur la découverte et la caractérisation de ces éléments répresseurs, qui a donné lieu à une publication soumis le 9 mai 2022 à la revue Nucleic Acids Research. Cet article propose un catalogue des silencers chez la souris dans lequel nous proposons des mécanismes d'actions [Hussain et al.,]. Saadat Hussain est à l'origine des manipulations expérimentales, et je suis à l'origine des analyses bioinformatiques présentées dans cet article. Ainsi Saadat a construit plusieurs vecteurs de CapSTARR-seq basé sur 4 promoteurs différents : SCP1, pPGK, pEF1α et pR-Eα qui est un promoteur couplé à un enhancer tissu-spécifique et en testant systématiquement les régions sensibles à la DNase 1 (DHS) dans la lignée P5424. La lignée cellulaire P5424 est issue de lymphocytes T précoce en développement et ressemble aux thymocytes double positif aux niveaux phénotypique et transcriptomique. Les données issues du promoteur EF1α n'ont pas été retenues car elle n'était pas consistantes : elles n'apportaient aucune plus-value à l'analyse.

Par la suite, j'ai réalisé le traitement des données CapSTARR via STARR Track que j'ai développé pendant ma thèse, qui a conduit à une liste de silencers potentiels par stratégie, une faible proportion d'entre eux étant partagés entre les conditions. Une liste de silencers identifiés par CapSTARR-seq a directement été validée par un test de luciférase montrant que les candidats identifiés avaient une activité répressive. Toutes les analyses ont été menées systématiquement sur les trois stratégies. Par la suite, j'ai mené la caractérisation la plus complète possible des éléments identifiés, dans le temps qui m'était imparti et dont les résultats sont consultables dans la publication. J'ai effectué toute les analyses bioinformatiques présentées dans l'article et organisé les résultats sous formes de figures que j'ai générées. Nous nous sommes intéressés à la distribution génomique des silencers putatifs, avons comparé l'enrichissement en marques d'histones, et étudié l'enrichissement en termes GO des silencers. J'ai également mené des analyses multi-omiques en couplant les résultats obtenu via CapSTARR-seq avec des données de RNA-seq afin de comparer le niveau d'expression des gènes à proximité des silencers avec des régions ne possédant aucun signal et d'enhancer. On peut observer que les gènes à proximité des régions silencers sont moins exprimés que les gènes à proximité des régions contrôles et enhancers. J'ai également mis en évidence la fixation de facteurs de transcription possédant une activité répressive tel que REST et GFIb et comparé l'activité des régions possédant les sites de fixations des facteurs de transcription. En mutant le site de fixation de REST, une perte du signal silencer a été détecté.

En recherchant des silencers régulant des gènes impliqués dans la différenciation cellulaire des lymphocytes, nous avons identifié un candidat arborant des sites de fixation du facteur de transcription ZNF263 répété en tandem. ZNF263 est un motif présentant un domaine en doigt de zinc (zinc finger) associé avec un domaine KRAB et Scan (Section : 1.2.2.4). Grâce au pipeline développé et en visualisant les données, on a pu repérer qu'il y avait une forte colocalisation entre le signal core silencer et ces motifs de ZNF263 répété et étant classé par RepeatMasker [Chen, 2004] comme STR. Plus étonnant encore nous avons pu mettre en évidence une corrélation entre le nombre de répétition et l'activité silencer plus particulièrement chez les silencers identifiés via pPGK. Ces résultats obtenus *in silico* ont été confirmés expérimentalement par Saadat Hussain qui a vérifié par une expérience de mutagenèse qu'il existait bel et bien un rapport entre le nombre de répétition et l'activité silencer. Ainsi en servant de plateforme de

recrutement pour le facteur de transcription ZNF263, les STR pourraient jouer un rôle majeur dans l'activité inhibitrice des éléments silencers.

ZNF263 tandem repeats are important contributors to silencer elements in T cells

Authors: Saadat Hussain[1,2,#], Nori Sadouni[1,2,#], Dominic van Essen[3], Lan T.M. Dao[1,2,4], Quentin Ferré[1,2], Guillaume Charbonnier[1,2], Magali Torres[1,2], Charles Lecellier[5], Tom Sexton[6], Simona Saccani[3], Salvatore Spicuglia[1,2*]

[1]Aix-Marseille University, Inserm, TAGC, UMR1090, Marseille, France.

[2]Equipe Labélisée Ligue Contre le Cancer, Marseille, France.

[3]Institute for Research on Cancer and Ageing, IRCAN 06107 Nice.

[4]Present address: Vinmec Research Institute of Stem cell and Gene technology (VRISG), Hanoi, Vietnam.

[5]Institut de Génétique Moléculaire de Montpellier - CNRS-UMR 5535 Montpellier, France.

[6]Institut de Génétique et de Biologie Moléculaire et Cellulaire – IGBMC (CNRS UMR 7104, INSERM U1258, Université de Strasbourg) 67404 Illkirch, France.

# Equal contribution

[*]**Correspondence:** salvatore.spicuglia@inserm.fr

Abstract

The action of *cis*-regulatory elements with either activation or repression functions underpins the precise regulation of gene expression during normal development and cell differentiation. Gene activation by the combined activities of promoters and distal enhancers has been extensively studied in normal and pathological contexts. In sharp contrast, gene repression by *cis*-acting silencers, defined as genetic elements that negatively regulate gene transcription in a position-independent fashion, is less well understood. Here, we repurpose the STARR-seq approach as a novel high-throughput reporter strategy to quantitatively assess silencer activity in mammals. We assessed silencer activity from DNase hypersensitive I sites in a mouse T cell line. Identified silencers were associated with repressive histone modifications and enriched for binding motifs of known transcriptional repressors. CRISPR-mediated gene editing validated the repressive function of distinct silencers involved in the repression of non-T cell genes and of genes regulated during T cell differentiation. Finally, we unravel an association of silencer activity with tandem repeats of the ZNF263 binding site, highlighting the role of repetitive elements in silencer activity. Our results provide a general strategy for genome-wide identification and characterization of silencer elements.

Introduction

The precise regulation of gene expression during normal development and cell differentiation requires the action of *cis*-regulatory elements with either activation or repression functions (1-3). Gene activation by the combined activities of promoters and distal enhancers has been extensively studied in normal and pathological contexts. In sharp contrast, gene repression by *cis*-acting silencers, defined as genetic elements that negatively regulate gene transcription in a position-independent fashion, is less well understood. Silencers were first described more than three decades ago in yeast and vertebrates (4-6). Since then, several silencers have been discovered to control the expression of key developmental and immunological model genes, and some progress has been made to characterize various features of a few of these individual silencers (3,7-13). Nevertheless, despite the widely-held belief that silencers likely represent critical general regulators of gene expression, this view is still largely conjectural, and their genome-wide distribution, mechanisms of action and involvement in disease are largely unknown. Noticeably, among the silencers that have been described in the literature, many are associated with the regulation of T cell specific genes. These included silencers associated with the expression of *TCA3/CCL1* (14), *Il2* (15), *CD4* (16), *Tcrb* (17), *ThPOK* (18), *Rag1-Rag2* (19), *CD8* (20) and *Spi1* loci (21). Several of these silencers have been shown to play an important role in cell lineage restriction: for instance, the CD4 and CD8 silencers repress the expression of the associated genes in $CD8^+$ and $CD4^+$ T cells, respectively. T cell differentiation thus provides an excellent model for the implementation of a high-throughput strategy to identify silencers.

Compared to other types of *cis*-regulatory elements, such as enhancers and insulators, silencers have been challenging to map genome-wide (22). Recent efforts included the development of a negative selection strategy (23) or prediction strategies based on chromatin signatures and 3D interactions with repressed genes (9,24-27). Episomal reporter assays have been widely used to characterize putative regulatory regions (1). The development of high-throughput reporter assays for enhancer function has enabled the testing of thousands of distinct DNA sequences simultaneously, by cloning variable DNA fragments into common reporter constructs and using high-throughput sequencing to quantify fragment activity (28). These functional approaches have led to an explosion of discoveries, including their roles in the regulation of many disease-related genes and their involvement in the development of diverse pathologies and cancer (29). In particular, the Self Transcribing Active Regulatory Region Sequencing (STARR-seq) method allows direct genome-wide investigation of

enhancer activity using DNA fragments directly collected from genomic DNA (30). Based on this technique, we previously developed the CapSTARR-seq approach (31), coupling capture of regions of interest to STARR-seq reporter assay, providing a cost-effective method to assess *cis*-regulatory function in mammals.

In line with their operational definitions, assays for silencers could measure their ability to silence gene expression in *cis*, when driven by an independent "strong" promoter (1,10,32). Therefore, to identify silencers genome-wide we suited to repurposed the STARR-seq approach by systematically testing DNase hypersensitive sites (DHS) from developing T cells, using three distinct promoter-based reporter vectors. We compared the set of silencers identified with the different vectors and evaluated their association with genomic and epigenomic features. The robustness of the approaches was extensively assessed by independent episomal reporter assays, and CRISPR/Cas9 genomic manipulation demonstrated the involvement of two endogenous silencers in the repression of neighboring genes. Repetitive elements (REs) have been suggested to contribute to *cis*-regulation, including gene repression (33,34). Here we found that silencers where enriched in REs such as SINEs and Short Tandem Repeats (STRs), while a subset of STRs recognized by the KRAB zinc- finger protein (KZFP) ZNF263 plays an important role in silencing activity. Overall, we provide a general, scalable, and high-throughput approach for the high-resolution experimental dissection of silencer elements in the context of human biology and disease.

Methods

Cell culture

Mouse P5424 T cells (35) and mouse embryonic fibroblast NIH-3T3 (3T3, ATCC: CRL-1658) cells were cultured in RPMI medium (Thermo Fisher Scientific) supplemented with heat-inactivated 10% FBS (PAA) at 37 °C, 5% CO_2. Cells were passaged every 2-3 days and frequently tested for mycoplasma contamination.

Stimulation of P5424 cells

P5424 cells were grown at a density of 3×10^5 cells/ml. Cells were treated with DMSO or PMA at 10 ng/ml (P1585, Sigma) and ionomycin at 0.5 μg/ml (I3909, Sigma) for 4 hours in triplicates as previously described (36).

Cloning of the STARR-Seq vectors

The STARR-seq mammalian screening vector (30) has been kindly provided by Alexander Stark (Vienna, Austria). The synthetic SCP1 promoter present in the enhancer-STARR-seq vector was replaced by the promoter of the ubiquitous PGK gene as defined in (37) or the lymphoid specific promoter of the *Rag2* gene as defined in (38) using In-Fusion homologous recombination. In the R-Ea construct the TCRα enhancer (38) was cloned downstream of the GFP cassette.

CapStarr-seq library generation

Construction and capture of the genomic library have been described previously (31,39). Genomic library was generated from mouse C57BL/6 genomic DNA. For target enrichment, a custom-designed 3 nt resolution oligonucleotide microarray covering 28,055 DHSs identified in mouse $CD4^+$ $CD8^+$ double-positive (DP) thymocytes (31) was constructed using the SureSelect technology (Agilent, 1M format) and the eArray tool default settings (https://earray.chem.agilent.com/earray/). In addition, 437 randomly-selected non-DHS regions were included. The three screening vectors were linearized with AgeI-HF and SalI-HF (New England Biolabs) by 6 h digestion, followed by agarose gel electrophoresis, extracted by QIAquick gel extraction (Qiagen), and cleaned up with Qiagen Minelute PCR purification Kit (Qiagen). After, 500 ng of amplified captured DNA was recombined by with 2000 ng of linearized screening vectors in a total of 10 μl per reactions (each having 50 ng of captured DNA and 200 ng of screening vector) (Clontech In-Fusion HD). All the recombination reactions were pooled together and purified with Agencourt AMPureXP DNA beads (Thermo Fisher Scientific) and then eluted in 29 μl. Thirteen aliquots (20 μl each) of MegaX DH10B Electrocompetent Bacteria of (Thermo Fisher Scientific) were transformed with 2 μl of DNA each, according to the manufacturer's recommendation. After 1 h recovery at 37 °C, the transformations were pooled together and transferred into 2 liters of LB overnight for each specific promoter vector. An aliquot of each transformation was plated on LB AMP medium to estimate the number of cloned fragments. A total of 6-8 million clones were achieved by each library. Finally, plasmid libraries were purified using Qiagen Plasmid Plus Maxi Kit (Qiagen).

CapStarr-seq library transfection

For each library, a total of $50x10^6$ cells was transfected in triplicate (5 µg/$1x10^6$ cells) using the Neon Transfection System (Thermo Fisher Scientific). The P5424 cells were transfected with 1600V-20ms-1 pulse conditions. After transfection, cells were transferred to a complete growth medium and incubated for 24 h before isolation of the RNA.

CapStarr-seq RNA isolation from transfected cells

RNA extraction was performed using the RNeasy miniprep kit (Qiagen) with the on-column DNaseI treatment. The PolyA RNA fraction was isolated by µMACS mRNA isolation kit (Miltenyi Biotec) following the manufacturer's recommendations. PolyA RNA was treated with Ambion turbo DNase (Thermo Fisher Scientific) and then purified with RNeasy Minelute kit (Qiagen). Finally, mRNA was quantified by using Qubit RNA HS Kit (Thermo Fisher Scientific).

CapStarr-seq reverse transcription and sequencing library preparation

cDNA first-strand synthesis was performed for non-stimulated (NS) and stimulated (PMI/I) cells with superscript III (Thermo Fisher Scientific) using a reporter-specific primer (5'-CAAACTCATCAATGTATCTTATCATG-3') and 0.2 to 0.3 µg of polyA RNA per reaction for a total of 10 reactions. After the reverse transcription, 1 µl of RNaseH was added and incubated at 37°C for 1 h. The cDNA was then purified with QIAquick PCR purification kit and determined concentration using Qubit ssDNA Kit (Thermo Fisher Scientific). The cDNA was amplified using the KAPA Hifi Hot Start Ready Mix in a 2-step nested PCR. In the first PCR (98°C, 2min; followed by 15 cycles of 98°C for 20 s, 65°C for 20 s, 72°C for 30 s), cDNA of 5 ng per reaction was amplified using two reporter-specific primers (fw: 5'-GGGCCAGCTGTTGGGGTG*T*C*C*A*C-3)' and rw: 5'-CTTATCATGTCTGCTCGA*A*G*C-3'), one of which spans the splice junction of the synthetic intron, in a total of 10 reactions. Purification of the PCR products was performed on gel using QIAquick gel extraction kit (Qiagen) followed by a clean-up with QIAquick mini elute PCR purification kit (Qiagen), to remove any residual contamination of plasmid or cDNA. Generating Ion Torrent libraries, purified PCR product was used as a template for the second PCR (5 ng/PCR, for a total of 10 PCR reactions; 98°C, 2min; followed by 10 cycles of 98°C for 20 s, 65°C for 20 s, 72°C for 30 s) with KAPA Hifi Hot Start Ready Mix and Ion Torrent library amplification primer mix (Thermo Fisher Scientific, T_PCR_A: 5'-CCA TCT CAT CCC TGC GTG TC-3' and P1amp: 5'-CCA CTACGC CTC CGC TTT CCT CTC TAT G-3'). Generating the INPUT control, 10 reactions with 5 ng of reporter constructs (library)

per reaction were amplified using the same conditions as mentioned above, except for forward primer in the first PCR (fw: 5'-GGGCCAGCTGTTGGGGTG*A*G*T*A*C-3'). To assess potential biases in library composition caused by electroporation, 10 reactions with 5 ng per reaction of the reporter constructs isolated from transfected cells were amplified as explained above. For sequencing the libraries, the sequencing indexes are added to the libraries by one simple PCR reaction using Multiplex oligos for Illumina (NEBNext Ultra RNA library prep). So, the non-transfected (plasmid input) and transfected libraries were sequenced on the Illumina NextSeq 500 platform (**Table S1**).

CapStarr-seq data processing

FastQ files of transfected (cDNA) and non-transfected (input) libraries were trimmed using sickle with -q 20 option and mapped to the mm9 reference genome using bowtie2 with default parameters. Sam files were converted using samtools to BAM files. BAM files of all replicates were pooled and converted into bed files using BedTools (v2.28.0) (40). In this approach, each read is considered as a captured region and independent clone. First using basic bash commands, we create for each condition a unique list of clones present in cDNA and input libraries. To count the frequency of each unique clone, the length was set to 1 bp to avoid multiple counting clones, then it was count using BedTools intersect -c using the original bed files with the unique list. This led us to get the number of clones present in cDNA and input. The length of all clones was restored to 314 nt, corresponding to the average size of the captured fragments. Each DHS region was annotated with a unique ID, then each clone was annotated by the ID of the overlapping region. We extended the original DHS coordinates by considering the start of the first clone and the end of the last clone overlapping the DHS region using BedTools and python homemade scripts. These regions are considered as the extended DHS. The extended DHS regions were split using bins of 50 bp. Using a R homemade script, we computed the activity of the extended DHS regions computing the fold change of the sum of the clones overlapping the same region in the transfected condition over the non-transfected condition. The count of clones was normalized using FPKM and activities were centred. We excluded the region which has an FPKM < 1 in the input. The subregion activity was computed in the same way, using the clone that overlaps the subregion created using the bins of 50 nt. We defined silencers as the extended DHS regions with $\log_2$(activity) lower than -1. In these regions, we defined the core silencer as the consecutive subregions with $\log_2$(activity) lower than -1. Then, we identified the edge silencer as the consecutive subregions with the lowest activity of the core silencer. A silencer can have only one edge of

N subregion with strict equal minimal activity. All the regions were annotated with the two nearby genes using GREAT web-service (41). All results are summarized in the **Table S2** and silencers are summarized in the **Table S3**. To visualize the CapStarr-seq signal per individual cloned fragments or by regions, we generated a bed file with a color code proportional to the activity ranging from green for positive activity to red for region with negative activity (silencer). To assess reproducibility between replicates and conditions, we generated a correlogram using ggplot2 (https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.147) and corrplot package (https://cran.r-project.org/web/packages/corrplot) with a Spearman non-parametric test.

Definition of silencers

Putative silencers were defined as DHS regions displaying a CapStarr-seq signal ($\log_2$(FC)) lower than or equal to -1 in any of the CapStarr-seq experiments. For each library, a control group of the same number of DHS as the silencers were created by randomly selecting DHS with $\log_2$(FC) ranging between -0.1 and 0.1. A list of active enhancers was created using the SCP1 condition taking the DHS with a $\log_2(FC) > 1$. The relative proportion of proximal (≤ 1 kb) and distal (> 1 kb) silencer was computed using the distance to TSS of the genes provided by RefSeq (42).

Epigenomic analyses

ChIP-seq datasets from P5424 cells and DNaseI-seq and ChIP-seq datasets from mouse DP thymocytes were downloaded from Gene Expression Omnibus (GEO) database as detailed in **Table S1.** ChIP-seq data were processed as described in (31). For average profiles, Wiggle files were converted to BigWiggle files using wigTo BigWig (43). Average profiles were generated with deepTools (44) using MNase-seq, DNaseI-seq and ChIP-seq signals from bigwig files around the DHS center (± 5 kb for histone modifications, ± 1 kb for TFs and chromatin features).

RNA-seq data

Public RNA-seq data from P5424 cells (36) and mouse T cell differentiation (45) were downloaded from GEO database (accession numbers GSE120655 and GSE48138, respectively). The raw RNA-seq data was processed as described in (36).

Association of silencers with variable genes

RefSeq genes in a window of ±100 kb around all silencer candidates were retrieved using BedTools window (v2.28.0). The different conditions were annotated using the mm9 file from UCSC. To retrieve the top 5% most variable genes, the variance through T cell differentiation was computed for each gene using merged replicates of the RNA-seq dataset (45). Those 5% most variable genes were considered as T cell regulated genes.

Tissue specific genes

Gene tissue-specificity score was computed by adapting the calculation method of entropy of Shannon to data expression from GNF Gene Atlas (46). This led us to compare the distribution of expression of genes across different tissues. High entropy score indicates uniform distribution meaning that the gene is not tissue-specific. Low entropy (entropy ≤ 3) score indicates high tissue-specific genes.

Motif research analysis and clustering

The HOMER (47) software was used to perform research motif analysis with the option "findMotifsGenome.pl input.file.bed mm9 output.file.bed –len 6,8,10,12,15 –size given –bg dhs". We choose as background all the DHS sites in order to isolate the core silencer-specific binding sites. This method was also used within STR contained in the core silencers subregions. From the Known motif output files, a list of transcription factors was obtained. The relative matrix files were downloaded from JASPAR 2018 database (48). In order to reduce the redundancy, the RSAT matrix clustering tool (49) was used with default parameters to cluster motifs based on their sequences. The best transcription factor *P*-values provided by HOMER was kept by cluster (**Table S4**). Then activity of silencers containing TFBS based on genomic track from JASPAR 2018 identify through HOMER was compared. Visualization of ZNF263 tandem repeats was made with TFmotifView tool (50)

Genomic and RE enrichment of candidate regions

Genomic distribution was analyzed using OLOGRAM (51), included in pygtftk (52). Regions outside DHS were excluded to compare our candidates with the DHS distribution. mm9 GTF files from Ensembl were provided to assess the genomic distribution of our candidates. We ploted with ggplot2 the $\log_{10}$ (*P*-value) with a factor ± 1 depending on whether it is enriched or depleted to obtain the enrichment score. Enrichment of REs was performed as described above using RepeatMasker (v4.0.8) file for mm9 (53).

RE analysis

The genomic position of STR elements within the core silencer were obtained using BedTools intersect to get overlapping regions between STR repetitive elements based on RepBase (54). The impact of the number of repeats on silencer activity was investigated using HipSTR reference file (55) in mice giving the details of simple repeat elements in the genome. Reference file was converted from mm10 to mm9 using LiftOver. Using BedTools, silencers containing simple repeats were identified and then separated by the number of repeats. In order to investigate ZNF263 as STRs, the number of ZNF263 TFBS based on the JASPAR 2018 database was counted in the extended DHS, the core silencers and the subregion flanking the core silencer within the silencer region using BedTools intersect with the -c option to report count number of overlaps.

Functional analysis

Functional enrichment analyses of putative silencers identified in each condition and putative enhancers coming from the SCP1 strategy were produced using a custom pipeline to automate multi-sample queries to the GREAT web-service. In summary, putative silencers regions and putative enhancers regions were queried with rGREAT R package (https://github.com/jokergoo/rGREAT,http://great.stanford.edu/public/html/) against GO Biological Process. Genomic association rules with genes were done using default GREAT "Basal plus extension". Significant terms were filtered as those having Binomial Fold Enrichment higher than 2 and both Binomial and Hypergeometric tests Benjamini-Hochberg adjusted *P*-values lower than 0.05. GOSemSim R package (56) was used to compute the Wang similarity distance between all terms. Terms with a similarity distance higher than 0.8 were grouped. Terms were further filtered for heatmap display to keep only the best 5 terms for each sample according to the binomial *P*-value. The color scale on heatmaps displays Binomial Benjamini-Hochberg adjusted (BH) *P*-values.

Luciferase reporter assays

Silencer candidates or control regions were amplified from mouse genomic DNA or synthesized *in vitro* and cloned downstream of the luciferase gene at the BamHI and Sal1 restriction site in the pGL3-Promoter vector (Promega) (**Table S5**) and verified by Sanger sequencing. A total of 1×10^6 P5424 cells were co-transfected with 1 µg of the tested construct and 200 ng of *Renilla* vector using the Neon Transfection System (Thermo Fisher

Scientific). Electroporation conditions for P5424 cells were maintained at 1600V-20ms-1 pulse and for NIH-3T3 were maintained at 1350V-20ms-2 pulse. After 24 hours of transfection, luciferase activity was measured using the Dual-Luciferase Reporter Assay kit (Promega) on a TriStar LB-941 Reader. For all measurements, firefly luciferase values were first normalized to *Renilla* luciferase values (controlling for transfection efficiency and cell number). Data are represented as the fold decrease in relative luciferase signal over the pGL3-Promoter vector (pSV40). All experiments were performed in triplicates.

FACS analysis

A total of 1×10^6 P5424 cells were transfected, as described above, with the indicated vectors. Twenty-four hours post electroporation, GFP expression was assessed on a FACS Calibur (BD Biosciences).

Site-Specific Mutagenesis

Mutagenesis was performed using Q5 Site-Directed Mutagenesis Kit (New England biolabs) following the manufacturer's instructions. The mutagenesis primers were designed using the NEBase Changer tool and are shown in (**Table S6)**. All the mutations were verified by sanger sequencing.

Hi-C processing

Raw Hi-C data from primary DP thymocytes were taken from Hu et al. (57) and processed with FAN-C (58), entailing iterative mapping to the mm9 genome assembly with bowtie2, filtering self-ligation events and PCR duplicates, binning the data to 10 kb bins and balancing the chromosome-wide matrices with the Knight-Ruiz method. TAD boundaries were identified by computing insulation scores (59) with windows of 100 kb (10 bins), normalizing to chromosome-wide averages of insulation scores, then filtering the local minima with the delta vector calculated for the three bins flanking the computed one, and with the difference of the minima and maxima of the delta vector being at least 0.7.

CRISPR-Cas9 genome editing

For the CRISPR–Cas9 experiments of targeted silencer regions, two gRNAs were designed for each end of the targeted region using the CRISPRdirect tool (60). The designed gRNAs were cloned into a gRNA cloning vector (Addgene, 41824) as described previously (61). One million cells were transfected with 1 µg of each gRNA and 1 µg of hCas9 vector (Addgene,

41815) using the Neon Transfection System (Thermo Fisher Scientific) and cultured in 5 mL. After two days of transfection, the transfected cells were plated in 96-well plates at limiting dilution (0.5 cells per 100 µl per well) for the clonal expansion. Individual cell clones were screened for homologous allele deletion after 10-14 days, by direct PCR using Phire Tissue Direct PCR Master Mix (Thermo Fisher Scientific) according to the manufacturer's protocol. For the detection of knockout or wild-type alleles, forward and the reverse primers were designed to bracket the targeted regions. The clones were considered to have undergone homologous allele deletion if they had no wild-type band and at least one deletion band of the expected size. The homozygous deletions were verified by sanger sequencing. All the gRNAs and primers are listed in **Table S6.**

Gene expression analyses

Total RNA from P5424 from WT and deleted clones in non-stimulated or PMA/Ionomycin treated conditions was extracted using the RNAeasy mini kit (Qiagen). Three micrograms of RNA was treated with DNase I (Thermo Fisher Scientific) and was quickly reverse transcribed into cDNA using Superscript VILO Master Mix (Thermo Fisher Scientific). The Real-time PCR was performed using Power SYBR Master Mix (Thermo Fisher Scientific) on a Quant Studio 6 Flex instrument (Thermo Fisher Scientific). Primer sequences are listed in **Table S6.** The expression of the gene was normalized to that of *Rlp32*. Relative expression was calculated by the ΔC_T method, and all the data shown are reported as fold change over the control. For each of the cell clones, from the three independent RNA/cDNA preparations, the Student's t-test was performed (unpaired, two-tailed, 95% confidence interval). Data are represented with s.d. For the conventional RT–PCR, one-twentieth of the synthesized cDNA was used as the template for the reaction.

Results

Experimental strategies to identify silencer elements

To set up an experimental strategy to quantify silencer activity, we repurposed the CapSTARR-seq technique (31), an approach coupling capture of defined regions to the previously developed STARR-seq technique (30) (**Figure 1a**; see also Methods section). We replaced the basal synthetic Super Core Promoter 1 (SCP1) promoter present in the original STARR-seq vector with a ubiquitous strong promoter from the human *PGK* gene (pPGK) (37) and a T cell-specific promoter enhancer pair pRag2-Eα (pR-Ea) (38) (**Figure 1A**).

Analyses of GFP expression by FACS demonstrated the expected promoter activity of the derived STARR-seq vectors (**Supplementary Figure S1A-B**). Although little is known about the general biochemical properties of silencers, it seems reasonable to assume that a subset of silencers may be occupied by sequence-specific transcription factors and/or lie within nucleosome-depleted genomic regions, and consequently may overlap with DNase I Hypersensitive Sites (DHS). Indeed, several known silencers have been identified as laying within DHS sites (e.g., (18,32,62)). Therefore, to isolate silencer elements genome-wide, we designed a captured library containing 28,055 DHSs from mouse double-positive (DP) thymocytes, plus 437 randomly-selected non-DHS regions as negative controls.

In brief, DNA fragments of ~400 bp were captured on a custom-designed microarray covering all the DHS and cloned by homologous recombination into the three different STARR-seq vectors (hereafter named, SCP1, pPGK and pR-Ea libraries). The STARR-seq libraries were transfected in triplicate into the mouse T-cell line P5424 and sequenced by targeted RNA-seq (**Table S2**). The P5424 cell line originated from early developing T cells and resembles DP thymocytes at phenotypic and transcriptomic levels (35,36) and has been previously used in STARR-seq experiments (31,63). As controls, we sequenced the libraries before transfection (hereafter named, input). The *cis*-regulatory activity was assessed by computing the $\log_2$ ratio between the targeted RNA-seq and the corresponding input signals (hereafter referred to as STARR-seq signal) after normalization and filtering (See Methods section; **Figure 1B; Table S2**). A good correlation was obtained between the replicates of the same library (Spearman's correlation coefficient (ρ) ranging between 0.38 and 0.87; **Supplementary Figure S1C**). For subsequent analyses, the signals from the replicates of the same library were merged.

Comparison of STARR-seq signal between DHS and random regions for the different libraries showed that enhancers are significantly overrepresented in DHSs as compared to random regions in the SCP1 library (**Figure 1B**). In contrast, DHSs with silencer activity were significantly detected with the pPGK and pR-Ea libraries. These results suggested that using a strong promoter-based library is indeed an effective strategy to detect silencer elements. We defined the putative silencers as DHS regions with a $\log_2$ STARR-seq signal lower than or equal to -1 (**Figure 1B**). The silencer definition resulted in a set of 1249, 672 and 413 putative silencers for the SCP1, pPGK, and pR-Ea libraries, respectively (**Table S3)**. Given the comparison of the STARR-seq signal between DHSs and random regions (**Figure 1B**), we expect that many of putative silencers identified with the SCP1 library might be false

positive (Figure 1C provided an example of a silencer region identified with the pPGK library showing the STARR-seq signal for the individual cloned fragments. Core silencers were identified as a region containing successive bins of 50 bp with a $\log_2$ STARR-seq signal lower than or equal to -1 (**Figure 1C**; **Table S2**). Correlation of silencer activity between the different libraries was relatively low, indicating that the use of distinct promoters enables the identification of distinct, yet overlapping, sets of silencers, and suggesting that silencers may exhibit promoter-specific activities, as has been shown for enhancers (64). Consistently, 15%, **27.5%** and 18% of silencers identified with the SCP1, pPGK and pR-Ea libraries, respectively, were also found within another library (**Figure 1D-E**), while only 15 putative silencers were shared between all the libraries (two examples are displayed in **Figure 1F**).

To assess whether there was a bias in the genomic location of silencers, we computed the specific enrichment of putative silencers obtained with each library with respect to all DHS (**Figures 1G and H**). While SCP1 and pR-Ea based silencers were depleted from promoter regions, the pPGK based silencers were enriched. Importantly, none of the silencer sets were enriched for terminator sequences (**Figure 1H**), which could represent a potential bias of the approach by artificially interfering with the quantification of the STARR-seq vector-derived transcripts.

Validation of STARR-seq identified silencers

To independently evaluate the accuracy of STARR-seq to identify silencers, we selected 24 DHS candidate silencers (13 common in at least two libraries, 8 specifics to the pPGK and 3 specific to the pR-Ea libraries), as well as 12 DHS control regions ($\log_2$ STARR-seq signal close to 0). The selected DHS were tested in a classical luciferase reporter assay in the P5424 cell line (**Figure 2A**). We found that 87.5% (21 out of 24) putative silencers and 33.3% (4 out of 12) control DHS regions displayed significant silencer activity in the luciferase assay. Overall, the STARR-seq identified silencers displayed a higher silencer activity in the luciferase assay as compared with the control DHS set (P -value = 0.0001; **Figure 2B**). Importantly, several STARR-seq-defined silencers (*DHS26112, DHS2610, DHS10824, DHS5667, DHS12366* and *DHS23650*) displayed a strong silencer effect resulting in luciferase expression close to background levels, while this was not observed for any of the control regions. Candidate silencers identified in more than one condition tended to display higher silencer activity (**Figure 2A**). Indeed, 6 out of the 10 DHS with the strongest silencer activity were found in more-than-one STARRseq library, whereas only 3 out of the 10

weakest candidates were. Moreover, silencer activity was independent of the orientation of the tested region with respect to the luciferase gene (**Figure 2C**). To further assess the silencer activity, we cloned one of the validated silencers (DHS12366) into a GFP-containing reporter vector. After transfection in P5424 cells, the vector with the *DHS12366* silencer displayed reduced GFP expression as compared to the control vector (Figure 2D). Thus, consistency between the independent reporter assays indicates that the high-throughput assessment of silencer activity by STARR-seq is highly accurate.

Chromatin features and gene functions associated with silencers

To assess the meaningfulness of the STARR-seq identified silencers, we assessed their association with chromatin features and expression of neighbor genes as compared with corresponding control DHS (see Methods section) and enhancers identified by the SCP1 library (**Figure 1A**). To explore whether silencer activity reflects the epigenetic status of the endogenous DHSs, we analyzed several chromatin features available from the P5424 cell line (**Figure 3A**) and the primary DP thymocytes (**Supplementary Figure S2**). Silencers were associated with a lower level of active histone marks (H3K4me3, H3K27ac and H3K4me1) (**Figure 3A** and **Supplementary Figure S2A**), as compared with control DHSs and SCP1-enhancers, but displayed a similar level of DNase I accessibility and CTCF binding (**Supplementary Figure S2B**). Both the repressive marks H3K27me3 and H3K9me3 were present at elevated levels at SCP1-silencers compared to control DHSs, while pPGK-silencers and pR-Ea-silencers were modestly enriched in H3K27me3 and H3K9me3, respectively (**Supplementary Figure S2A**). Overall, these results suggested that silencer elements are found in a relatively open chromatin configuration and might be associated with different types of repressive mechanisms.

We next used the GREAT tool (41) to associate each DHS to their neighbor genes and assess gene expression using available RNA-seq data from P5424 cells (36) (**Figure 3B**). Silencers identified by the three different libraries were significantly associated with genes expressed at lower levels than those associated with control DHS. As expected, enhancer regions identified by the SCP1 library were associated with genes expressed at a higher level. Functional enrichment analysis showed that the identified silencers were associated with genes involved in immune and T cell phenotypes and functions, but generally different from those associated with SCP1-enhancers (**Figure 3C-D**), suggesting that at least a subset of silencers might be involved in the repression of T cell associated genes.

Transcription factors associated with silencers

To assess whether putative silencers were enriched for transcription factor binding sites (TFBS) we performed motif enrichment analyses using the HOMER tool (47) on the three sets of identified core silencers, as well as, the set of active enhancers (**Figure 4A**; **Supplementary Figure S3**). Active SCP1-defined enhancers were enriched in binding sites of TFs involved in T cell differentiation, including MYB, TCF, RUNX, ROR and ETS/NFATC, consistent with previous results (31). In contrast, all three sets of silencers were enriched in TFBS bound by known transcriptional repressors, including CTCF, HOX TF family, SMAD and ZNF263. Strikingly, the pPGK and SCP1 based silencers, but not pR-Ea silencers, were strongly enriched in binding sites for the RE1-Silencing Transcription factor (REST), a major transcriptional repressor involved in the repression of neural genes in non-neuronal cells (65-67). To better evaluate the impact of TF binding sites on silencer activity, we plotted the significance of the enrichment of each TFBS against the mean activity of the core silencers (**Figure 4B**). We found that several TFBS were associated with strong silencer activity, including REST, SMAD3/4, MAFK and the HOX TF family. The strongest silencer activity was associated with the presence of a REST or MAFK motifs in silencers found with either SCP1 or pPGK vectors, or the presence of SMAD3 in SCP1-detected silencers. Finally, we assessed the relevance of REST and SMAD binding sites by mutating these sites in silencer candidates and assessing the silencer activity by the luciferase assay (**Figures 4C and 4D**). Mutation of either REST or SMAD3 binding sites significantly de-repressed the luciferase expression, indicating that these two TF binding sites are indeed important contributors to the silencer activity.

Dynamic silencer activity mediated by TCR signaling

To explore whether silencer activity can be regulated by T cell stimulation, we performed STARR-seq experiments using the pPGK library in P5424 cells treated with PMA and Ionomycin (PMA/I), previously shown to partially mimic TCR signaling and consecutive T-cell differentiation (36) (**Supplementary Figure S4A; Tables S1-3**). We observed that a majority of silencers active in PMA/I-treated cells were stimulation specific (59%) (**Supplementary Figure S4B**) and were generally associated with genes downregulated after T cell stimulation (**Supplementary Figure S4C**). Luciferase reporter assays for three induced silencers validated their stimulation-dependent activity (**Supplementary Figure S4D)**.

Analysis of motif enrichment did not reveal any TFBS specifically enriched in PMA/I-dependent silencers (**Supplementary Figure S4E**). However, we observed that REST-containing motifs were exclusively enriched in constitutive silencers, suggesting that binding of REST transcriptional repressor provides general silencer activity.

REST-containing silencers contribute to the repression of non-T cell genes

To gain insight into the contribution of REST to silencer activity, we determined the set of identified silencers harboring a REST binding motif (**Figure 5A**). Only 5% of identified silencers with SCP1 and pPGK libraries contained REST binding sites, however, those silencers were associated with significantly lower levels of expression of neighboring genes (**Figure 5B**). Moreover, REST-containing silencers were associated with a higher proportion of tissue-specific genes not expressed in T cell precursors (**Figure 5C;** chi-square test, *P* -value = 0.001 and 0.04 for the pPGK and SCP1 libraries, respectively), and also with a reduced proportion of T-cell regulated genes compared to non-REST-containing silencers (**Figure 5D**; i.e., genes whose expression is highly variable across T cell differentiation; see Methods section).

To experimentally explore the contribution of REST-containing silencers to the regulation of gene expression, we performed CRISPR/Cas9-mediated deletion of DHS12366, a silencer candidate identified with the SCP1 and pPGK vectors and displaying REST-dependent silencer activity by luciferase assay (**Figure 4C**). We obtained two P5424 clones with homozygous deletion of the DHS12366 region (**Supplementary Figure S5A)**. The DHS12366 silencer was located close to the 3' side of a TAD (**Figure 5E**). Consequently, we compared the expression of all genes present in the two adjacent TADs. Strikingly, we observed that seven genes displayed significantly increased expression in the two DHS12366-deleted clones as compared with wild-type P5424 cells (**Figure 5F**), including two genes consistently displaying more than 2-fold increased expression (*Plin4* and *Arrdc5*). Interestingly, several of the deregulated genes were expressed in a tissue-specific manner in non-T-cell tissues (**Figure 5G**). Consistent with a ubiquitous activity, the DHS12366 silencer displayed silencer activity also in the fibroblast cell line NIH-3T3 (**Supplementary Figure S6C**). Overall, our results support the idea that REST-containing silencers repress tissue-specific genes in other unrelated tissues.

The DHS23650 silencer regulates two genes involved in T cell function

Next, we sought to identify silencers involved in normal T cell function. To this aim, we searched for putative silencers that might control the expression of genes regulated across T cell development and differentiation. We reasoned that lymphoid genes regulated by silencers might have a high expression variance across T cell populations. Thus, we isolated the top 5% of highly variable genes based on available RNA-seq from different stages of thymic and peripheral T cell differentiation (45) (**Figure 6A**). We then retrieved the top variable genes located in a window of 100 kb around any STARR-seq-defined silencer (**Figure 6B**). We obtained 516 genes associated with 615 silencers. Of these, we identified the DHS23650 silencer associated with *Hcst* (Hematopoietic cell signal transducer, also known as DAP10) and *Nfkbid* (Nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, delta), two genes involved in T cell differentiation and activation. The *Hcst* gene encodes for a transmembrane signaling adaptor which forms part of the immune recognition receptor complex (68,69). This receptor complex has a role in cell survival and proliferation by activation of NK and T cell responses. The *Nfkbid* gene encodes for a member of the atypical inhibitors of NF-κB TF and is particularly involved in the regulation of T cell activation and development of regulatory T cells (70,71). Strikingly, the expression of both genes is anticorrelated through T cell differentiation. In particular, *Hcst* is induced during T helper (Th) maturation, while *Nfkbid* is repressed (**Figure 6C**).

Hcst and *Nfkbid* were both located within the TAD containing the DHS23650 silencer (**Figure 6D**), suggesting that they could be a direct target of this silencer. Luciferase reporter assay demonstrated a strong silencer activity for the DHS23650 silencer in both orientations (**Figure 6E**). To explore the *in vivo* function of the DHS23650 silencer, we deleted this element in the P5424 cell line using CRISPR/Cas9 genome editing (**Figure 6F; Supplementary Figure S5B**). We analyzed the expression of all genes contained in the same TAD as the DHS23650 silencer (**Figure 6F**; only genes with detectable expression are shown). Only the *Hcst* gene appeared to be significantly up-regulated in the two DHS23650-deleted clones. Interestingly, we observed that *Nfkbid* was specifically upregulated by the PMA/ionomycin stimulation (**Figure 6D**), in agreement with its induction between the $CD4^-$ $CD8^-$ double negative (DN) and DP stages (**Figure 6C**). Analysis of expression in wild-type and mutated P5424 cells stimulated by PMA/ionomycin revealed a role of DHS23650 silencer in limiting *Nfkbid* induction (**Figure 6G**), consistent with the conserved silencer activity of DHS23650 after PMA/ionomycin stimulation (**Supplementary Figure S6A**). Therefore, the DHS23650 silencer regulates both *Hcst* and *Nfkbid* genes in different stimulatory contexts.

However, the effect of DHS23650 on *Hcst* and *Nfkbid* expression appeared to be mild (<2 fold) in comparison with the expression changes that these genes display during T cell differentiation (**Figure 6C**). This might be due to a limitation of our cell line model or indicates that DHS23650 modulates the transcription level of the target genes rather than being responsible for their complete repression.

As both *Hcst* and *Nfkbid* are important regulators of T cell differentiation and activation we analyzed the expression of several T cell markers previously validated in the P5424 cell line (**Supplementary Figure S6B**). Interestingly, the regulation of *Lef1*, *Ptcra* and *Bcl2* after PMA/ionomycin stimulation of P5424 cells is significantly altered in the two DHS23650 deleted clones, suggesting that the DHS23650 silencer might be required for normal T cell differentiation. Consistent with the regulation of T-cell specific genes, the silencer activity of DHS23650 was not conserved in the fibroblast cell line NIH-3T3, in contrast to the REST-containing DHS12366 silencer (**Supplementary Figure S6C**). Overall, these results show that our STARR-seq derived approach is able to identify a silencer element involved in the control of T-cell regulated genes.

Repetitive elements (REs) associated with silencers

With the aim of identifying regulators associated with DHS23650 silencer activity, we looked for potential repressor candidates based on Jaspar TFBS (**Supplementary Figure S7A**). We identified GFI1 (Growth Factor Independent 1 Transcriptional Repressor) and ZNF263 binding sites as potentially relevant repressor elements, as they were present in the DHS23650 silencer and their motifs were generally enriched in silencer elements (**Figure 4A-B**). GFI1 is a nuclear zinc finger protein that functions as a transcriptional repressor and is essential for hematopoiesis and involved in Th2 differentiation pathway and T-cell receptor signaling (72-75). However, mutation of the GFI1 binding site did not affect the silencer activity of DHS23650 (**Supplementary Figure S7B**). The ZNF263 motif was found as a tandem repeat of 16 binding sites in the 5' side of DHS23650 (**Figure 7A and Supplementary Figure S7A**). ZNF263 is a C2H2 ZNF that contains 9 Zinc finger domains and a KRAB repression domain (76) and is ubiquitously expressed. It was found amongst the top enriched *de novo* motifs in SCP1 and pPGK silencers (**Supplementary Figure S3**). Interestingly, the presence of the ZNF263 tandem repeat in the DHS23650 silencer was reminiscent of a previous study showing that the high density of identical motifs of ZEB1 repressor in tandem repeats can make them suitable platforms for recruitment of transcriptional repressors (77). To

experimentally validate the impact of ZNF263 tandem repeats on silencer activity, we generated a series of DHS23650 mutations and tested the silencer activity by luciferase reporter assays (**Figure 7B**). Deletion of the ZNF263 tandem repeat resulted in a significant reduction of DHS23650 silencer activity while the ZNF263 tandem repeat alone displayed similar silencer activity as the full-length DHS23650. To assess whether the number of ZNF263 repeats is important for the silencer activity of DHS23650 we mutated four, eight or all of the ZNF263 repeats (**Figure 7B**). Strikingly, increased mutations of ZNF263 repeats resulted in the progressive decrease of the silencer activity. Thus, ZNF263 tandem repeats appear to play an essential role in the silencer activity of DHS23650. To assess whether the ZNF263 binding site was frequently present in tandem, we analyzed the density of the ZNF263 binding site (i.e., the number of ZNF263 binding sites per 100 bp) found in pPGK-defined core silencers (**Figure 7C**). Core silencers were significantly associated with a higher density of ZNF263 repeats as compared with other DHS regions or the regions flanking the core silencers (see examples of ZNF263 tandem repeats in **Supplementary Figure S7C**). Moreover, we found that the increased number of ZNF263 sites was associated with a significant increase in silencer activity (**Figure 7D**). Thus, ZNF263 tandem repeats appeared to play an important role in the activities of multiple silencers.

The above results raised the question of whether REs could be involved in silencer activity. Large portions of mammalian genomes are derived from REs, which are linked to TF binding (78-80). RE elements have been associated with both enhancer (78,81) and repressive activities (82). To more generally assess STARR-seq silencers for the occurrence of RE sequences, we used the RepeatMasker annotation (83). The number of RE-derived sequences in silencer regions was compared to the number detected in all DHS regions (**Figure 7E**). pPGK-based silencers were highly enriched in SINE, STR and small RNAs. These results indicate that certain families of REs are overrepresented at silencers. However, different type of REs are enriched depending on the library used to identify the silencers.

STRs are short sequences of DNA, normally of length 2-5 base pairs, that are generally repeated 5-50 times. It has been shown that STRs could regulate gene expression by diverse mechanisms, including recruitment of transcriptional repressors (34,84). Given the observation that ZNF263 binding sites are frequently found in tandem, we assessed their overlap with annotated STRs (85). As shown in Figure 7F, roughly half of pPGK-defined silencers containing STRs also contains ZNF263 BS. Consistently, ZNF263 is the most

enriched TFBS in STRs present in silencers (**Figure 7G**). Finally, we assessed whether the length of the STRs could impact silencer activity. Reminiscent of the impact of ZNF263 repeats length on silencer activity of the DHS23650 silencer, we observed that silencer activity of pPGK-based silencers significantly increased with the length of tandem repeats (**Figure 7H**). Therefore, the STRs, by serving as a platform for the recruitment of ZNF263, might play an important role in the silencer activity of *cis*-regulatory elements in T cells.

Discussion

Despite the widely-held belief that silencers represent critical general regulators of gene expression, their genome-wide distribution, mechanisms of action and involvement in disease are largely unknown. A breakthrough in the analysis of distal *cis*-regulatory elements was provided by the development of high-throughput reporter assays to assess promoters (86,87), enhancers (28) or insulator activities (88). Such a strategy for the identification of silencer elements has been missing. Here, we repurposed the widely used STARR-seq approach to identify silencers systematically and efficiently.

An operational assay for the silencer function of a genetic element is to isolate it and measure its ability to repress promoter activity in a given cell type (10,32). To screen for silencers in a manner optimizing throughput and functional information, we adapted a high throughput reporter assay based on the CapSTARR-seq strategy developed previously (31). We identified DNA fragments capable of negatively regulating their transcription in a minimal, episomal context in transfected cells, allowing for genome-wide screening of putative silencers from several tens of thousands of genomic regions. As an initial assessment of our approach, we analyzed a set of 28,055 DHS from primary developing mouse thymocytes in a T cell line previously used in STARR-seq (31). We tested STARR-seq vectors containing either a basic (SCP1), strong (pPGK) or T-cell specific (pR-Ea) promoter. We extensively assessed the accuracy of STARR-seq to quantify silencer activity and demonstrated its robustness to identify *bona fide* silencers. Overall, the basic SCP1 library allowed identifying both enhancers and silencers, although no significant enrichment for silencer elements was observed. In contrast, the libraries harboring the ubiquitous pPGK promoter or the T-cell specific promoter-enhancer pair (pR-Ea) were significantly enriched in DHS harboring silencer activity, thus suggesting that STARR-seq vectors with strong promoters perform well in the identification of silencer elements. We observed a relatively low level of overlap between the silencers identified with the three libraries, suggesting that silencers may exhibit

promoter-specific activities, as has been shown for enhancers (64). We observed that silencers identified by the SCP1 and pPGK libraries shared similar characteristics, such as enrichment for REST and ZNF263 tandem repeats, in contrast to the pR-Ea library, as discussed below. Another striking feature was the specific enrichment of silencers overlapping gene promoters observed with the pPGK library. This is reminiscent of a recent study showing that "poised" gene promoters exhibit a silencer-like function to repress the expression of distal genes via promoter-promoter interactions (89). Together with the recent findings that many promoters can act as distal enhancers of other genes, also named Epromoters (90-92), these observations support a unifying model whereby single DNA sequences can encode different types of regulatory functions, including being a promoter for immediate genes, or an enhancer or silencer for neighboring or distal genes via linear chromatin proximity or long-range chromatin interactions (89,93). Overall, we propose that modified STARR-seq vectors, replacing the basic SCP1 promoter with a constitutive strong promoter (such as pPGK), provide an effective strategy to discover and characterize silencers genome-wide.

Several previous works have predicted silencers using indirect approaches based on 3D interactions and epigenetic signatures. A study using Capture Hi-C (CHi-C) technology suggested that transcriptionally inactive genes interacting with previously uncharacterized elements marked by repressive features may act as long-range silencers (25). They further showed that a genomic region located 1.2 Mb from the *BCL6* promoter and associated with Polycomb repressive complex 2 (PRC2) displayed silencer activity when tested in a reporter assay. Similar approaches identified PRC2-bound silencers playing an important role in mouse development (24) or tumor growth (94). Another strategy to identify candidate silencer elements has been to correlate cross-tissue epigenetic profiles (26,95). However, all the above approaches did not directly assess the silencer activity but assumed an association with repressive epigenetic signatures and long-range interaction with the target genes that prevented an unbiased identification of silencer elements. In an attempt to directly identify silencer elements, Pang and Snyder developed a lentiviral screening approach named repressive ability of silencer elements (ReSE), to identify elements capable of repressing the pro-apoptotic protein, Caspase-9 (23). The ReSE approach identified *bona fide* silencers, but the lentivirus strategy is difficult to set up and has limited complexity, while a potential bias might be introduced by the random genomic integration of the lentivirus.

We identified several TFs associated with silencer activities. Of these, REST appeared to be associated with the strongest silencing activity in the pPGK and SCP1 libraries. Moreover, REST-containing silencers were associated with tissue-specific genes that are not normally expressed in T cells. This observation was consistent with the results obtained after the deletion of the REST-containing silencer DHS12366. Our results agree with the widespread role of REST repressor (96). REST was initially described to be involved in the repression of neural genes in non-neuronal cells (65-67). However, the function of the REST TF is not restricted to the repression of neuronal genes and might be involved in the regulation of distinct developmental pathways (97-99). Our results suggested the existence of two main types of silencer elements. One type of silencer contains the REST binding site and appears to have ubiquitous silencer activity. The other type of silencers might have more tissue-restricted silencer activity (*i.e.*, involved in the regulation of genes expressed during cell differentiation). The CTCF binding site was also enriched in our set of identified silencers which is intriguing given the general role of this protein as an insulator factor. This is, however, consistent with a previous study that showed that the T39 region bound by CTCF functions as a strong silencer, but is devoid of insulator activity, thus suggesting that in some cases CTCF might be specifically required for silencer activity (32,88).

Some known silencers have been shown to recruit tissue-specific transcription factors with repressive activities, but the overall set of proteins that collaborate to impart silencer function is essentially unknown. In our study, we identified several TFs involved in T cell differentiation and function for which binding sites were associated with strong silencer activity. These included the hematopoietic-specific RUNX1 transcription factor (100), the HOXA family of developmental repressors (101), as well as the TGFb-signaling dependent SMAD3/4 repressors (102). Members of the Runx family have been shown to repress T cell specific genes by binding to well-characterized silencers such as those found in the CD4 and ThPOK loci (7,8,103-106). We have previously shown that HOXA TFs play an important role during early T cell differentiation and their maintained expression induced T-acute lymphoblastic leukemia (107). SMAD-dependent TGFb signaling also plays an essential role in T cell differentiation and function (108,109). The direct involvement of HOXA and SMAD factor in the activity of T cell regulated silencers will need to be investigated in the future.

We also identified the ZNF263 tandem repeats as a potential silencer element in T cells. ZNF263 is a transcriptional repressor belonging to the family of KZFPs and encoded by the

ubiquitously expressed *Zfp263* gene. ZNF263 has been previously shown to be involved in the epigenetic repression of tumor suppressor genes in glioblastoma (110). KZFPs play a major role in the recognition and transcriptional silencing of REs (111,112). The majority of KZFPs bind to TEs, including LTR, L1, SINE, and SVA families, as well as simple repeats and other variable number tandem repeats, including zinc finger repeats. KZFP tether KRAB associated protein 1 (KAP1, also known as TRIM28) to the DNA. In turn, KAP1 can recruit different epigenetic effectors, such as histone methyltransferases, nucleosome remodeling and deacetylation (NuRD) complex or DNA methyltransferases (113,114). Besides ZNF263, several other KZFP motifs appeared to be enriched in the set of identified silencers (**Supplementary Figure S3**), suggesting that KZFP motifs might be a general feature of silencer elements. Amongst the silencer-enriched KZFP motifs, we found Zfp281, which has been shown to sustain $CD4^+$ T lymphocyte activation by directly repressing the Ctla-4 gene (115). As STRs have been suggested to regulate gene expression in mammalian cells through various molecular mechanisms and to contribute to gene expression variation in humans (34,84,85), it is tempting to speculate that STRs bound by KZFP family of transcriptional repressors have been co-opted in the mammalian genome as repressive *cis*-regulatory modules.

Besides STRs, our results suggest that other REs, namely TEs, might also play a role in silencer activity. TEs are silenced by the targeted deposition of repressive histone modifications (33). For example, in mESCs, TEs are silenced through H3K9 tri-methylation by SET domain bifurcated histone lysine methyltransferase 1 (SETDB1), which is recruited to TEs by KZFPs through interaction with KAP1 (116). SETDB1 knockout leads to widespread de-repression of class I and class II ERV elements and transcription of chimeric RNAs, suggesting that repression of these elements not only prevents mutagenic transposition but also deleterious *cis*-regulatory effects (117). The current model suggests that, rather than a defense system against transposition, the KZFP system may enable the genomic accumulation of TEs with strong *cis*-regulatory elements (such as LTR elements), which increases the likelihood of these elements being subsequently co-opted for host functions (33). Recent examples of TEs co-opted as silencer elements included a SINE element involved in the silencing of a T-cell specific gene (21) and endogenous retrovirus (ERV) involved in the repression of immune genes (82). Similarly, SETDB1-mediated repression of SINE B2 repeats restricts the usage of functional CTCF sites (118). These examples reveal that TE

silencing not only affects TE activity but can also have collateral effects on the regulation of host-gene transcription.

Overall, our study represents an initial step toward the understanding of the molecular basis driving silencer activity, including epigenetic features and binding of transcription factors. We provided experimental evidence that the STARR-seq approach is able to identify silencers functioning in the endogenous context and likely playing a key physiological role in the regulation of T cell differentiation and function. Thus, further implementation of STARR-seq in different cellular systems will help in the functional assessment of mammalian silencers that are active in specific pathways or induced by specific stimuli.

Data availability

The CapSTARR-seq data generated in this study have been submitted to the Gene Expression Omnibus (GEO) under the accession code GSE202547. All other datasets used in this study are listed in Supplemental Table S1.

Acknowledgments

We thank the Transcriptomics and Genomics Marseille-Luminy (TGML) platform for sequencing the CapSTARR-seq samples and the Marseille-Luminy cell biology platform for the management of cell culture. TGML is a member of the France Genomique consortium (ANR-10-INBS-0009). Work in the laboratory of S. Spicuglia was supported by recurrent funding from Institut National de la Santé et de la Recherche Médicale and Aix-Marseille University and by specific grants from Canceropôle PACA, A*MIDEX (ANR-11-IDEX-0001-02), INCA (PLBIO018-031 INCA_12619), Ligue contre le Cancer (Equipe Labellisée 2018), ANR (ANR-18-CE12-0019 and ANR-17-CE12-0035) and Bettencourt Schueller Foundation (Prix coup d'élan pour la recherche française). SH and NS were supported by PhD fellowships from the Pakistani's government and from Marseille Institute of Rare diseases (MARMARA), respectively.

Author contributions

SSp and SSa conceived the project and secured funding. SSp supervised the project. SH, NS, and SSp conceptualized and designed the experiments. SH performed the majority of the experimental work, with the help of LTMD and MT. NS performed the majority of the bioinformatic analyses, with the help of QF and GC. CL contributed to the analyses of

repetitive sequences. TS performed analyses of Hi-C data. DvE and SSa contributed to the design of the CapStarr-seq approach as well as bioinformatics identification of silencers. SH, NS, and SSp wrote the manuscript. All authors contributed to reading, discussion, and commenting on the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Chatterjee, S. and Ahituv, N. (2017) Gene Regulatory Elements, Major Drivers of Human Disease. *Annu Rev Genomics Hum Genet*.
2. Maston, G.A., Evans, S.K. and Green, M.R. (2006) Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, **7**, 29-59.
3. Ogbourne, S. and Antalis, T.M. (1998) Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem J*, **331 (Pt 1)**, 1-14.
4. Baniahmad, A., Muller, M., Steiner, C. and Renkawitz, R. (1987) Activity of two different silencer elements of the chicken lysozyme gene can be compensated by enhancer elements. *Embo j*, **6**, 2297-2303.
5. Brand, A.H., Breeden, L., Abraham, J., Sternglanz, R. and Nasmyth, K. (1985) Characterization of a "silencer" in yeast: a DNA sequence with properties opposite to those of a transcriptional enhancer. *Cell*, **41**, 41-48.
6. Kadesch, T., Zervos, P. and Ruezinsky, D. (1986) Functional analysis of the murine IgH enhancer: evidence for negative control of cell-type specificity. *Nucleic Acids Res*, **14**, 8209-8221.
7. Taniuchi, I., Sunshine, M.J., Festenstein, R. and Littman, D.R. (2002) Evidence for distinct CD4 silencer functions at different stages of thymocyte differentiation. *Mol Cell*, **10**, 1083-1096.
8. Setoguchi, R., Tachibana, M., Naoe, Y., Muroi, S., Akiyama, K., Tezuka, C., Okuda, T. and Taniuchi, I. (2008) Repression of the transcription factor Th-POK by Runx complexes in cytotoxic T cell development. *Science.*, **319**, 822-825.
9. Huang, Z., Liang, N., Goni, S., Damdimopoulos, A., Wang, C., Ballaire, R., Jager, J., Niskanen, H., Han, H., Jakobsson, T. *et al.* (2021) The corepressors GPS2 and SMRT control enhancer and silencer remodeling via eRNA transcription during inflammatory activation of macrophages. *Mol Cell*, **81**, 953-968 e959.
10. Petrykowska, H.M., Vockley, C.M. and Elnitski, L. (2008) Detection and characterization of silencers and enhancer-blockers in the greater CFTR locus. *Genome Res*, **18**, 1238-1246.
11. Liu, Z., Widlak, P., Zou, Y., Xiao, F., Oh, M., Li, S., Chang, M.Y., Shay, J.W. and Garrard, W.T. (2006) A Recombination Silencer that Specifies Heterochromatin Positioning and Ikaros Association in the Immunoglobulin kappa Locus. *Immunity*, **24**, 405-415.
12. Yadav, D.K., Shrestha, S., Dadhwal, G. and Chandak, G.R. (2018) Identification and characterization of cis-regulatory elements 'insulator and repressor' in PPARD gene. *Epigenomics*, **10**, 613-627.
13. Tran, T.H., Nakata, M., Suzuki, K., Begum, N.A., Shinkura, R., Fagarasan, S., Honjo, T. and Nagaoka, H. (2010) B cell-specific and stimulation-responsive enhancers derepress Aicda by overcoming the effects of silencers. *Nat Immunol*, **11**, 148-154.

14. Oh, C.K., Neurath, M., Cho, J.J., Semere, T. and Metcalfe, D.D. (1997) Two different negative regulatory elements control the transcription of T-cell activation gene 3 in activated mast cells. *Biochem J*, **323 (Pt 2)**, 511-519.
15. Williams, T.M., Moolten, D., Burlein, J., Romano, J., Bhaerman, R., Godillot, A., Mellon, M., Rauscher, F.J., 3rd and Kant, J.A. (1991) Identification of a zinc finger protein that inhibits IL-2 gene expression. *Science*, **254**, 1791-1794.
16. Sawada, S., Scarborough, J.D., Killeen, N. and Littman, D.R. (1994) A lineage-specific transcriptional silencer regulates CD4 gene expression during T lymphocyte development. *Cell*, **77**, 917-929.
17. Dombret, H., Font, M.P. and Sigaux, F. (1996) A dominant transcriptional silencer located 5' to the human T-cell receptor V·2.2 gene segment which is activated in cell lines of thymic phenotype. *Nucleic Acids Res.*, **24**, 2782-2789.
18. He, X., Park, K., Wang, H., Zhang, Y., Hua, X., Li, Y. and Kappes, D.J. (2008) CD4-CD8 lineage commitment is regulated by a silencer element at the ThPOK transcription-factor locus. *Immunity*, **28**, 346-358.
19. Yannoutsos, N., Barreto, V., Misulovin, Z., Gazumyan, A., Yu, W., Rajewsky, N., Peixoto, B.R., Eisenreich, T. and Nussenzweig, M.C. (2004) A cis element in the recombination activating gene locus regulates gene expression by counteracting a distant silencer. *Nat Immunol*, **5**, 443-450.
20. Yao, X., Nie, H., Rojas, I.C., Harriss, J.V., Maika, S.D., Gottlieb, P.D., Rathbun, G. and Tucker, P.W. (2010) The L2a element is a mouse CD8 silencer that interacts with MAR-binding proteins SATB1 and CDP. *Mol Immunol*, **48**, 153-163.
21. Hosokawa, H., Koizumi, M., Masuhara, K., Romero-Wolf, M., Tanaka, T., Nakayama, T. and Rothenberg, E.V. (2021) Stage-specific action of Runx1 and GATA3 controls silencing of PU.1 expression in mouse pro-T cells. *J Exp Med*, **218**.
22. Della Rosa, M. and Spivakov, M. (2020) Silencers in the spotlight. *Nat Genet*, **52**, 244-245.
23. Pang, B. and Snyder, M.P. (2020) Systematic identification of silencers in human cells. *Nat Genet*, **52**, 254-263.
24. Ngan, C.Y., Wong, C.H., Tjong, H., Wang, W., Goldfeder, R.L., Choi, C., He, H., Gong, L., Lin, J., Urban, B. *et al.* (2020) Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. *Nat Genet*, **52**, 264-272.
25. Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A. *et al.* (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet*, **47**, 598-606.
26. Doni Jayavelu, N., Jajodia, A., Mishra, A. and Hawkins, R.D. (2020) Candidate silencer elements for the human and mouse genomes. *Nat Commun*, **11**, 1061.
27. Zhang, P., Xia, J.H., Zhu, J., Gao, P., Tian, Y.J., Du, M., Guo, Y.C., Suleman, S., Zhang, Q., Kohli, M. *et al.* (2018) High-throughput screening of prostate cancer risk loci by single nucleotide polymorphisms sequencing. *Nat Commun*, **9**, 2022.
28. Santiago-Algarra, D., Dao, L.T.M., Pradel, L., Espana, A. and Spicuglia, S. (2017) Recent advances in high-throughput approaches to dissect enhancer function. *F1000Res*, **6**, 939.
29. Gasperini, M., Tome, J.M. and Shendure, J. (2020) Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat Rev Genet*, **21**, 292-310.
30. Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M. and Stark, A. (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, **339**, 1074-1077.
31. Vanhille, L., Griffon, A., Maqbool, M.A., Zacarias-Cabeza, J., Dao, L.T., Fernandez, N., Ballester, B., Andrau, J.C. and Spicuglia, S. (2015) High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nat Commun*, **6**, 6905.
32. Qi, H., Liu, M., Emery, D.W. and Stamatoyannopoulos, G. (2015) Functional validation of a constitutive autonomous silencer element. *PLoS One*, **10**, e0124588.
33. Fueyo, R., Judd, J., Feschotte, C. and Wysocka, J. (2022) Roles of transposable elements in the regulation of mammalian transcription. *Nat Rev Mol Cell Biol*.

34. Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., Daly, M.J., Price, A.L., Pritchard, J.K., Sharp, A.J. *et al.* (2016) Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat Genet*, **48**, 22-29.
35. Mombaerts, P., Terhorst, C., Jacks, T., Tonegawa, S. and Sancho, J. (1995) Characterization of immature thymocyte lines derived from T-cell receptor or recombination activating gene 1 and p53 double mutant mice. *Proc Natl Acad Sci U S A*, **92**, 7420-7424.
36. Saadi, W., Kermezli, Y., Dao, L.T.M., Mathieu, E., Santiago-Algarra, D., Manosalva, I., Torres, M., Belhocine, M., Pradel, L., Loriod, B. *et al.* (2019) A critical regulator of Bcl2 revealed by systematic transcript discovery of lncRNAs associated with T-cell differentiation. *Scientific Reports*, **9**, 4707.
37. Qin, J.Y., Zhang, L., Clift, K.L., Hulur, I., Xiang, A.P., Ren, B.Z. and Lahn, B.T. (2010) Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. *PloS one*, **5**, e10611.
38. Wei, X.C., Dohkan, J., Kishi, H., Wu, C.X., Kondo, S. and Muraguchi, A. (2005) Characterization of the proximal enhancer element and transcriptional regulatory factors for murine recombination activating gene-2. *Eur J Immunol*, **35**, 612-621.
39. Dao, L.T.M., Vanhille, L., Griffon, A., Fernandez, N. and Spicuglia, S. (2015) CapStarr-seq protocol. *Protocol Exchange*.
40. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.
41. 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.
42. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, **44**, D733-745.
43. Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S. and Karolchik, D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204-2207.
44. Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A. and Manke, T. (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research*, **42**, W187-191.
45. Hu, G., Tang, Q., Sharma, S., Yu, F., Escobar, T.M., Muljo, S.A., Zhu, J. and Zhao, K. (2013) Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. *Nat Immunol*, **14**, 1190-1198.
46. Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, **101**, 6062-6067.
47. 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.
48. Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S.R., Tan, G. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res*, **46**, D260-D266.
49. Castro-Mondragon, J.A., Jaeger, S., Thieffry, D., Thomas-Chollier, M. and van Helden, J. (2017) RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res*, **45**, e119.
50. Leporcq, C., Spill, Y., Balaramane, D., Toussaint, C., Weber, M. and Bardet, A.F. (2020) TFmotifView: a webserver for the visualization of transcription factor motifs in genomic regions. *Nucleic Acids Res*, **48**, W208-W217.

51. Ferré, Q., Charbonnier, G., Sadouni, N., Lopez, F., Kermezli, Y., Spicuglia, S., Capponi, C., Ghattas, B. and Puthier, D. (2019) OLOGRAM: Determining significance of total overlap length between genomic regions sets. *Bioinformatics*.
52. Lopez, F., Charbonnier, G., Kermezli, Y., Belhocine, M., Ferre, Q., Zweig, N., Aribi, M., Gonzalez, A., Spicuglia, S. and Puthier, D. (2019) Explore, edit and leverage genomic annotations using Python GTF toolkit. *Bioinformatics*.
53. Chen, N. (2004) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*, **Chapter 4**, Unit 4.10.
54. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, **110**, 462-467.
55. Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M. and Erlich, Y. (2017) Genome-wide profiling of heritable and de novo STR variations. *Nat Methods*, **14**, 590-592.
56. Yu, G. (2020) Gene Ontology Semantic Similarity Analysis Using GOSemSim. *Methods Mol Biol*, **2117**, 207-215.
57. Hu, G., Cui, K., Fang, D., Hirose, S., Wang, X., Wangsa, D., Jin, W., Ried, T., Liu, P., Zhu, J. *et al.* (2018) Transformation of Accessible Chromatin and 3D Nucleome Underlies Lineage Commitment of Early T Cells. *Immunity*, **48**, 227-242 e228.
58. Kruse, K., Hug, C.B. and Vaquerizas, J.M. (2020) FAN-C: A Feature-rich Framework for the Analysis and Visualisation of C data. *bioRxiv*, 2020.2002.2003.932517.
59. Crane, E., Bian, Q., McCord, R.P., Lajoie, B.R., Wheeler, B.S., Ralston, E.J., Uzawa, S., Dekker, J. and Meyer, B.J. (2015) Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, **523**, 240-244.
60. Naito, Y., Hino, K., Bono, H. and Ui-Tei, K. (2015) CRISPRdirect: software for designing CRISPR/Cas guide RNA with reduced off-target sites. *Bioinformatics*, **31**, 1120-1123.
61. Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E. and Church, G.M. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823-826.
62. Donda, A., Schulz, M., Burki, K., De Libero, G. and Uematsu, Y. (1996) Identification and characterization of a human CD4 silencer. *Eur J Immunol*, **26**, 493-500.
63. Maqbool, M.A., Pioger, L., El Aabidine, A.Z., Karasu, N., Molitor, A.M., Dao, L.T.M., Charbonnier, G., van Laethem, F., Fenouil, R., Koch, F. *et al.* (2020) Alternative Enhancer Usage and Targeted Polycomb Marking Hallmark Promoter Choice during T Cell Differentiation. *Cell Rep*, **32**, 108048.
64. Zabidi, M.A., Arnold, C.D., Schernhuber, K., Pagani, M., Rath, M., Frank, O. and Stark, A. (2015) Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, **518**, 556-559.
65. Schoenherr, C.J. and Anderson, D.J. (1995) The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science*, **267**, 1360-1363.
66. Chong, J.A., Tapia-Ram°rez, J., Kim, S., Toledo-Aral, J.J., Zheng, Y., Boutros, M.C., Altshuller, Y.M., Frohman, M.A., Kraner, S.D. and Mandel, G. (1995) REST: A mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell*, **80**, 949-957.
67. Coulson, J.M. (2005) Transcriptional regulation: cancer, neurons and the REST. *Curr Biol*, **15**, R665-668.
68. Wu, J., Song, Y., Bakker, A.B., Bauer, S., Spies, T., Lanier, L.L. and Phillips, J.H. (1999) An activating immunoreceptor complex formed by NKG2D and DAP10. *Science*, **285**, 730-732.
69. Diefenbach, A., Tomasello, E., Lucas, M., Jamieson, A.M., Hsia, J.K., Vivier, E. and Raulet, D.H. (2002) Selective associations with signaling proteins determine stimulatory versus costimulatory activity of NKG2D. *Nat Immunol*, **3**, 1142-1149.
70. Schuster, M., Annemann, M., Plaza-Sirvent, C. and Schmitz, I. (2013) Atypical IkappaB proteins - nuclear modulators of NF-kappaB signaling. *Cell Commun Signal*, **11**, 23.
71. Schuster, M., Glauben, R., Plaza-Sirvent, C., Schreiber, L., Annemann, M., Floess, S., Kuhl, A.A., Clayton, L.K., Sparwasser, T., Schulze-Osthoff, K. *et al.* (2012) IkappaB(NS) protein

mediates regulatory T cell development via induction of the Foxp3 transcription factor. *Immunity*, **37**, 998-1008.
72. Ebihara, T. and Taniuchi, I. (2019) Transcription Factors in the Development and Function of Group 2 Innate Lymphoid Cells. *Int J Mol Sci*, **20**.
73. Anguita, E., Candel, F.J., Chaparro, A. and Roldan-Etcheverry, J.J. (2017) Transcription Factor GFI1B in Health and Disease. *Front Oncol*, **7**, 54.
74. Chiang, C. and Ayyanathan, K. (2013) Snail/Gfi-1 (SNAG) family zinc finger proteins in transcription regulation, chromatin dynamics, cell signaling, development, and disease. *Cytokine Growth Factor Rev*, **24**, 123-131.
75. Moroy, T. and Khandanpour, C. (2011) Growth factor independence 1 (Gfi1) as a regulator of lymphocyte development and activation. *Semin Immunol*, **23**, 368-378.
76. Frietze, S., Lan, X., Jin, V.X. and Farnham, P.J. (2010) Genomic targets of the KRAB and SCAN domain-containing zinc finger protein 263. *J Biol Chem*, **285**, 1393-1403.
77. Balestrieri, C., Alfarano, G., Milan, M., Tosi, V., Prosperini, E., Nicoli, P., Palamidessi, A., Scita, G., Diaferia, G.R. and Natoli, G. (2018) Co-optation of Tandem DNA Repeats for the Maintenance of Mesenchymal Identity. *Cell*, **173**, 1150-1164.e1114.
78. Barakat, T.S., Halbritter, F., Zhang, M., Rendeiro, A.F., Perenthaler, E., Bock, C. and Chambers, I. (2018) Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell*, **23**, 276-288 e278.
79. Glinsky, G.V. (2015) Transposable Elements and DNA Methylation Create in Embryonic Stem Cells Human-Specific Regulatory Sequences Associated with Distal Enhancers and Noncoding RNAs. *Genome Biol Evol*, **7**, 1432-1454.
80. Kunarso, G., Chia, N.Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.S., Ng, H.H. and Bourque, G. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet*, **42**, 631-634.
81. Wang, X., He, L., Goggin, S.M., Saadat, A., Wang, L., Sinnott-Armstrong, N., Claussnitzer, M. and Kellis, M. (2018) High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat Commun*, **9**, 5380.
82. Adoue, V., Binet, B., Malbec, A., Fourquet, J., Romagnoli, P., van Meerwijk, J.P.M., Amigorena, S. and Joffre, O.P. (2019) The Histone Methyltransferase SETDB1 Controls T Helper Cell Lineage Integrity by Repressing Endogenous Retroviruses. *Immunity*, **50**, 629-644.e628.
83. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res*, **12**, 996-1006.
84. Bagshaw, A.T.M. (2017) Functional Mechanisms of Microsatellite DNA in Eukaryotic Genomes. *Genome Biol Evol*, **9**, 2428-2443.
85. Grapotte, M., Saraswat, M., Bessiere, C., Menichelli, C., Ramilowski, J.A., Severin, J., Hayashizaki, Y., Itoh, M., Tagami, M., Murata, M. *et al.* (2021) Discovery of widespread transcription initiation at microsatellites predictable by sequence-based deep neural network. *Nat Commun*, **12**, 3297.
86. van Arensbergen, J., FitzPatrick, V.D., de Haas, M., Pagie, L., Sluimer, J., Bussemaker, H.J. and van Steensel, B. (2017) Genome-wide mapping of autonomous promoter activity in human cells. *Nat Biotechnol*, **35**, 145-153.
87. Weingarten-Gabbay, S., Nir, R., Lubliner, S., Sharon, E., Kalma, Y., Weinberger, A. and Segal, E. (2019) Systematic interrogation of human promoters. *Genome Res*, **29**, 171-183.
88. Liu, M., Maurano, M.T., Wang, H., Qi, H., Song, C.Z., Navas, P.A., Emery, D.W., Stamatoyannopoulos, J.A. and Stamatoyannopoulos, G. (2015) Genomic discovery of potent chromatin insulators for human gene therapy. *Nat Biotechnol*, **33**, 198-203.
89. Wei, X., Xiang, Y., Peters, D.T., Marius, C., Sun, T., Shan, R., Ou, J., Lin, X., Yue, F., Li, W. *et al.* (2022) HiCAR is a robust and sensitive method to analyze open-chromatin-associated genome organization. *Mol Cell*, **82**, 1225-1238 e1226.

90. Dao, L.T.M., Galindo-Albarran, A.O., Castro-Mondragon, J.A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., Charbonnier, G., Griffon, A., Vanhille, L., Stephen, T. *et al.* (2017) Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat Genet*, **49**, 1073-1081.
91. Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., Lin, K.C., Huang, H., Liu, T., Marina, R.J. *et al.* (2017) A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat Methods*, **14**, 629-635.
92. Engreitz, J.M., Haines, J.E., Perez, E.M., Munson, G., Chen, J., Kane, M., McDonel, P.E., Guttman, M. and Lander, E.S. (2016) Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*, **539**, 452-455.
93. Andersson, R. (2015) Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *Bioessays*, **37**, 314-323.
94. Cai, Y., Zhang, Y., Loh, Y.P., Tng, J.Q., Lim, M.C., Cao, Z., Raju, A., Lieberman Aiden, E., Li, S., Manikandan, L. *et al.* (2021) H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. *Nat Commun*, **12**, 719.
95. Huang, D., Petrykowska, H.M., Miller, B.F., Elnitski, L. and Ovcharenko, I. (2019) Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. *Genome Res*, **29**, 657-667.
96. Ooi, L. and Wood, I.C. (2007) Chromatin crosstalk in development and disease: lessons from REST. *Nat Rev Genet*, **8**, 544-554.
97. Kuwahara, K., Saito, Y., Takano, M., Arai, Y., Yasuno, S., Nakagawa, Y., Takahashi, N., Adachi, Y., Takemura, G., Horie, M. *et al.* (2003) NRSF regulates the fetal cardiac gene program and maintains normal cardiac structure and function. *Embo j*, **22**, 6310-6321.
98. Martin, D., Tawadros, T., Meylan, L., Abderrahmani, A., Condorelli, D.F., Waeber, G. and Haefliger, J.A. (2003) Critical role of the transcriptional repressor neuron-restrictive silencer factor in the specific control of connexin36 in insulin-producing cell lines. *J Biol Chem*, **278**, 53082-53089.
99. Rockowitz, S., Lien, W.H., Pedrosa, E., Wei, G., Lin, M., Zhao, K., Lachman, H.M., Fuchs, E. and Zheng, D. (2014) Comparison of REST cistromes across human cell types reveals common and context-specific functions. *PLoS Comput Biol*, **10**, e1003671.
100. Collins, A., Littman, D.R. and Taniuchi, I. (2009) RUNX proteins in transcription factor networks that regulate T-cell lineage choice. *Nat Rev Immunol*, **9**, 106-115.
101. Cain, B. and Gebelein, B. (2021) Mechanisms Underlying Hox-Mediated Transcriptional Outcomes. *Front Cell Dev Biol*, **9**, 787339.
102. Hill, C.S. (2016) Transcriptional Control by the SMADs. *Cold Spring Harb Perspect Biol*, **8**.
103. Jiang, H. and Peterlin, B.M. (2008) Differential chromatin looping regulates CD4 expression in immature thymocytes. *Mol Cell Biol*, **28**, 907-912.
104. Wildt, K.F., Sun, G., Grueter, B., Fischer, M., Zamisch, M., Ehlers, M. and Bosselut, R. (2007) The transcription factor Zbtb7b promotes CD4 expression by antagonizing Runx-mediated activation of the CD4 silencer. *J Immunol*, **179**, 4405-4414.
105. Jiang, H., Zhang, F., Kurosu, T. and Peterlin, B.M. (2005) Runx1 Binds Positive Transcription Elongation Factor b and Represses Transcriptional Elongation by RNA Polymerase II: Possible Mechanism of CD4 Silencing. *Mol Cell Biol*, **25**, 10675-10683.
106. Telfer, J.C., Hedblom, E.E., Anderson, M.K., Laurent, M.N. and Rothenberg, E.V. (2004) Localization of the domains in runx transcription factors required for the repression of CD4 in thymocytes. *J Immunol*, **172**, 4359-4370.
107. Cieslak, A., Charbonnier, G., Tesio, M., Mathieu, E.L., Belhocine, M., Touzart, A., Smith, C., Hypolite, G., Andrieu, G.P., Martens, J.H.A. *et al.* (2020) Blueprint of human thymopoiesis reveals molecular mechanisms of stage-specific TCR enhancer activation. *J Exp Med*, **217**, (9): e20192360.

108. Gu, A.D., Wang, Y., Lin, L., Zhang, S.S. and Wan, Y.Y. (2012) Requirements of transcription factor Smad-dependent and -independent TGF-β signaling to control discrete T-cell functions. *Proc Natl Acad Sci U S A*, **109**, 905-910.
109. Takimoto, T., Wakabayashi, Y., Sekiya, T., Inoue, N., Morita, R., Ichiyama, K., Takahashi, R., Asakawa, M., Muto, G., Mori, T. *et al.* (2010) Smad2 and Smad3 are redundantly essential for the TGF-beta-mediated regulation of regulatory T plasticity and Th1 development. *J Immunol*, **185**, 842-855.
110. Yu, Z., Feng, J., Wang, W., Deng, Z., Zhang, Y., Xiao, L., Wang, Z., Liu, C., Liu, Q., Chen, S. *et al.* (2020) The EGFR-ZNF263 signaling axis silences SIX3 in glioblastoma epigenetically. *Oncogene*, **39**, 3163-3178.
111. Yang, P., Wang, Y. and Macfarlan, T.S. (2017) The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. *Trends Genet*, **33**, 871-881.
112. Ecco, G., Imbeault, M. and Trono, D. (2017) KRAB zinc finger proteins. *Development*, **144**, 2719-2729.
113. Helleboid, P.Y., Heusel, M., Duc, J., Piot, C., Thorball, C.W., Coluccio, A., Pontis, J., Imbeault, M., Turelli, P., Aebersold, R. *et al.* (2019) The interactome of KRAB zinc finger proteins reveals the evolutionary history of their functional diversification. *EMBO J*, **38**, e101220.
114. Jang, S.M., Kauzlaric, A., Quivy, J.P., Pontis, J., Rauwel, B., Coluccio, A., Offner, S., Duc, J., Turelli, P., Almouzni, G. *et al.* (2018) KAP1 facilitates reinstatement of heterochromatin after DNA replication. *Nucleic Acids Res*, **46**, 8788-8802.
115. Guo, J., Xue, Z., Ma, R., Yi, W., Hui, Z., Guo, Y., Yao, Y., Cao, W., Wang, J., Ju, Z. *et al.* (2020) The transcription factor Zfp281 sustains CD4. *Cell Mol Immunol*, **17**, 1222-1232.
116. Imbeault, M., Helleboid, P.Y. and Trono, D. (2017) KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, **543**, 550-554.
117. Karimi, M.M., Goyal, P., Maksakova, I.A., Bilenky, M., Leung, D., Tang, J.X., Shinkai, Y., Mager, D.L., Jones, S., Hirst, M. *et al.* (2011) DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell Stem Cell*, **8**, 676-687.
118. Gualdrini, F., Polletti, S., Simonatto, M., Prosperini, E., Pileri, F. and Natoli, G. (2022) H3K9 trimethylation in active chromatin restricts the usage of functional CTCF sites in SINE B2 repeats. *Genes Dev*.

Legends

Figure 1. CapSTARR-seq for silencer assessment

(A) Schematic of the CapSTARR-seq strategy to assess the silencer activity in the P5424 mouse T cell line.

(B) Distribution of CapSTARR-seq signal ($\log_2$) in the different conditions for DHS and random captured regions. The threshold for putative silencers ($\log_2$ (FC) $\leq$ -1) and enhancers (log2(FC) $\geq$ 1) is indicated. Statistical analysis was performed using Wilcoxon test, *P*-values are displayed (ns: not significant).

(**C**) The UCSC genomic track of Mouse NCBI37/mm9 around DHS19266 showing the $\log_2$ STARR-seq signal of individual clones, the captured DHS region, the core silencer and the silencer score of the region.

(D) Bar plot showing the number of unique and shared silencer candidates identified with the three different CapSTARR-seq libraries.

(**E**) Venn diagram displaying the overlap between the silencer candidates identified with the three different CapSTARR-seq libraries.

(F) Example of a putative silencer identified with the three promoter-based CapSTARR-seq strategies. The signal for each CapSTARR-seq experiment and the corresponding Input are displayed.

(G) Bar plot displaying the proportion of silencer candidates that are proximal (< 1 kB) or distal (> 1 kb) to the closer TSS.

(H) Genomic distribution of silencer candidates compare to the whole set of DHS. Bar plots represent the -$\log_{10}$(*P*-value) of negative binomial test computed by OLOGRAM. Depletion is represented by negative values.

Figure 2. Validation of the CapSTARR-seq approach.

(A) Luciferase reporter assays in P5424 cells of DHSs defined as putative silencers by CapSTARR-seq (green) or with $\log_2$ STARR-seq signal close to zero (black). The promoter-based CapSTARR-seq where the silencer was identified is indicated in the bottom panel. Data represent the normalized fold change over the pSV40 vector control. Error bars show s.d. from three independent transfections (***$P < 0.001$, **$P < 0.01$, *$P < 0.05$; two-sided Student's *t* test).

(B) Comparison of luciferase activity between silencer candidates and control regions. The two-sided Student's *t* test is shown.

(C) Assessment of orientation dependent silencer activity for a subset of identified silencers (F: forward; R: reverse).

(D) FACS analysis for GFP expression assessing DHS12366 silencer activity.

Figure 3. Chromatin features and genes associated with silencers

(A) Average profiles of H3K27ac and H3K4me3 ChIP-seq signal from P5424 cells centered on putative silencers (red), control regions (black) and putative enhancers (green). **(B)** Violin plot comparing the expression in P5424 cells of genes associated with putative silencers (red), control regions (black) and putative enhancers (green). Statistical analysis was performed using Wilcoxon test, *P*-values are displayed.

(C-D) Heatmap of top five GO terms analysis for mouse phenotype (**C**) and MSigDB Pathway **(D)** enriched in genes associated with silencers or SCP1 enhancers.

Figure 4. TFBS associated with silencers and site directed mutagenesis.

(**A**) Heatmap displaying the enrichment score of the top 10 clustered TF motifs enriched in each of the silencer sets, as well as, in the SCP1 enhancers.

(**B**) Dot plots displaying the mean activity of silencers carrying a given TFBS against the enrichment score for the same TFBS. Only significantly enriched TFBS are displayed. TFs of interest are highlighted.

(C-D) Validation of REST (C) and SMAD3 (D) binding sites impact on silencer activity. Left panels display the mutated nucleotides. Right panels display the luciferase reporter assay in wild-type and mutated silencers (***$P < 0.001$, **$P < 0.01$, *$P < 0.05$; two-sided Student's t test).

Figure 5. Functional validation of DHS12366 silencer by CRISPR–Cas9 system.

(**A**) Percentage of silencers harboring REST binding sites.

(**B**) Expression level of genes associated with silencers containing or not REST binding sites. Significance was assessed by Wilcoxon test.

(**C-D**) Proportion of genes associated with all DHS, all silencers or silencers containing or not REST binding sites that are tissue-specifics excluding T-cells (C) or T-cell regulated genes (D).

(E) Top panel: Hi-C data and TADs in DP thymocytes surrounding the DHS12366 silencer. Bottom panel: Genomic tracks displaying the indicated RNA-seq and ChIP-seq signals as well as DHS and TADs from DP thymocytes surrounding the DHS12366 silencer.

(F) Gene expression analysis of genes around the DHS12366 locus in wild-type and ΔDHS12366 R1 and R2 P5424 clones. Error bars, s.d.: ***P < 0.001, **P < 0.01, *P < 0.05, two-sided Student's t test.

(G) Heatmap displaying the relative gene expression of genes around DHS12366 locus in T cell populations and other cell types.

Figure 6. Functional validation of DHS23650 silencer by CRISPR/Cas9 system

(A) Genes ranked in function of their expression variance across T-cell differentiation. The top 5% of variable genes are highlighted in blue. The *Nfkbid* and *Hcst* genes associated with DHS23650 silencer are also shown.

(B) All putative silencers were associated with genes located in a window of 100 kB upstream and downstream. Of these, 516 genes (corresponding to 615 silencers) were part of the top 5% variable genes during T-cell differentiation.

(C) Relative expression of the *Nfkbid* and *Hcst* genes during T-cell differentiation.

(**D**) Top panel: Hi-C data and TADs in DP thymocytes surrounding the DHS12366 silencer. Bottom panel: Genomic tracks displaying the indicated RNA-seq and ChIP-seq signals in P5424 cells stimulated or not with PMA and Ionomycin (36).

(**E**) Luciferase reporter assays testing the silencer activity of DHS23650 in both orientations. Data represent the normalized fold change over the pSV40 vector. Error bars show s.d. from three independent replicates (***P < 0.001, **P < 0.01, *P < 0.05; two-sided Student's t test).

(**F**) Gene expression analysis of genes around the DHS23650 locus in wild-type, ΔDHS23650 R1 and ΔDHS23650 R2 p5424 clones.

(**G**) Gene expression analysis of *Nfkbid* in wild-type, ΔDHS23650 R1 and ΔDHS23650 R2 p5424 clones stimulated or not with PMA and Ionomycin. Error bars, s.d.: ***P < 0.001, **P < 0.01, *P < 0.05 two-sided Student's t test.

Figure 7. Analysis and validation of Simple Tandem Repeats

(A) The UCSC genomic track of Mouse NCBI37/mm9 around the DHS23650 silencer displaying the individual clonal activity, the DHS region (silencer), the core silencer, the RepeatMasker track and the ZNF263 binding sites.

(B) Luciferase reporter assay of wild-type and mutated DHS23650 silencer. The impact of ZNF263 repeats on silencer activity was assessed by either deleting the ZNF263 tandem repeat region or by mutating the indicated number of ZNF263 binding sites. Data represent the normalized fold change over the pSV40 vector. Error bars show s.d. from three independent transfections (***P < 0.001, **P < 0.01, *P < 0.05; two-sided Student's t test).

(C) Empirical cumulative density of ZNF263 binding sites per 100 bp in all DHS, pPGK core silencers or DHS regions flanking the core silencers. Significance was assessed by a KS test.

(D) STARR-seq signal of pPGK silencers in function of the number of ZNF263 binding sites. Significance was assessed by a Wilcoxon test.

(E) Enrichment of repetitive elements at putative silencers. Bar plots represent the $-\log_{10}$(p-value) of negative binomial test computed by OLOGRAM.

(F) Proportion of STRs in pPGK core silencers that overlap with ZNF263 binding sites.

(G) *De novo* motif discovered in STRs within pPGK core silencers using the HOMER tool. Best matched TFBS are indicated.

(H) STARR-seq signal of silencers in function of the number of repeat units within the STRs. Significance was assessed by a Wilcoxon test.

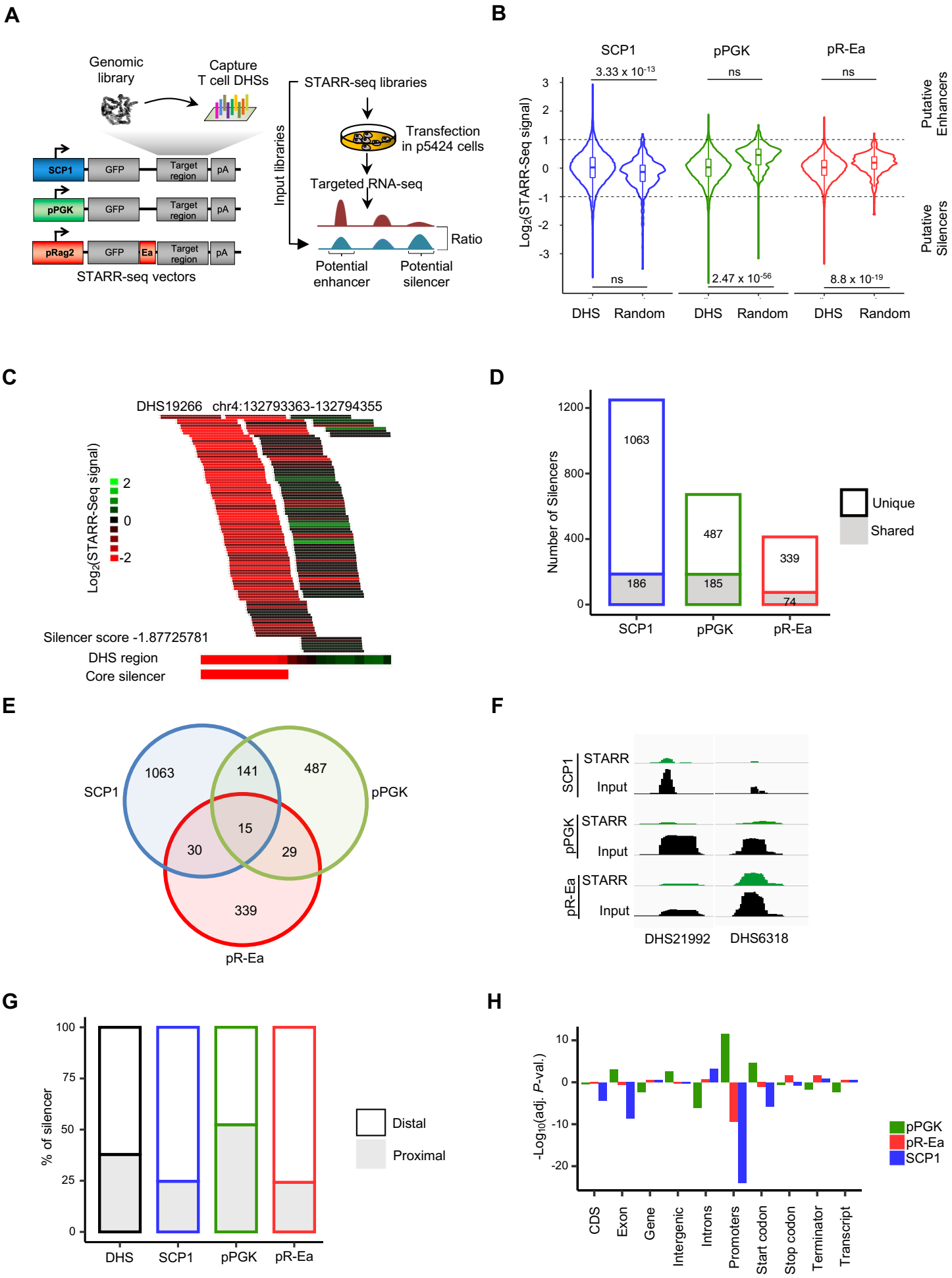

A
Genomic library
Capture T cell DHSs
STARR-seq libraries
Transfection in p5424 cells
Targeted RNA-seq
Ratio
Potential enhancer
Potential silencer
Input libraries
SCP1
pPGK
pRag2
GFP
Ea
Target region
pA
STARR-seq vectors
B
SCP1
pPGK
pR-Ea
3.33 x 10-13
ns
ns
ns
2.47 x 10-56
8.8 x 10-19
Log2(STARR-Seq signal)
Putative Enhancers
Putative Silencers
DHS
Random
C
DHS19266 chr4:132793363-132794355
Log2(STARR-Seq signal)
Silencer score -1.87725781
DHS region
Core silencer
D
Number of Silencers
1063
186
487
185
339
74
Unique
Shared
SCP1
pPGK
pR-Ea
E
SCP1
pPGK
pR-Ea
1063
141
487
15
30
29
339
F
STARR
Input
DHS21992
DHS6318
G
% of silencer
Distal
Proximal
DHS
SCP1
pPGK
pR-Ea
H
-Log10(adj. P-val.)
pPGK
pR-Ea
SCP1
CDS
Exon
Gene
Intergenic
Introns
Promoters
Start codon
Stop codon
Terminator
Transcript

Figure 1

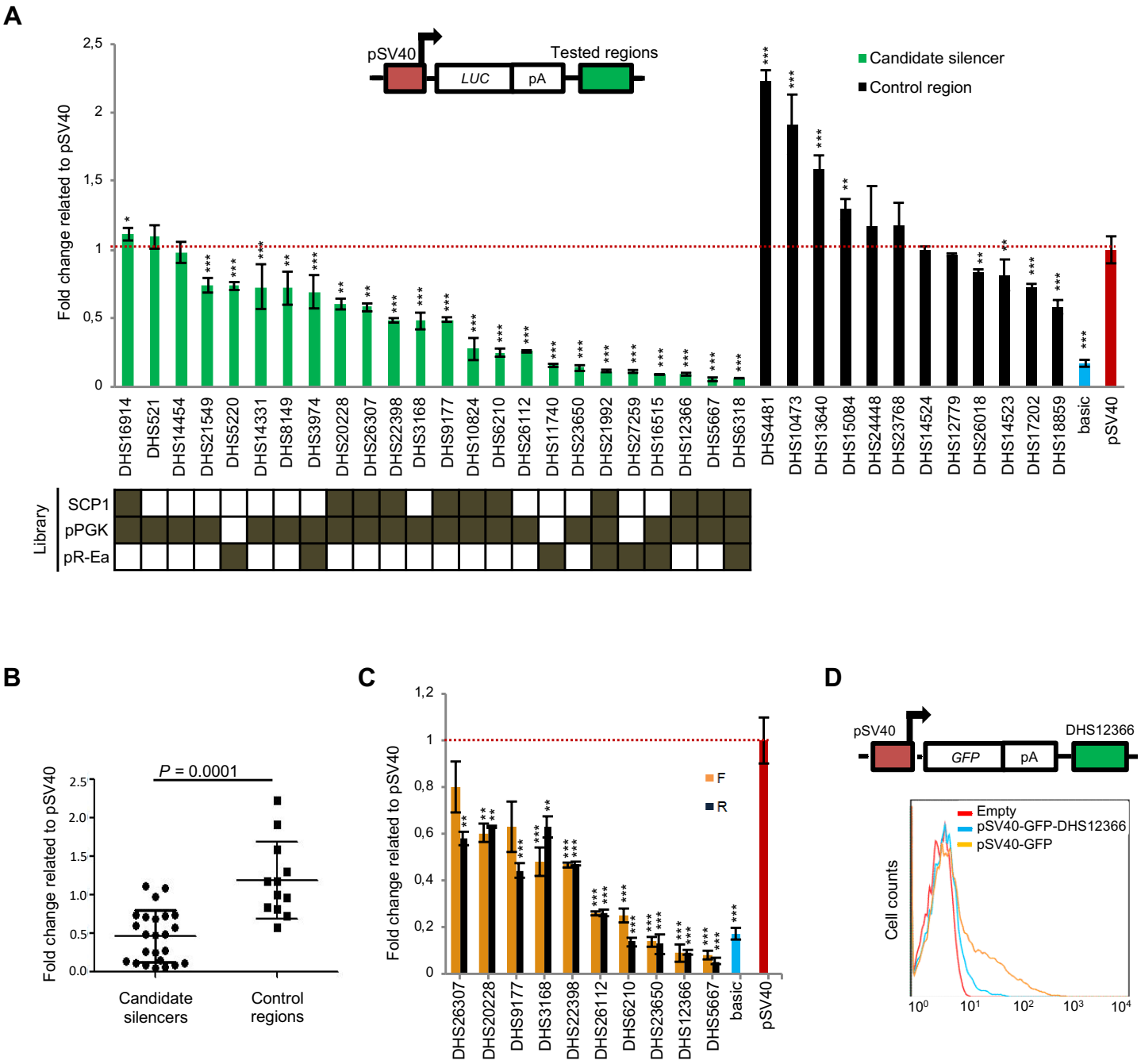
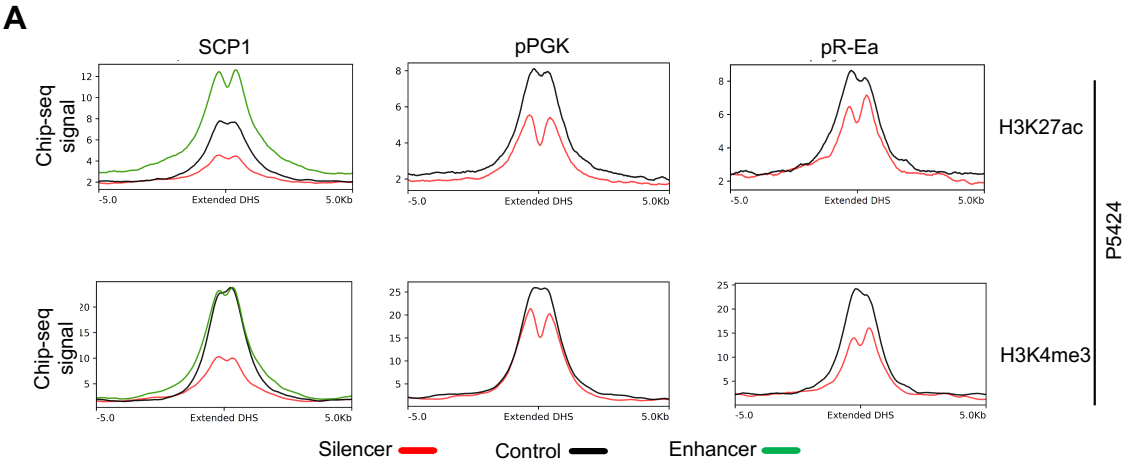
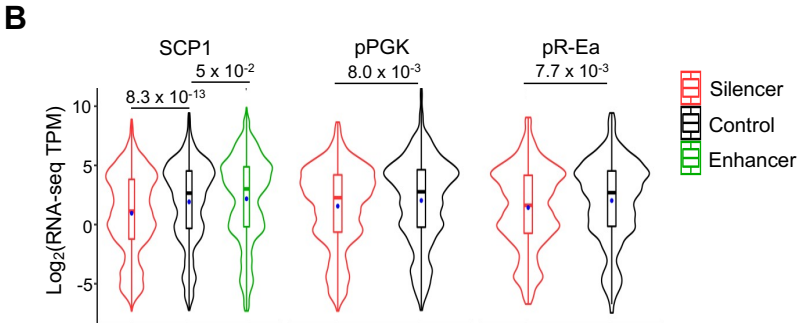
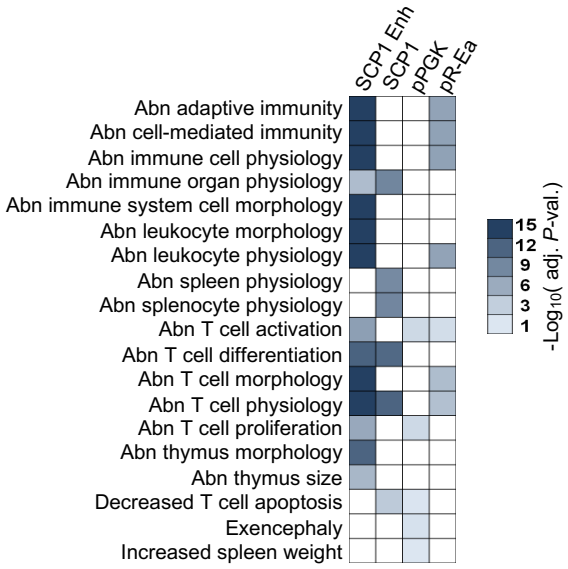

A
pSV40
LUC
pA
Tested regions
Candidate silencer
Control region
Fold change related to pSV40
2,5
2
1,5
1
0,5
0
DHS16914
DHS521
DHS14454
DHS21549
DHS5220
DHS14331
DHS8149
DHS3974
DHS20228
DHS26307
DHS22398
DHS3168
DHS9177
DHS10824
DHS6210
DHS26112
DHS11740
DHS23650
DHS21992
DHS27259
DHS16515
DHS12366
DHS5667
DHS6318
DHS4481
DHS10473
DHS13640
DHS15084
DHS24448
DHS23768
DHS14524
DHS12779
DHS26018
DHS14523
DHS17202
DHS18859
basic
pSV40
Library
SCP1
pPGK
pR-Ea
B
P = 0.0001
Fold change related to pSV40
2.5
2.0
1.5
1.0
0.5
0.0
Candidate silencers
Control regions
C
Fold change related to pSV40
1,2
1
0,8
0,6
0,4
0,2
0
F
R
DHS26307
DHS20228
DHS9177
DHS3168
DHS22398
DHS26112
DHS6210
DHS23650
DHS12366
DHS5667
basic
pSV40
D
pSV40
GFP
pA
DHS12366
Empty
pSV40-GFP-DHS12366
pSV40-GFP
Cell counts
10^0
10^1
10^2
10^3
10^4

Figure 2


A
SCP1
pPGK
pR-Ea
Chip-seq signal
Extended DHS
-5.0
5.0Kb
H3K27ac
H3K4me3
P5424
Silencer
Control
Enhancer


B
SCP1
pPGK
pR-Ea
5 x 10-2
8.3 x 10-13
8.0 x 10-3
7.7 x 10-3
Log2(RNA-seq TPM)
Silencer
Control
Enhancer


C
SCP1 Enh
SCP1
pPGK
pR-Ea
Abn adaptive immunity
Abn cell-mediated immunity
Abn immune cell physiology
Abn immune organ physiology
Abn immune system cell morphology
Abn leukocyte morphology
Abn leukocyte physiology
Abn spleen physiology
Abn splenocyte physiology
Abn T cell activation
Abn T cell differentiation
Abn T cell morphology
Abn T cell physiology
Abn T cell proliferation
Abn thymus morphology
Abn thymus size
Decreased T cell apoptosis
Exencephaly
Increased spleen weight
-Log10(adj. P-val.)

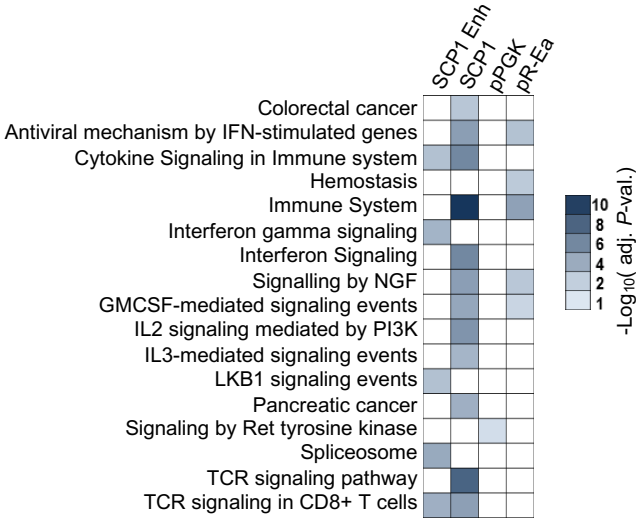

D
SCP1 Enh
SCP1
pPGK
pR-Ea
Colorectal cancer
Antiviral mechanism by IFN-stimulated genes
Cytokine Signaling in Immune system
Hemostasis
Immune System
Interferon gamma signaling
Interferon Signaling
Signalling by NGF
GMCSF-mediated signaling events
IL2 signaling mediated by PI3K
IL3-mediated signaling events
LKB1 signaling events
Pancreatic cancer
Signaling by Ret tyrosine kinase
Spliceosome
TCR signaling pathway
TCR signaling in CD8+ T cells
-Log10(adj. P-val.)

Figure 3

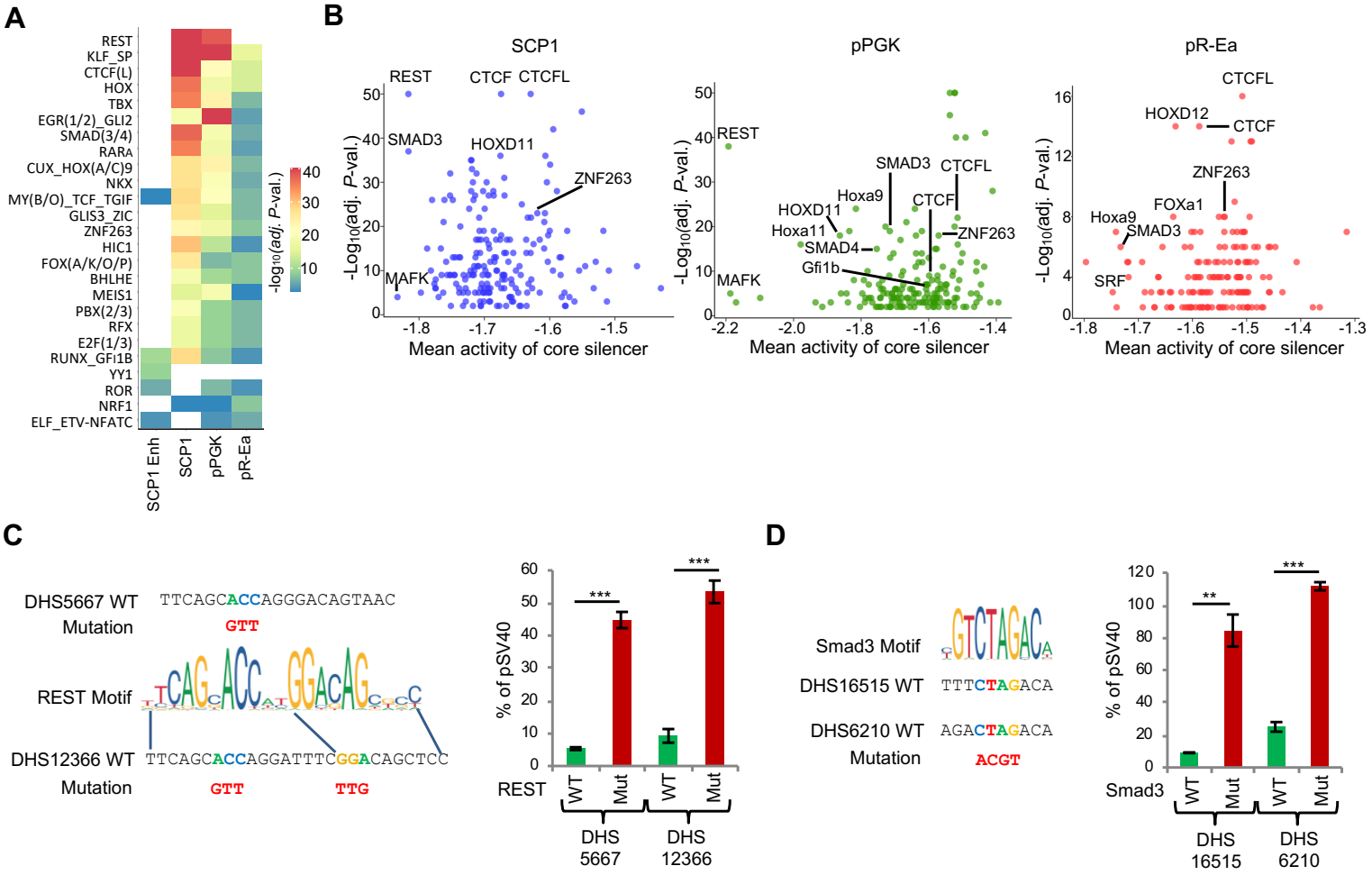

A
REST
KLF_SP
CTCF(L)
HOX
TBX
EGR(1/2)_GLI2
SMAD(3/4)
RARa
CUX_HOX(A/C)9
NKX
MY(B/O)_TCF_TGIF
GLIS3_ZIC
ZNF263
HIC1
FOX(A/K/O/P)
BHLHE
MEIS1
PBX(2/3)
RFX
E2F(1/3)
RUNX_GFI1B
YY1
ROR
NRF1
ELF_ETV-NFATC
SCP1 Enh
SCP1
pPGK
pR-Ea
-log10(adj. P-val.)
B
SCP1
REST
CTCF CTCFL
SMAD3
HOXD11
ZNF263
MAFK
-Log10(adj. P-val.)
Mean activity of core silencer
pPGK
REST
SMAD3 CTCFL
Hoxa9
CTCF
HOXD11
Hoxa11
ZNF263
SMAD4
Gfi1b
MAFK
pR-Ea
CTCFL
HOXD12
CTCF
ZNF263
FOXa1
Hoxa9
SMAD3
SRF
C
DHS5667 WT
TTCAGCACCAGGGACAGTAAC
Mutation
GTT
REST Motif
DHS12366 WT
TTCAGCACCAGGATTTCGGACAGCTCC
Mutation
GTT
TTG
% of pSV40
REST
WT
Mut
DHS 5667
DHS 12366
D
Smad3 Motif
DHS16515 WT
TTTCTAGACA
DHS6210 WT
AGACTAGACA
Mutation
ACGT
Smad3
DHS 16515
DHS 6210

Figure 4

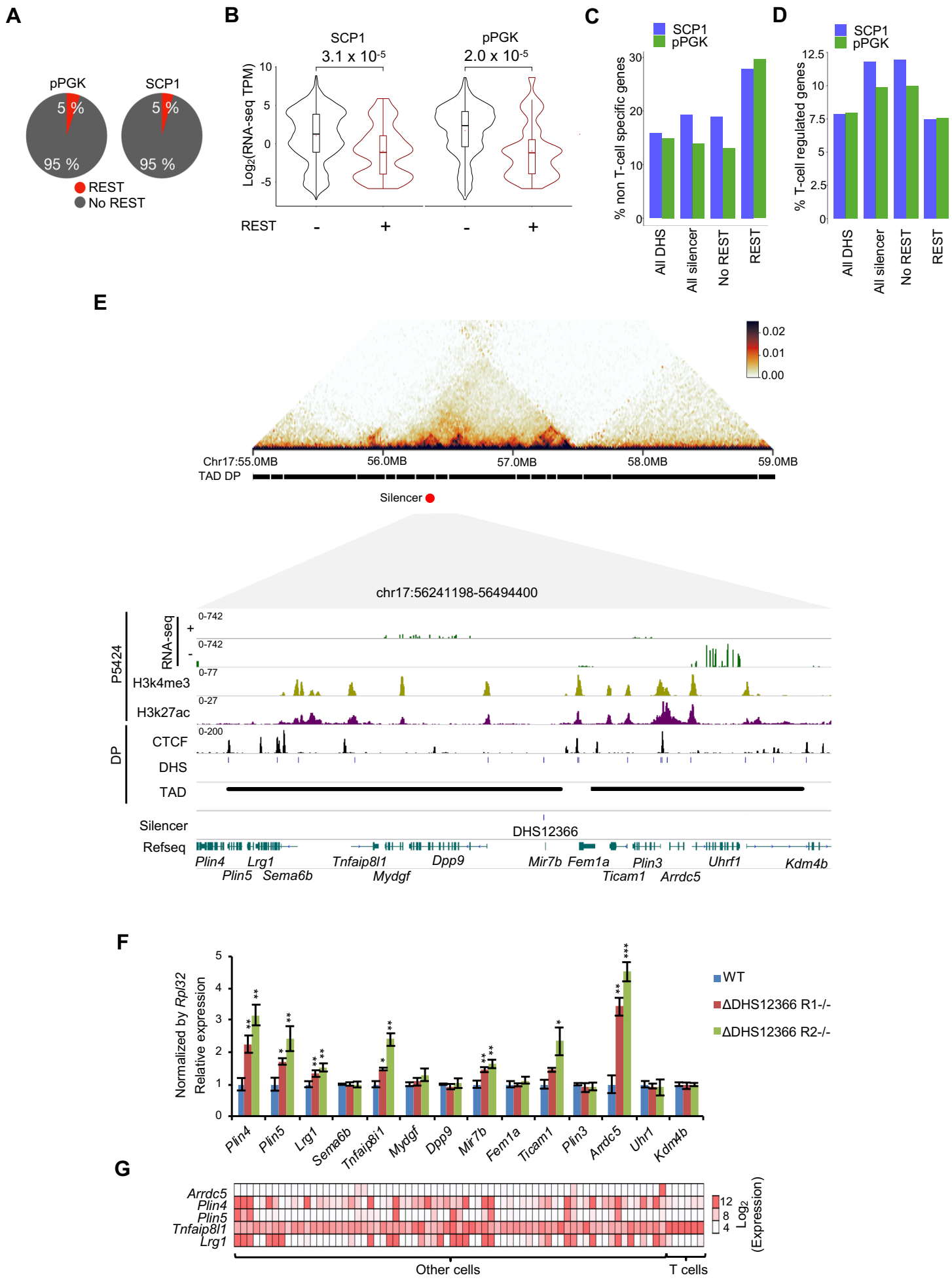

A
pPGK
SCP1
5 %
95 %
REST
No REST
B
SCP1
3.1 x 10-5
pPGK
2.0 x 10-5
Log2(RNA-seq TPM)
REST
-
+
C
SCP1
pPGK
% non T-cell specific genes
All DHS
All silencer
No REST
REST
D
% T-cell regulated genes
E
0.02
0.01
0.00
Chr17:55.0MB
56.0MB
57.0MB
58.0MB
59.0MB
TAD DP
Silencer
chr17:56241198-56494400
P5424
RNA-seq
H3k4me3
H3k27ac
DP
CTCF
DHS
TAD
Silencer
Refseq
DHS12366
Plin4
Lrg1
Tnfaip8l1
Dpp9
Mir7b
Fem1a
Plin3
Uhrf1
Kdm4b
Plin5
Sema6b
Mydgf
Ticam1
Arrdc5
F
Normalized by Rpl32
Relative expression
WT
ΔDHS12366 R1-/-
ΔDHS12366 R2-/-
Plin4
Plin5
Lrg1
Sema6b
Tnfaip8l1
Mydgf
Dpp9
Mir7b
Fem1a
Ticam1
Plin3
Arrdc5
Uhrf1
Kdm4b
G
Arrdc5
Plin4
Plin5
Tnfaip8l1
Lrg1
Log2 (Expression)
Other cells
T cells

Figure 5

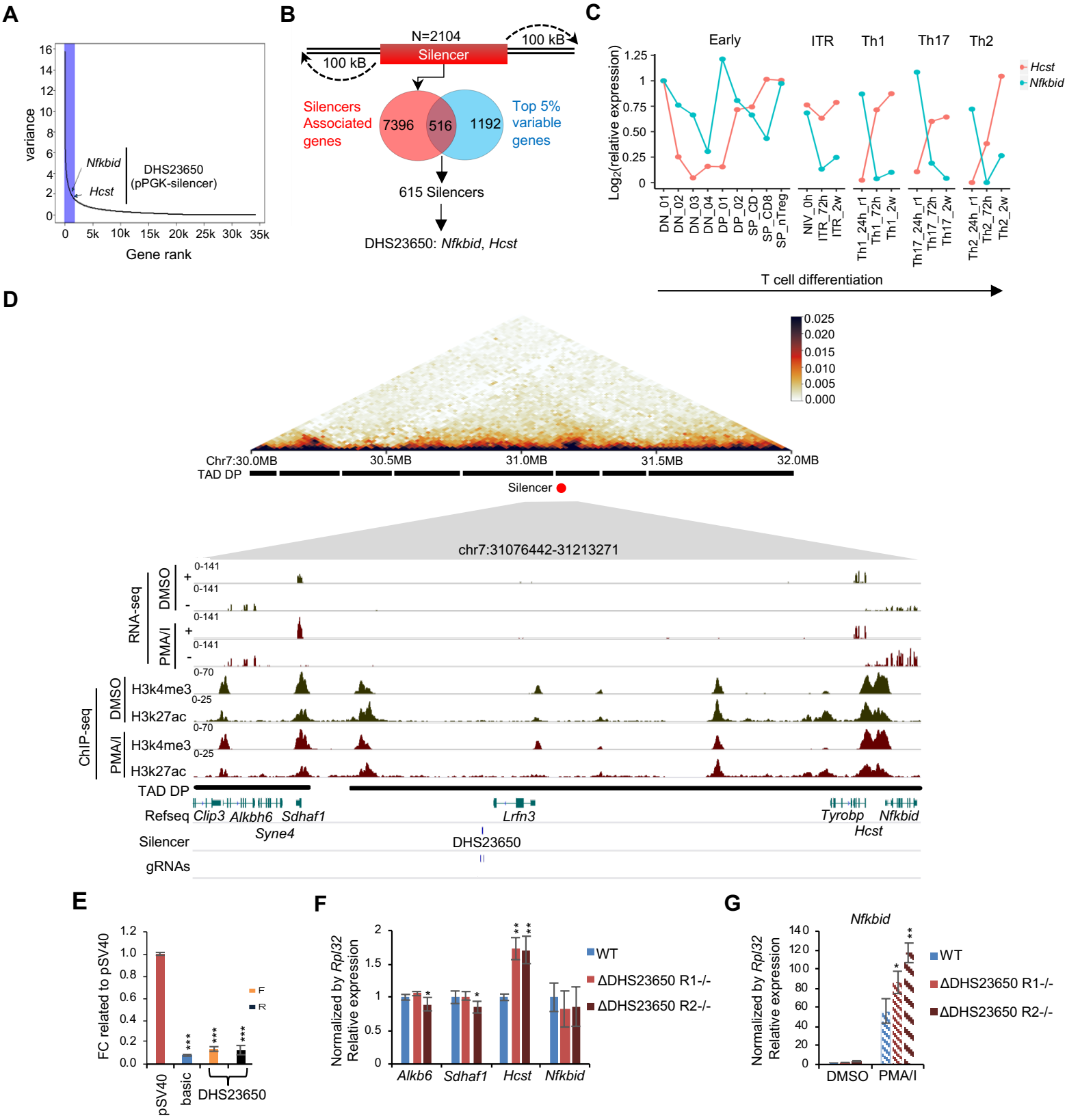

A
variance
Gene rank
Nfkbid
Hcst
DHS23650
(pPGK-silencer)
B
N=2104
Silencer
100 kB
100 kB
Silencers
Associated
genes
7396
516
1192
Top 5%
variable
genes
615 Silencers
DHS23650: Nfkbid, Hcst
C
Early
ITR
Th1
Th17
Th2
Log2(relative expression)
Hcst
Nfkbid
T cell differentiation
D
Chr7:30.0MB
30.5MB
31.0MB
31.5MB
32.0MB
TAD DP
Silencer
chr7:31076442-31213271
RNA-seq
DMSO
PMA/I
ChIP-seq
H3k4me3
H3k27ac
TAD DP
Refseq
Clip3
Alkbh6
Sdhaf1
Syne4
Lrfn3
Tyrobp
Nfkbid
Hcst
Silencer
DHS23650
gRNAs
E
FC related to pSV40
pSV40
basic
DHS23650
F
R
F
Normalized by Rpl32
Relative expression
WT
ΔDHS23650 R1-/-
ΔDHS23650 R2-/-
Alkb6
Sdhaf1
Hcst
Nfkbid
G
Nfkbid
DMSO
PMA/I

Figure 6

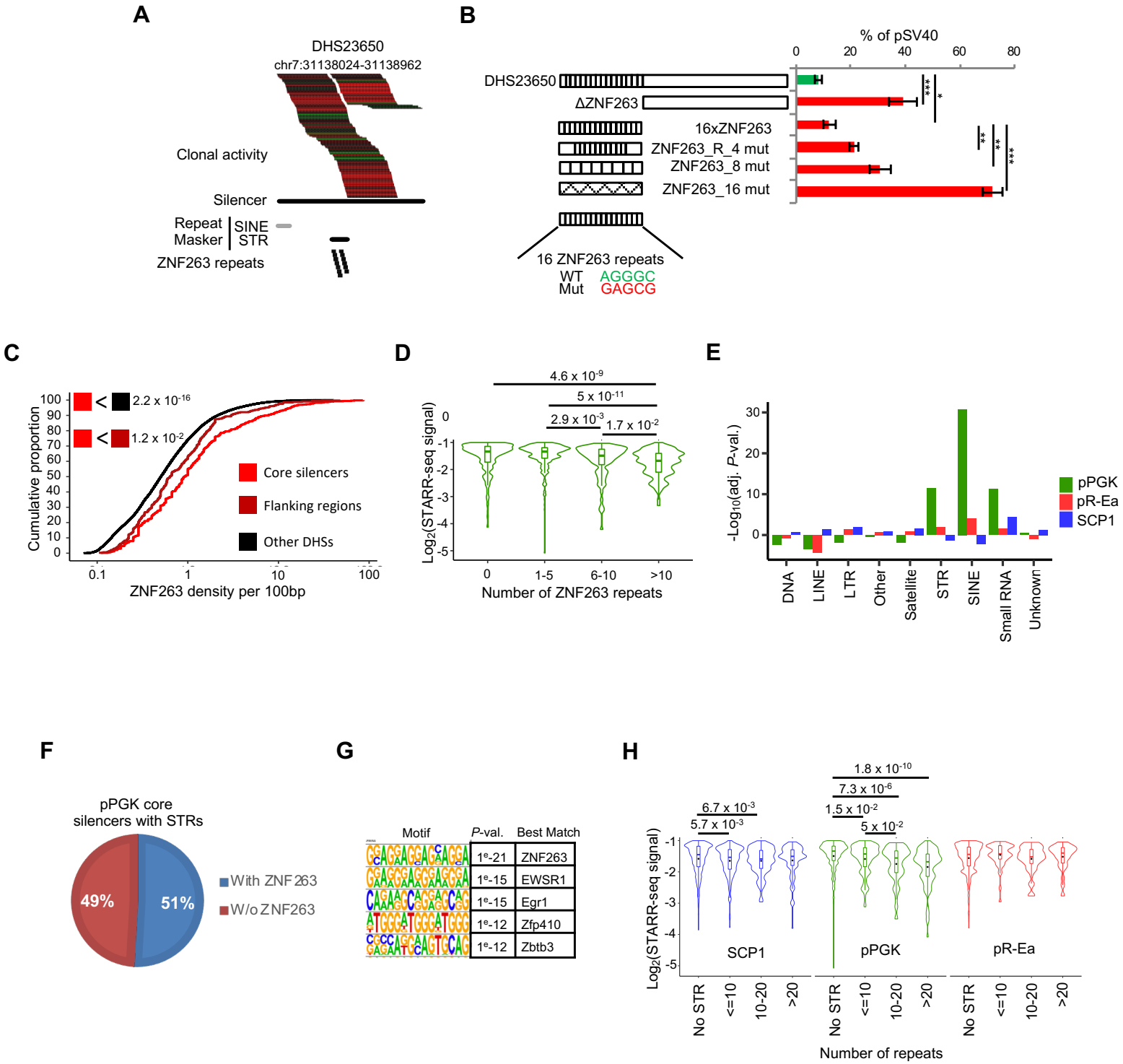
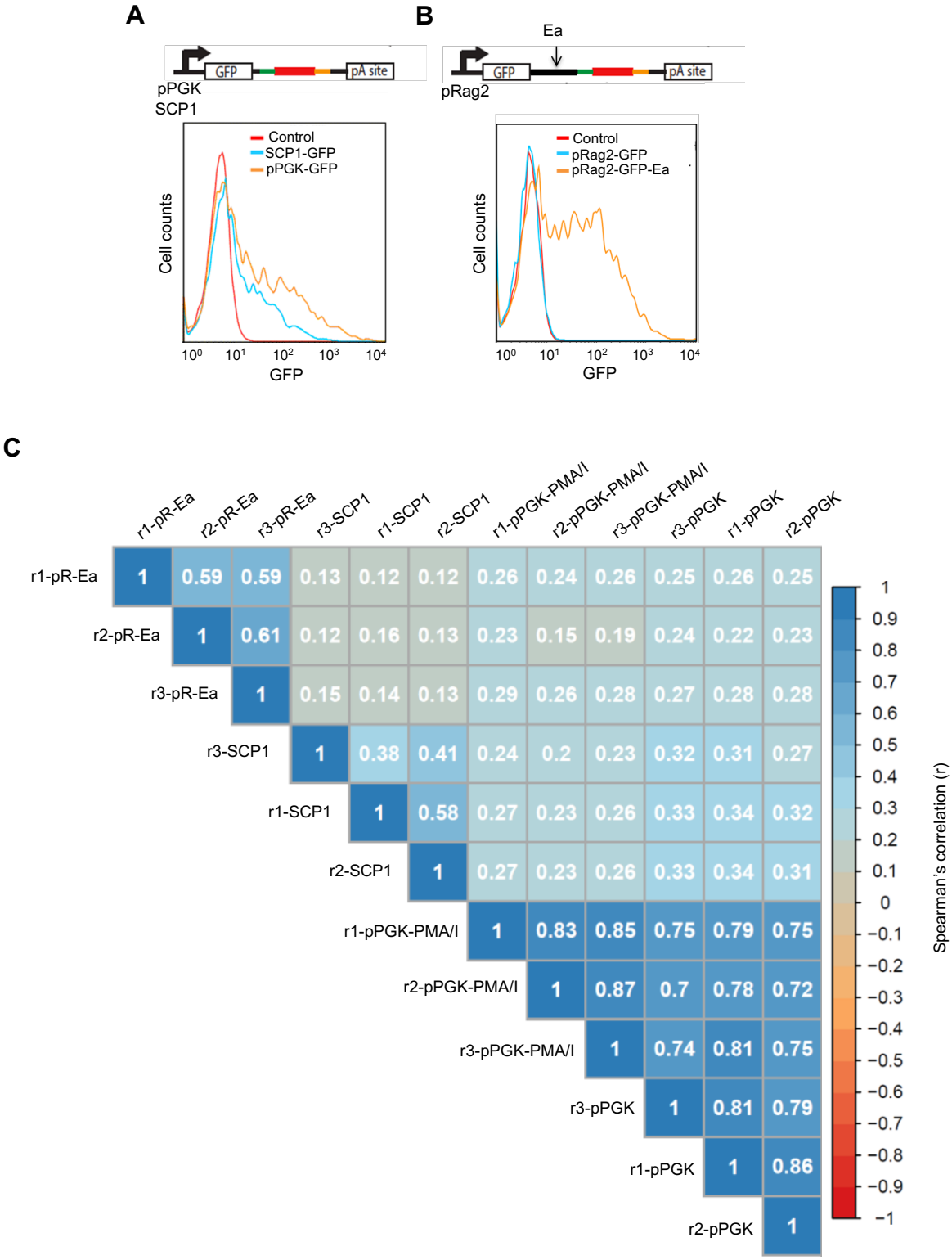
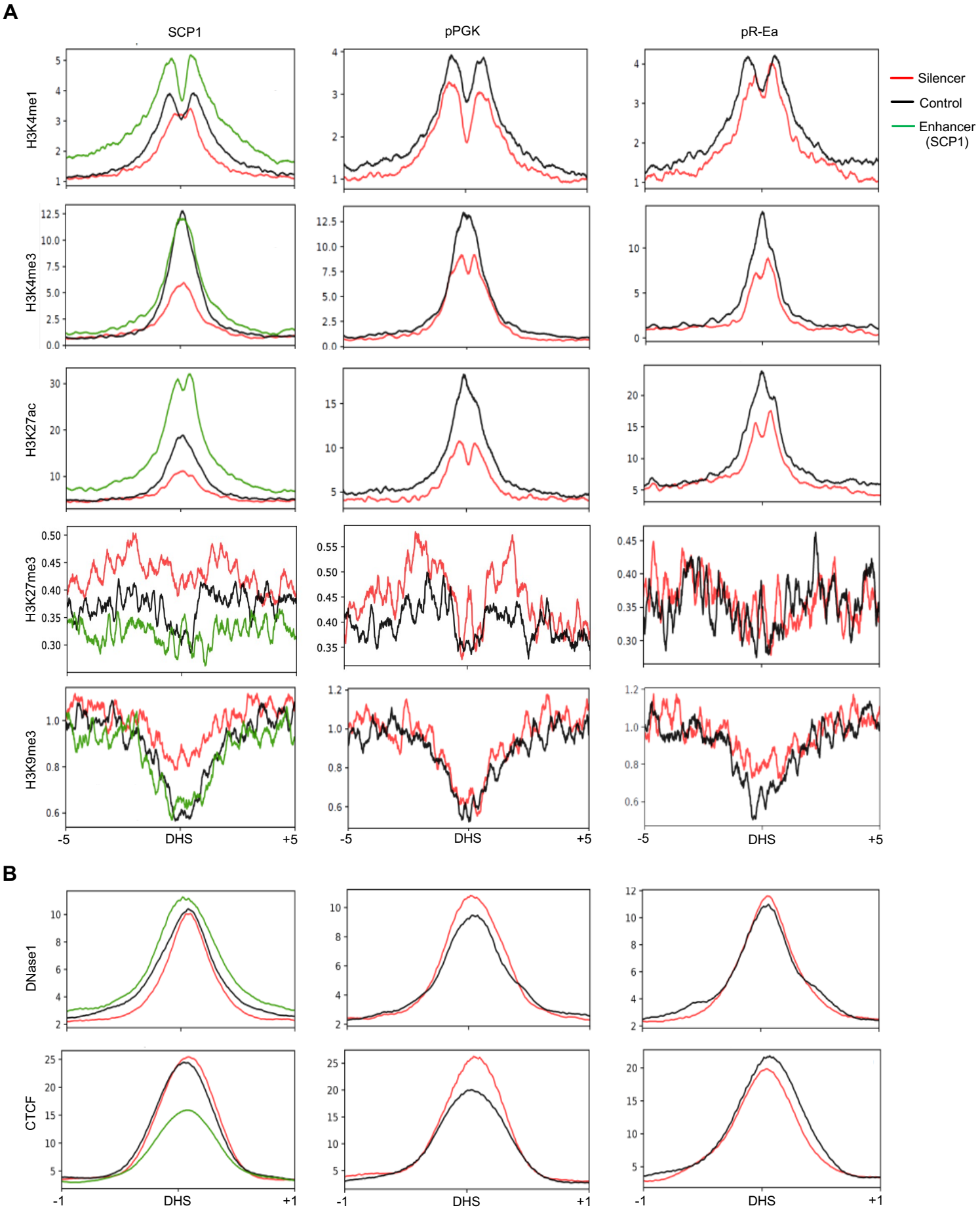

A
DHS23650
chr7:31138024-31138962
Clonal activity
Silencer
Repeat Masker
SINE
STR
ZNF263 repeats
B
% of pSV40
0
20
40
60
80
DHS23650
ΔZNF263
16xZNF263
ZNF263_R_4 mut
ZNF263_8 mut
ZNF263_16 mut
16 ZNF263 repeats
WT AGGGC
Mut GAGCG
C
Cumulative proportion
2.2 x 10-16
1.2 x 10-2
Core silencers
Flanking regions
Other DHSs
ZNF263 density per 100bp
D
4.6 x 10-9
5 x 10-11
2.9 x 10-3
1.7 x 10-2
Log2(STARR-seq signal)
0
1-5
6-10
>10
Number of ZNF263 repeats
E
-Log10(adj. P-val.)
pPGK
pR-Ea
SCP1
DNA
LINE
LTR
Other
Satellite
STR
SINE
Small RNA
Unknown
F
pPGK core silencers with STRs
49%
51%
With ZNF263
W/o ZNF263
G
Motif
P-val.
Best Match
1e-21
ZNF263
1e-15
EWSR1
1e-15
Egr1
1e-12
Zfp410
1e-12
Zbtb3
H
1.8 x 10-10
7.3 x 10-6
6.7 x 10-3
5.7 x 10-3
1.5 x 10-2
5 x 10-2
Log2(STARR-seq signal)
SCP1
pPGK
pR-Ea
No STR
<=10
10-20
>20
Number of repeats

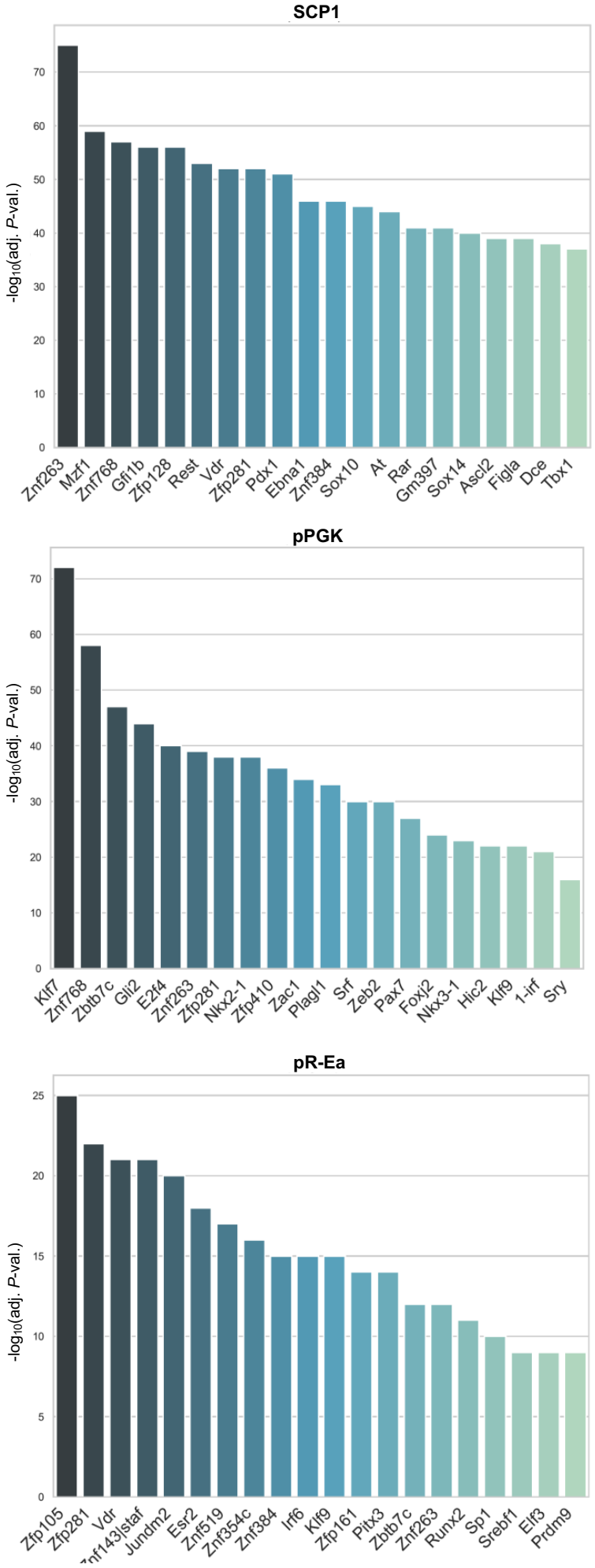
Figure 7



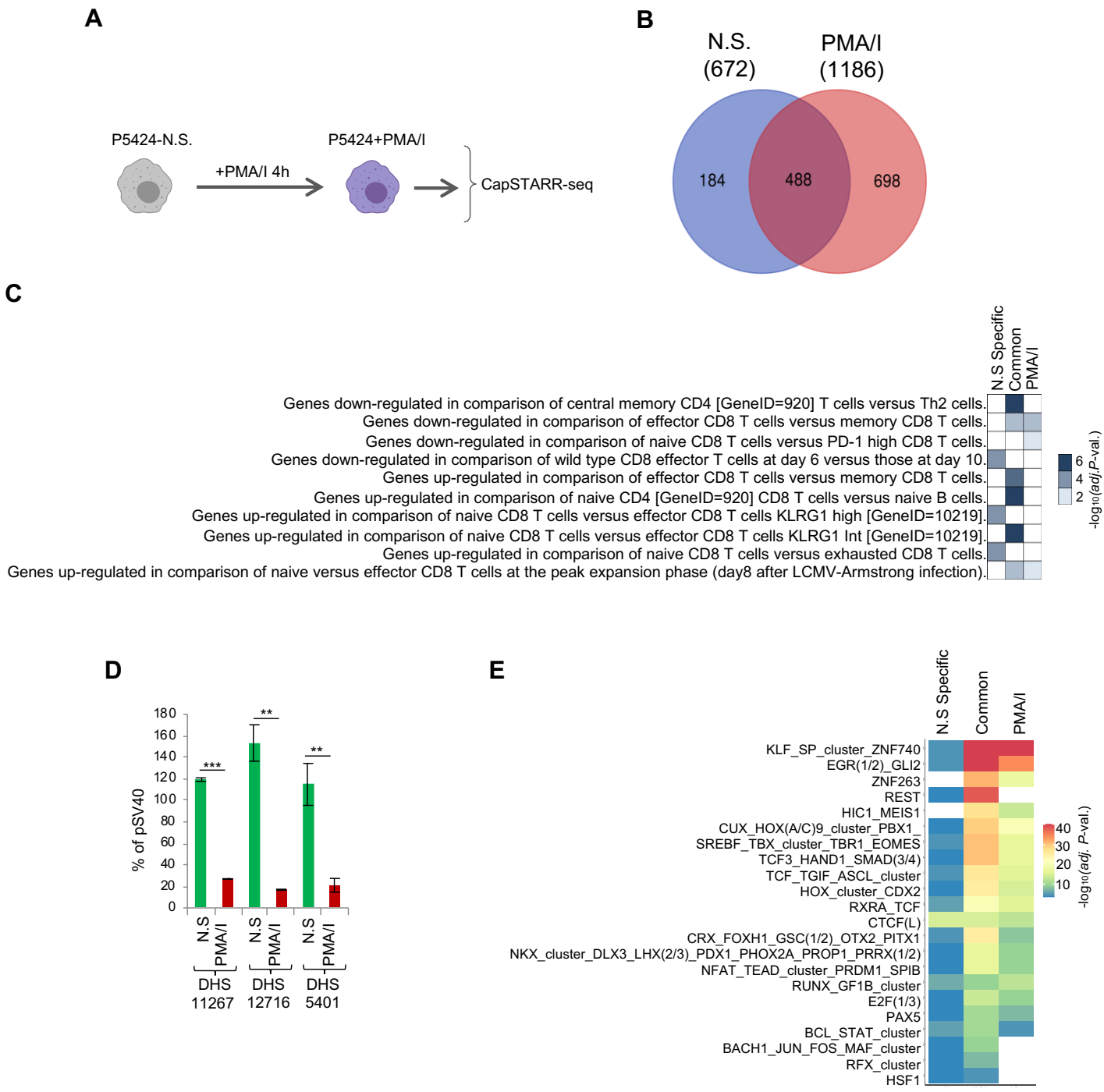
Supplementary Figure S1. Validation of the STARR-seq vectors and Correlogram of STARR-seq experiments. (**A-B**) FACS analysis of GFP expression of STARR-seq vectors containing the SCP1 or pPGK promoters (A) or the pRag2-Tcra enhancer pair (pR-Ea) (B). **(C)** The Spearman's correlation coefficient is displayed between the different STARR-seq signals between all conditions and replicates. Samples are clustered using Hierarchical Clustering.



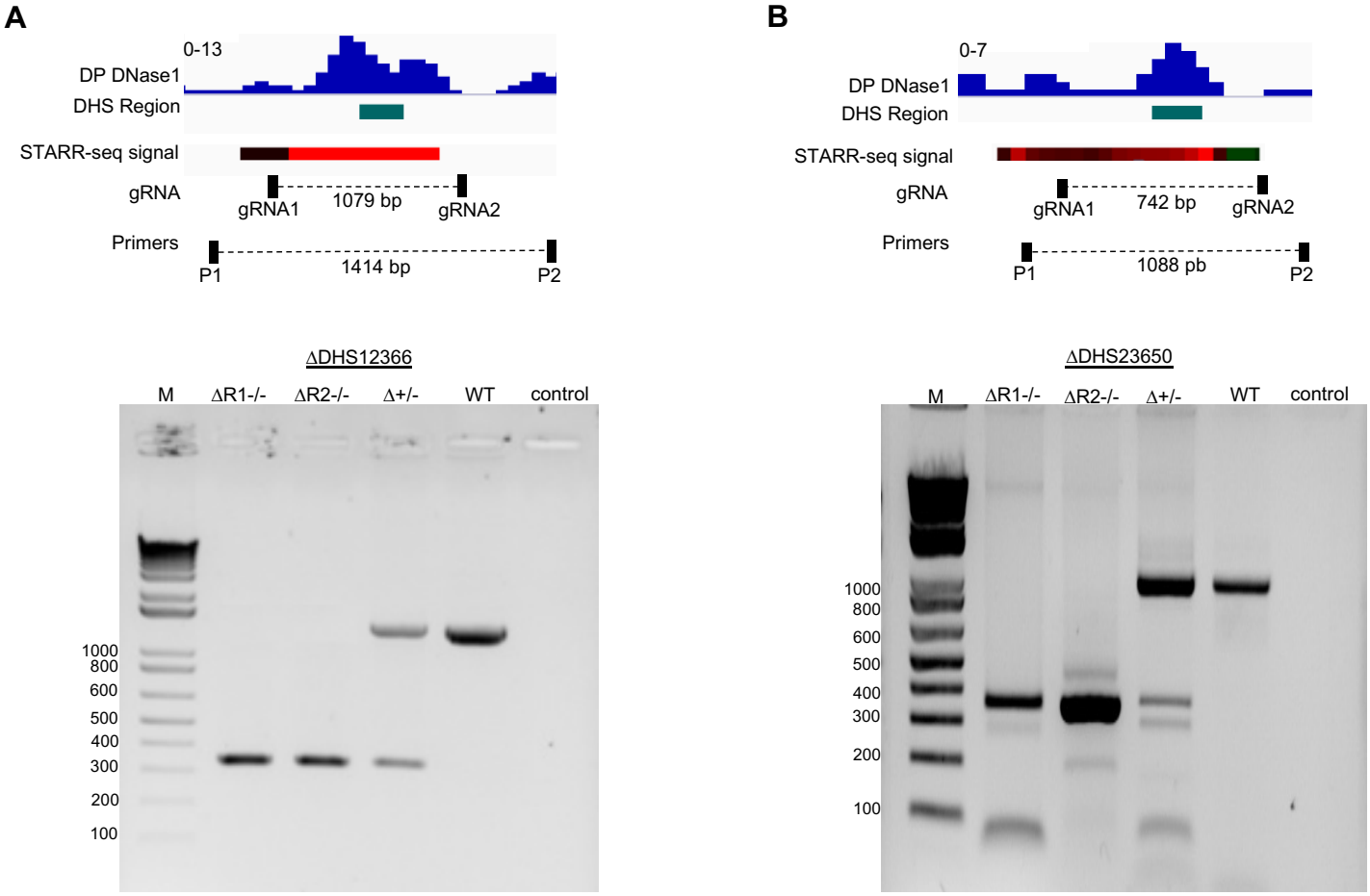
Supplementary Figure S2. Chromatin features associated with silencers. Average ChIP-seq profiles of different histone marks **(A)** and different chromatin features **(B)** based on STARR-seq signal in the mouse DP thymocytes centered on DHS regions identified as putative silencers (red), control regions (black) and SCP1 enhancers (green). Distance to the DHS are indicated in kb.



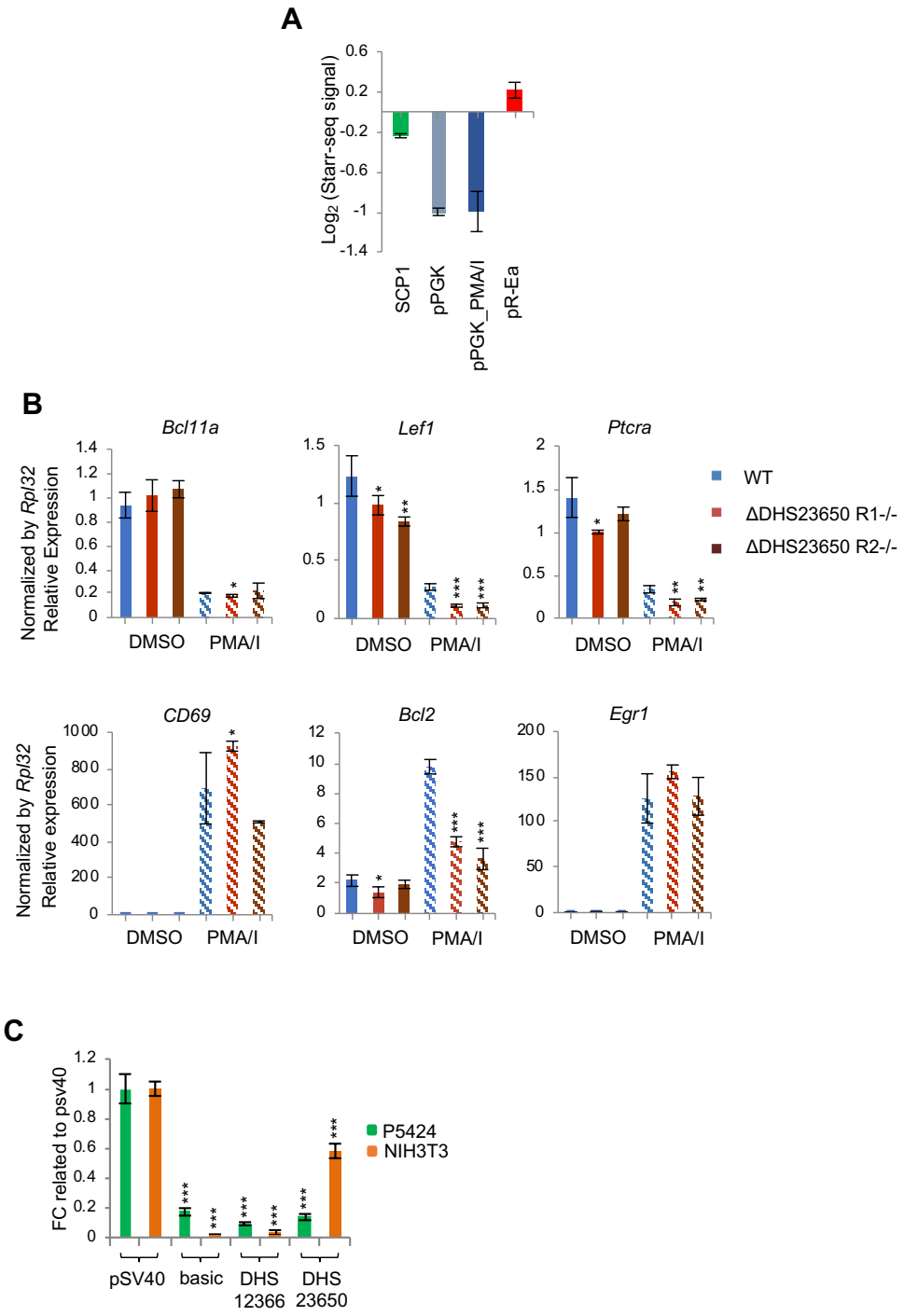
Supplementary Figure S3. *De novo* motifs enrichment analysis. Bar plots of -$\log_{10}$(adjusted *P*-value) of the top 20 most enriched *de novo* motifs identified using the HOMER software in the three set of silencers.



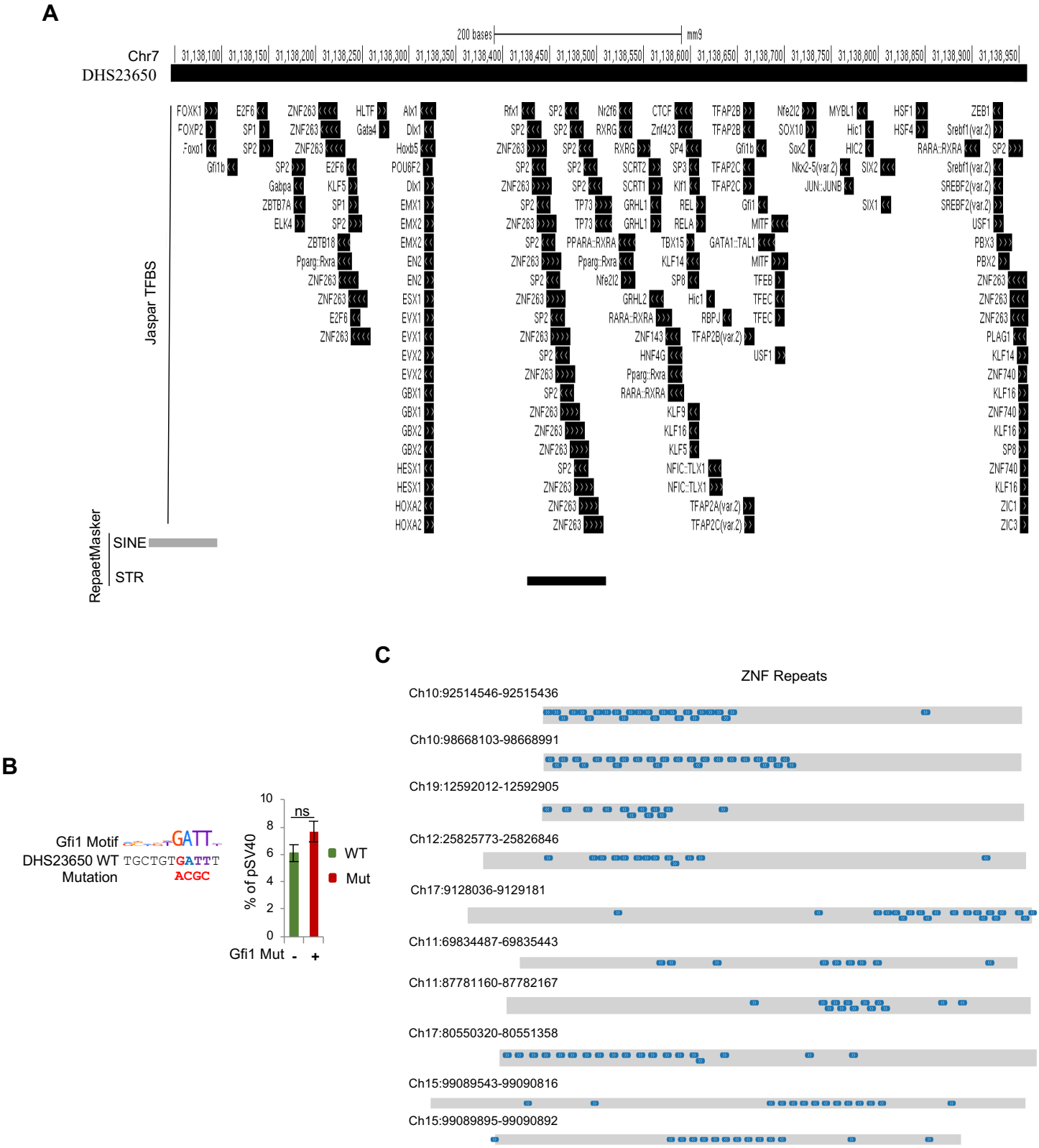
Supplementary Figure S4. Silencers regulated by PMA/ionomycin stimulation. (**A**) Experimental approach to identify putative silencers in non-stimulated (N.S.) and PMA/ionomycin stimulated (PMA/I) P5424 cells using the pPGK library. (**B**) Venn diagram displaying the overlap between silencers identified in both conditions. (**C**) Heatmap displaying the top five MsigDB immunological signatures identified in common and N.S. or PMA/I specific silencers. (**D**) Luciferase reporter assay of three PMA/I-specific silencers. Experiments were performed in stimulated and non stimulated conditions. Error bars show s.d. from three independent transfections (***P < 0.001, **P < 0.01, *P < 0.05; two-sided Student's t test). (**E**) Enrichment scores of top 10 most enriched motifs in common and N.S. or PMA/I specific silencers.



Supplementary Figure S5. Genomic deletion of DHS12366 and DHS23650 silencers using CRISPR/Cas9 gene editing. (**A, B; top pannels**) Genomic tracks showing DP DNase1 signals, DHS, STARR-seq signal, sgRNAs and primers. (**A, B; bottom pannels**) The homozygous and heterozygous genomic deletions were detected by PCR using the corresponding P1 and P2 **primers**. Details on the gRNA sequences and PCR primers are provided in the **Supplementary Table S6**. 1kb ladder plus is indicated.



Supplementary Figure S6. DHS23650 silencer activity and functional assays. (**A**) STARR-seq signal of DHS23650 silencer in the different libraries and conditions. (**B**) Expression analyses of the indicated T cells markers in wild-type and DHS23650-deleted P5424 cells treated with DMSO (control) or PMA/I. (**C**). Luciferase reporter assays in P5424 and NIH3T3 cell lines. Data represent the normalized fold change over the pSV40 vector. Error bars show s.d. from three independent experiments (***P < 0.001, **P < 0.01, *P < 0.05; two-sided Student's t test).



Supplementary Figure S7. Genomic tracks for TFs and repetitive elements around DHS23650 and ZNF263 binding sites. (**A**) UCSC genome browser (NCBI37/mm9) displaying the Jaspar TFBS and the RepeatMasker tracks at the DHS23650 silencer. (**B**) Luciferase reporter assay to assess the impact of Gfi1 binding site on DHS23650 silencer activity. The consensus Gfi1 sequence is indicated. Error bars show s.d. from three independent transfections (***P < 0.001, **P < 0.01, *P < 0.05; two-sided Student's t test). (**C**) Examples of pPGK silencers harboring ZNF263 tandem repeats using TFMotifview tool.

2.5 Projets annexes

Le pipeline d'analyse de données CapSTARR-seq a été utilisée dans plusieurs projets de l'équipe, ce qui a donné lieu à deux publications dans lesquelles j'ai été impliqué. Je présenterais brièvement ces publications dans lesquels j'ai contribué en tant que co-auteur.

2.5.1 Identification d'enhancers impliqué dans la régulation du gène *Ikzf1* via une technologie de test de gène rapporteur à grande échelle

Dans cette publication, nous nous sommes intéressés à la régulation du gène Ikzf1 impliqués dans la différenciation cellulaire des cellules lymphoblastiques de type T et B [Georgopoulos, 2017, Heizmann et al., 2018]. En effet, plusieurs régions régulatrices potentielles ont été identifiées à proximité d'Ikzf1, cependant, les mécanismes régulateurs du locus Ikzf1 ne sont pas encore totalement compris. Une dérégulation du gène Ikzf1 étant impliquée dans l'apparition de pathologie tel que la leucémie.

Nous avons pu identifier un enhancer situé à 120kb en amont du gène Ikzf1 en intégrant des données issues de différentes technologie notamment le CapSTARR-seq et le 4C-seq afin d'étudier les interactions. Dans ce projet, mon pipeline a servi à traiter les données de CapSTARR-seq, notamment afin de générer un fichier possédant un code RGB, où l'on peut visualiser l'activité par clone (Figure : 2.4).

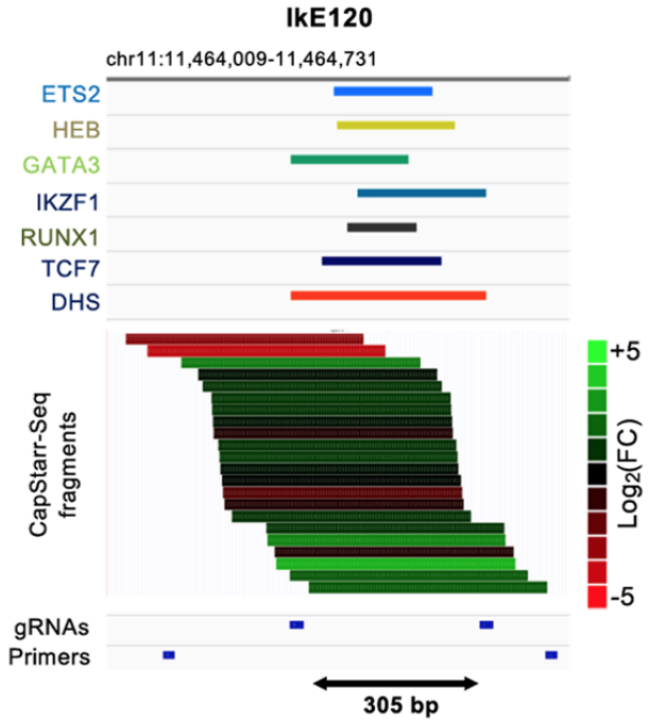


FIGURE 2.4 – L'enhancer IkE120 est situé à 120 kb en amont du gène Ikzf1. On peut visualiser l'activité des clones par un code RGB proportionnel à l'activité. L'activité enhancer principale est colocalisée avec les sites de fixations des facteurs de transcriptions au sein de cette région enhancer. La délétion de cette région à l'aide de la technologie CRISPR/Cas9 a démontré le rôle de cet enhancer dans le contrôle de l'expression du gène Ikzf1

La visualisation des fragments appartenant à cet enhancer a permis de confirmer un chevauchement entre les clones possédant une activité enhancer et les sites de fixations des facteur de transcription au sein de cette enhancer. Une délétion par CRISPR/Cas9 des régions chevauchant les clones possédant l'activité enhancer a permis de mettre en évidence une dérégulation du

gène Ikzf1. En effet, suite à cette délétion, une baisse du niveau de H3K27ac a été relevé à ce niveau, démontrant un impact dans la régulation épigénétique et la perte du pouvoir régulateur (Article : 4.1).

2.5.2 Les Epromoteurs : des plateformes pour le recrutement des facteurs de transcription nécessaire à la réponse inflammatoire

Différentes études ont mis en évidence le rôle des E-promoteurs dans la réponse aux stress [Dao et al., 2017, Dao and Spicuglia, 2018, Muerdter et al., 2018]. Dans ce projet, nous nous sommes intéressés aux mécanismes de réponse à l'Interféron et cette étude a permis de mettre en évidence l'implication des E-promoteurs en tant que plateforme de recrutement pour la fixation de facteurs de transcription nécessaire à la réponse inflammatoire. Afin d'étudier les E-promoteurs, la technologie du CapSTARR-seq a été utilisée sur les promoteurs des gènes afin de saisir leur activités enhancers en présence ou absence de stimulation à l'Interféron de type I. Le but était de mettre en évidence le rôle des E-promoteurs dans la régulation de l'induction génique après stimulation par l'IFNα, et plus généralement des mécanismes lors de la réponse inflammatoire.

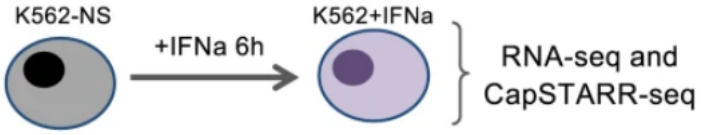


FIGURE 2.5 – Schéma de la stratégie expérimentale des cellules avec ou sans IFNa sur lesquels ont été mené un CapSTARR-seq et un RNA-seq

Cette étude a permis de mettre en évidence la régulation de clusters gènes régulés par des E-promoteurs localisés dans le même cluster. Un de ces clusters est le locus OAS, qui contient trois gènes impliqués dans la réponse antivirale [Hornung et al., 2014]. Dans ce locus, les trois gènes induits par l'IFNα étaient associés à des E-promoteurs inductibles. Nous avons remarqué que l'E-promoteur associé à OAS3 affichait une affinité plus élevée aux facteurs de transcription de réponse à l'interféron (STAT1/2 et IRF1/9), par rapport aux E-promoteurs OAS1 et OAS2. Ce résultat a été validé expérimentalement par des KO des différents OAS induit par CRISPR/Cas9, montrant qu'une délétion de OAS1 et OAS2 n'affecte pas l'expression de OAS3 alors qu'une délétion de OAS3 consiste en une réduction drastique de l'induction de OAS1 et OAS2 après stimulation par l'IFNα.

Dans cette étude, j'ai réalisé le traitement des données de CapSTARR-seq sous les différentes conditions que sont avec et sans stimulations par l'IFNα. Mon analyse a permis de mettre en évidence des promoteurs possédant une activité enhancer (E-promoteur) impliqué dans la réponse à la stimulation par l'IFN. La visualisation de l'activité des fragments et/ou des régions s'est révélé ici utile afin de se donner une idée rapidement et simplement de la distribution du signal ou de l'activité dans certaines régions (Article : 4.2).

2.5.3 OLOGRAM : Modélisation de la distribution des croisements de données génomiques

En génomique, la plupart des analyses vont aboutir à des fichiers possédant des coordonnées renseignant sur la position de l'élément étudié : fixation d'une protéine, marque épigénétique, etc. Une question récurrente et essentielle en bioinformatique génomique est de déterminer si les colocalisations identifiées sont significatives sur le génome qui pourraient suggérer une

relation fonctionnelle entre elles. Plusieurs approches et implémentations sont disponibles [Simovski et al., 2018], et diffèrent principalement sur le modèle statistique proposé. Parmi ces outils on peut citer : CEAS qui va reposer sur une loi binomiale ; [Shin et al., 2009], Bedtools Fisher qui est basé sur un modèle hypergéométrique [Quinlan, 2014] ou sur des modèles empiriques tel que HyperBrowser ou encore MULTOVL [Sandve et al., 2010, Aszódi, 2012]. Dans ce dernier cas, les régions des jeux de données à comparer sont mélangées homogènement sur le génome plusieurs fois afin d'obtenir une distribution empirique totalement aléatoire. La comparaison des chevauchements obtenus avec cette distribution aléatoire permet d'obtenir une P-valeur permettant éventuellement de rejeter l'hypothèse nulle. Cependant ce modèle ne tient pas compte des distances existantes entre les régions testées. D'après ce que j'ai pu présenter précédemment, on peut facilement poser l'hypothèse que la répartition des éléments au sein du génome n'est pas aléatoire, les éléments génomiques ne sont pas distribués au hasard. L'unique outil connu permettant cette approche est accessible seulement à l'intérieur de l'instance Galaxy du Genomic HyperBrowser, ce qui le rend impossible à intégrer dans pipeline développé localement. D'ailleurs OLOGRAM a été déployé au sein de PyGTFtk qui permet de manipuler des données au format GTF et ainsi d'enchainer les commandes.

Ainsi OLOGRAM a été développé dans le but de comparer des jeux de données génomiques, comparer leurs positions afin de savoir si les colocalisations observées sont significatives ou pas. De plus il tient compte de la distribution des distances entre régions et permet d'inclure ou d'exclure des régions données afin de mener une comparaison relative. Cet outil a été implémentée, dans la majeure partie par Quentin Ferré. J'ai contribué à ce projet par les nombreux tests et identification de bugs qu'il y a pu avoir au cours du développement (Article : 4.3).

2.5.4 Crible CRISPR à grande échelle

Parmi les techniques de tests fonctionnels à grande échelle, on a pu citer le CRISPR comme méthode récente pour l'étude des éléments cis-régulateurs. Au cours de ma thèse, j'ai pu contribuer à un projet dont le but est l'identification d'enhancers impliqués dans les processus de leucémie aiguë lymphoblastique de type T (T-ALL).

Pour cela, nous nous sommes basés sur les lignées cellulaires CEM et Jurkat qui sont des lignées cellulaires T-ALL. Par la suite, en me basant sur des données générées au sein du laboratoire, j'ai appliqué une liste de filtres afin de sélectionner les enhancers distaux ($> 2kB$du TSS) spécifique à la T-ALL (Figure : 2.6). Ainsi je me suis basé sur les pics d'ATAC-seq chevauchant les pics d'H3K27ac dans une fenêtre d'au moins 100pb. Les pics étant à moins de 2kB du TSS d'un gène étant préalablement exclu afin de ne retenir que les enhancers distaux. Par la suite j'ai exclu les pics qui étaient également retrouver dans les cellules thymiques le but étant de garder les enhancers spécifique à la T-ALL. Un filtre sur le score du pic a été appliqué afin de garder les 1300 pics possédant le meilleur score par lignée cellulaire (CEM et Jurkat). Parmi ces 1300 enhancers par lignée cellulaire, environ 257 sont partagés parmi CEM et Jurkat.

Par la suite, j'ai créé un pipeline afin d'automatiser le design de la librairie permettant la création de guides personnalisés. Le but était de fournir un fichier BED contenant les coordonnées des régions d'intérêts, peu importe leurs nombres, et de fournir en sortie un fichier contenant une liste de guide par région avec pour chaque guide, les coordonnées, la séquence du guide, et un identifiant unique.

Pour cela je me suis basé sur l'outil CHOPCHOP [Labun et al., 2019] qui est utilisable en ligne de commande. Contrairement à la majorité des outils générant des guides ARN, CHOPCHOP ne demande pas exclusivement les coordonnées du TSS de la région visée, ce qui dans notre cas était impossible à fournir car nous nous basions sur les enhancers distaux. En effet, la plupart des stratégies visent à impacter la transcription d'un gène et c'est souvent le TSS qui est visé par le guide et donc demandé par les outils pour créer des guides qui cibleront cette région.

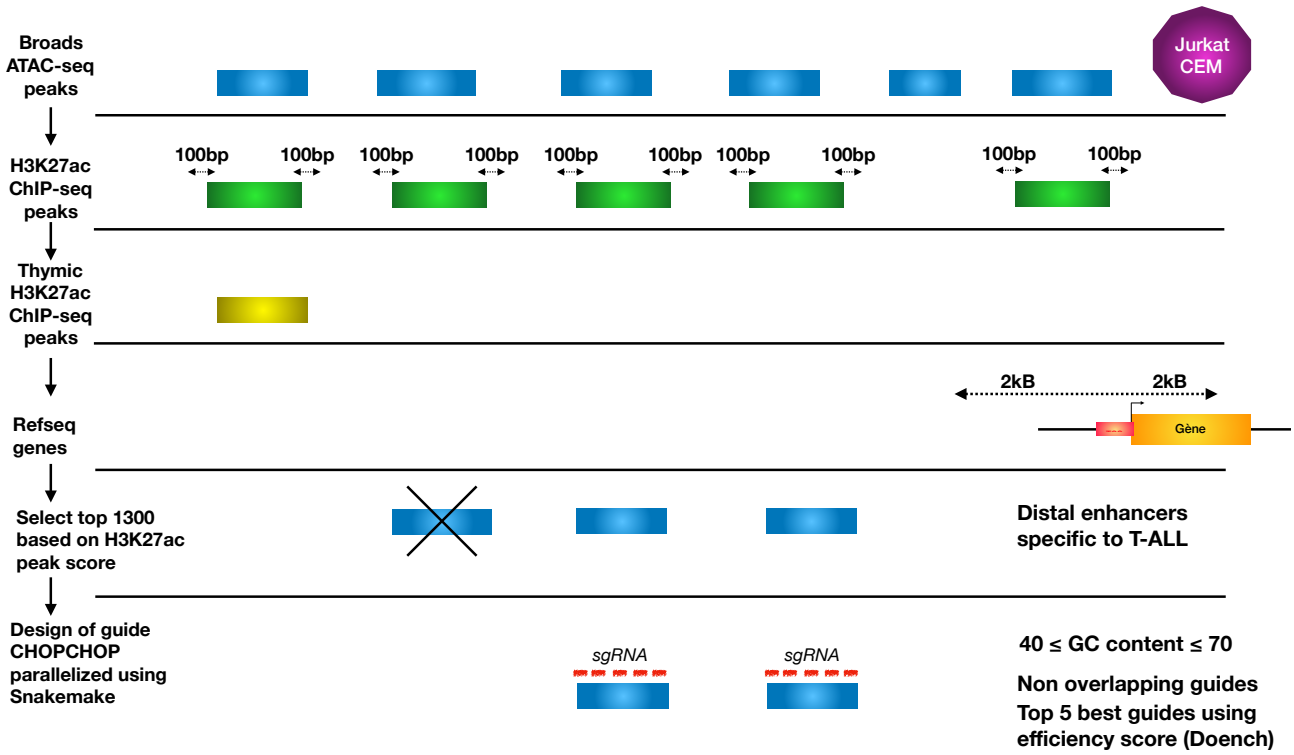


FIGURE 2.6 – Stratégie du crible CRISPRi sur les enhancers distaux afin de mettre en évidence les enhancers impliqués dans les T-ALL

CHOPCHOP s'affranchit de cette contrainte, il est possible d'effectuer des requêtes sur des coordonnées génomiques indépendamment de leurs rôles. Une contrainte est que CHOPCHOP ne prend en entrée qu'une coordonnée à la fois. Ainsi, s'il faut générer des guides pour 10.000 régions, il faudrait lancer manuellement 10.000 fois CHOPCHOP. C'est pour cela que j'ai parallélisé ce processus via Snakemake. En donnant en entrée un fichier BED contenant un nombre N de lignes, CHOPCHOP sera lancé N fois. CHOPCHOP n'est pas gourmand en ressources, en fonction de la puissance du serveur de calcul, il est possible de lancer parallèlement plusieurs processus CHOPCHOP. Cet outil dispose d'option intéressantes comme plusieurs méthodes de calcul pour le score du guide, ou encore la possibilité de choisir uniquement des guides qui ne se chevauchent pas afin d'étendre la zone couverte.

Dans notre cas, les régions font environ 200pb et la taille d'un guide RNA est de 20pb. Comme je souhaitais obtenir des guides qui ne se chevauchent pas, j'ai sélectionné 5 guides par enhancer afin de couvrir raisonnablement la région. Au total, j'ai obtenu environ 11,000 guides, auxquels nous avons ajouté 1,000 guides contrôle négatif provenant de la librairie CRISPRi V2 ce qui correspond à un total de 12000 guides, soit le nombre de guides que l'on pouvait commander. L'expérience est actuellement en cours.

Afin d'anticiper l'obtention des données CRISPR, j'ai été amené à expérimenter des outils de traitement de données CRISPR. En soit l'analyse de données CRISPR consiste à comparer le nombre de guides obtenues par gène entre la condition de départ et la condition de test. Un simple t-test peut se révéler suffisant afin d'obtenir les gènes significativement touchés par le CRISPR comme a pu le démontrer [Bodapati et al., 2020]. Dans mon cas, c'est l'outil MAGeCK

[Li et al., 2014] qui a retenu mon attention. En effet, MAGeCK calcul une p-valeur par guide basée sur une négative binomiale, par la suite il ne va considérer que les guides étant sous un seuil ne conservant pas les « mauvais » guides. J'ai trouvé cela particulèrement intéressant d'appliquer un test statistique par guide, car pas tous les outils ne l'effectuent. Par la suite un score est calculé basé sur l'algorithme RRA (Robust Ranking Algorithm), nous fournissant une p-valeur sur la significativité du score RRA et une FDR. MAGeCK contient toute une suite d'outil qui va également permettre d'analyser les ontologies des gènes mis en évidence par l'expérience de CRISPR. De plus il est utilisable en ligne de commande, donc localement et il est possible de l'inclure dans des "pipelines maisons".

À l'heure actuelle ces analyses ont été portées sur les enhancers car comme nous avons pu le dire à plusieurs reprises, ces éléments sont bien identifiés et caractérisés. Par la suite, il serait intéressant d'effectuer une analyse similaire sur les silencers afin de mettre en évidence les mécanismes dans lesquels ils sont impliqués.

3 Discussion & Perspectives

La compréhension des éléments régulateurs est un enjeu majeur en génomique, permettant de mieux comprendre l'organisation des cellules, la différenciation cellulaire et l'impact des dysfonctionnements sur l'apparition de pathologies. Alors que les études à grande échelle des éléments augmentant l'activité des gènes tel que les promoteurs et les enhancers ont plutôt été bien couvertes, l'étude à grande échelle des éléments répresseurs tel que les silencers peine à s'amorcer. Une façon d'étudier l'activité silencer est de piéger sa séquence et de la tester dans un test de gène rapporteur, en s'attendant à une diminution de l'activité par inhibition sur le promoteur du gène rapporteur [Qi et al., 2015, Petrykowska et al., 2008]. Afin de les étudier à grande échelle, nous avons adapté la technique de STARR-seq [Vanhille et al., 2015]. Cette technique s'est montrée efficace afin d'identifier à grande échelle les enhancers et Epromoteurs [Dao et al., 2017, Santiago-Algarra et al., 2021]. Nous avons identifié des fragments d'ADN capables de réguler négativement leur propre transcription dans un contexte épisomique minimal dans des cellules transfectées, permettant un criblage à l'échelle du génome des silencers. Comme première évaluation de notre approche, nous avons analysé un ensemble de 28,055 DHS provenant de thymocytes de souris en développement dans une lignée primaire de cellules de lymphocytes T précédemment utilisée dans STARR-seq [Vanhille et al., 2015].

3.1 Identification à grande échelle & mécanismes d'actions des silencers

3.1.1 Les différentes stratégies CapSTARR-seq pour l'identification des éléments silencers à l'échelle du génome

Plusieurs stratégies faisant varier les promoteurs dans les vecteurs du STARR-seq ont été testé afin d'identifier la meilleure condition permettant l'identification des éléments silencers. Nous avons testé succinctement SCP1, pPGK et un promoteur couplé à un enhancer tissu-spécifique pR-Eα (pRag2-Eα). Ces 3 promoteurs possèdent des caractéristiques distinctes. SCP1 avait déjà été testé pour l'identification d'enhancers [Vanhille et al., 2015] et il est intéressant de noter que nous avons pu identifier à nouveau ces éléments enhancers via SCP1. Ainsi, SCP1 est une bonne solution pour l'identification des silencers, cette stratégie étant celle nous ayant permis d'identifier le plus grand nombre de silencers. Ainsi, SCP1 est l'une des conditions nous ayant permis de déceler l'enrichissement du facteur de transcription répressif REST. La stratégie utilisant pPGK est également plus qu'acceptable pour l'identification des silencers. Les silencers obtenus par pPGK présentent un enrichissement en séquences répétées, notamment des SINEs et des STRs, nous permettant de proposer un mécanisme d'action pour ces silencers, d'autant que ces silencers possèdent une activité inhibitrice plus forte que les autres silencers identifié par pPGK. La paire pR-Eα tissus-spécifique permet l'identification de silencers dans une moindre mesure. Les silencers identifiés via cette stratégie ont tout de même pu être validés par un test de luciférase. Cependant, je ne considère pas cette stratégie comme optimale pour l'identification de silencers. Une hypothèse serait que l'enhancer interfèrerait avec la séquence testée, atténuant son signal. Il est intéressant de noter que le facteur de transcription REST n'est pas enrichi au sein des silencers identifiés par pR-Eα, ce dernier permettant très certainement l'identification

de silencers tissu-spécifiques.

3.1.2 REST un facteur de transcription répresseur pour les silencers ubiquitaires

REST est un facteur de transcription répresseur se fixant aux silencers inhibant les gènes neuronaux dans les cellules non-neuronales [Schoenherr and Anderson, 1995, Chong et al., 1995]. Ainsi on peut le définir comme étant ubiquitaire, et non tissu-spécifique. Grâce au MPRA, le site de fixation de REST a été étudié [Mouri et al., 2022] montrant qu'une délétion du site de fixation de REST allait être critique pour le silencer RE1 et pointant les sous domaines nécessaires à l'activité répressive. Dans notre cas, nous avons identifié les silencers possédant le site de fixation de REST comme étant 'plus fort' que les autres silencers. Cependant, les silencers possédant un site de fixation REST semblaient moins associés à des gènes spécifiques au lymphocyte T et plus associé à des gènes tissus spécifiques d'autres tissus. Ceci nous permet de suggérer un mécanisme d'action pour les silencers qui soit REST-dépendant. Ces derniers seraient plus ubiquitaires et impliqués dans la régulation de gènes spécifique de plusieurs tissus, ce qui concorde avec des études montrant le rôle de REST dans l'organisation de plusieurs tissus, tels que le tissu cardiaque [Kuwahara et al., 2003] ou encore le pancréas [Rovira et al., 2021]. De plus, les silencers contenant REST affichaient un très faible taux de faux positifs dans les tests de validation rétrovirale, indiquant que la présence du site de liaison REST fournit une activité silencer universelle (Figure 3.1).

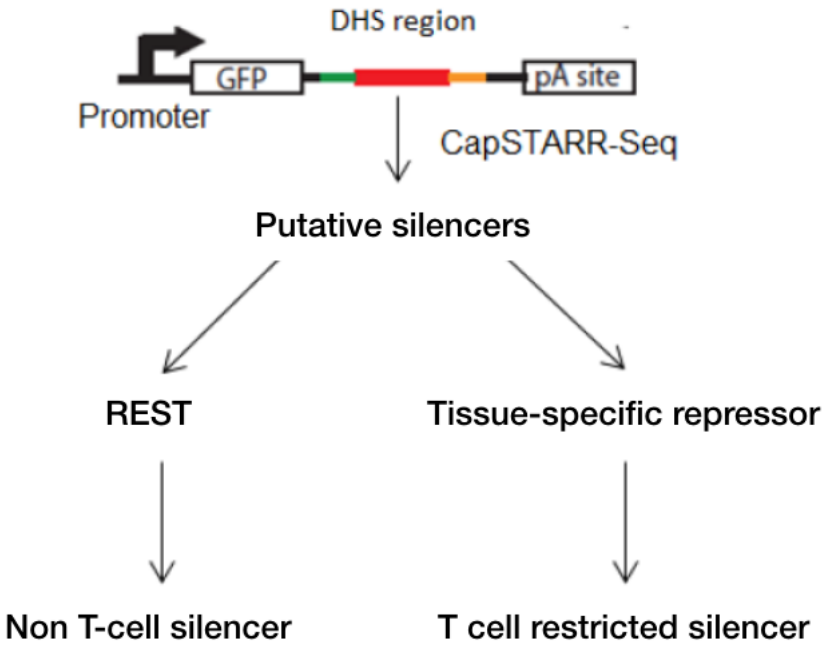


FIGURE 3.1 – Mécanisme d'action des silencers identifiés. Les silencers porteurs de REST semblent agir comme silencers ubiquitaires au contraire des autres qui semblent être tissu-spécifiques.

D'autres facteurs de transcription possédant un pouvoir répresseur ont été identifiés comme enrichis dans notre jeux de silencers tel que GFIb [Zweidler-McKay et al., 1996], SMAD3 [Takimoto et al., 2010, Li et al., 2006] et HOX [Cieslak et al., 2020]. Certains silencers possédant ces facteurs de transcriptions ont également été validés expérimentalement par test de gène rapporteur.

3.1.3 Autres mécanismes d'actions des silencers

L'étude des silencers que nous avons mené a été réalisée dans la lignée cellulaire P5424 (lignée de cellules au stade précoce des lymphocytes T). Ainsi notre support était la régulation des gènes impliqués dans la différenciation des lymphocytes T. Partant de ce postulat, nous avons recherché les candidats silencers possédant plusieurs gènes impliqués dans ce processus dans une fenêtre de 100kB autour de la région silencer. Grâce à cela, nous avons mis en évidence un silencer (DHS-23650) ayant à proximité NFκB et HCST, deux gènes impliqués dans la différenciation des lymphocytes T. Plus étonnant encore, en regardant les données RNA-seq pour ces deux gènes au cours de la différenciation des lymphocytes T, nous avons observé une anti-corrélation entre ces deux gènes, suggérant un changement de cible du silencer basculant d'un gène à l'autre. Expérimentalement, la délétion du DHS-23650 résulte en une surexpression de ces deux gènes (Figure : 3.2).

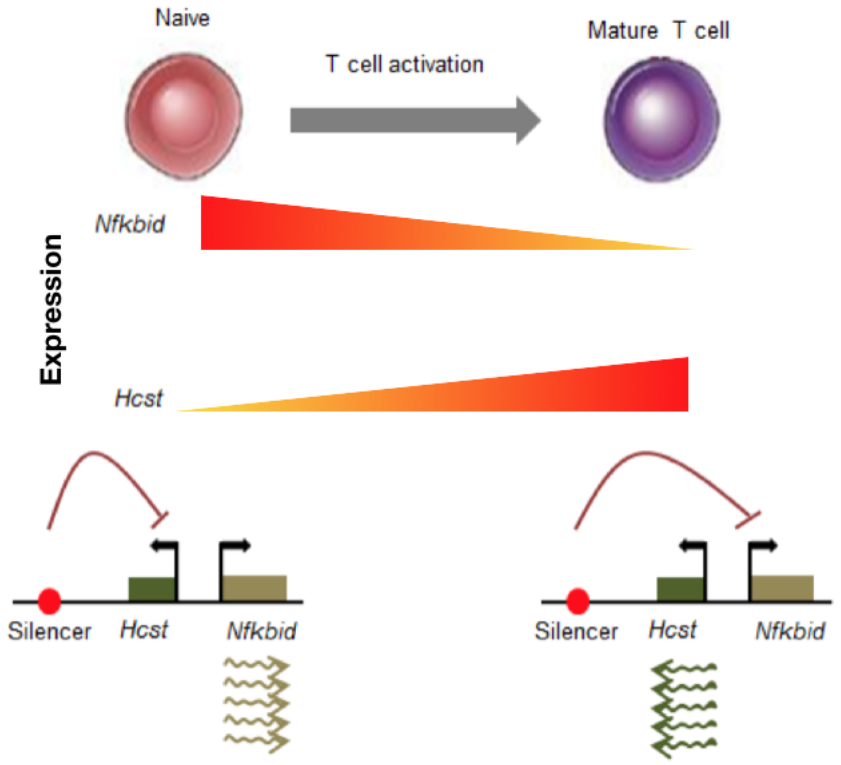


FIGURE 3.2 – Action d'un silencer sur deux gènes impliqués dans le développement des cellules lymphocytaires T. Les gènes présentes un expression anti-corrélées au cours de la différenciation cellulaire et sont sous le contrôle d'un même silencer.

Ce silencer possède des sites de fixation du facteur de transcription ZNF263 répétés en tandem et a été identifié grâce à la condition pPGK, la condition même présentant un enrichissement en séquences répétées. Ce mécanisme n'est pas sans rappeler le comportement de ZEB1, un facteur de transcription possédant un domaine en doigt de zinc (zinc finger) possédant une activité silencer lorsque ces sites de fixation présentent une répétition en tandem [Balestrieri et al., 2018]. On peut suggérer que les silencers ont coopté ces régions répétées afin de renforcer leur pouvoir répresseur, d'autant qu'une corrélation entre le nombre de répétition du site de fixation et l'activité silencer a été mise en évidence. En effet, il a été démontré que les régions répétées du génome tel que les ERV allaient pouvoir être détournées de leur usage afin d'intégrer le

réseau de régulation génique [Bakoulis et al., 2022], ce mécanisme leur permettant de pouvoir se répliquer et proliférer à travers le génome [Fueyo et al., 2022]. ZNF263 appartient à la famille des Krab Zinc Finger Proteins (KZNFP) et il a déjà été démontré qu'il était impliqué dans la répression épigénétique des gènes suppresseurs de tumeurs dans le glioblastome [Yu et al., 2020]. La majorité des KZNFP fixe les éléments transposables tel que les LINE, LTR, SINEs afin de les réprimer et les contenir.

Parmi les types d'éléments répétés enrichis dans les silencers obtenus par pPGK, nous avions également la famille des SINEs. Il s'agit d'un type de rétrotransposons, qui a évolué chez la souris en plateforme de fixation de CTCF [Schmidt et al., 2012]. Ceci est d'autant plus intéressant qu'il a été démontré récemment que la présence d'une séquence SINE possédant une insertion polymorphique était impliqué dans la restriction de la propagation de la marque H3K9ac qui est impliquée dans la transcription active, qui se traduit par une diminution de l'expression des gènes à proximité [Ichiyanagi et al., 2021]. D'une façon plus générale, des données ChIP-seq ont révélé que les SINE sont généralement enrichis aux limites des marques de chromatine active, et que la majorité de ces éléments ne sont pas liés par CTCF [Ichiyanagi et al., 2021]. Une hypothèse serait que le complexe répresseur KRAB empêcherait la fixation de CTCF sur les sites inclus dans les éléments SINEs. En effet, il a été démontré que SETDB1 qui fait partie du complexe KRAB-ZNF contrôle l'accessibilité des sites CTCF dans les éléments répétés SINEs [Gualdrini et al., 2022].

Une autre hypothèse pourrait être la présence du complexe ChAHP (CHD4, ADNP, HP1) qui reconnaît le même motif que CTCF et agit comme un concurrent direct à la liaison de ce dernier aux niveaux des éléments SINE et module ainsi localement l'architecture 3D de la chromatine [Kaaij et al., 2019]. Ce type de processus "anti-boucle" a déjà été décrit chez la Drosophile avec le facteur de transcription Snail. Le répresseur Snail bloque la formation d'interactions entre enhancer et promoteur lorsqu'il est lié à un enhancer distal [Chopra et al., 2012]. Le complexe ChAHP viserait environ 15.000 loci dans le génome des cellules souches embryonnaires de souris (ESCs) menant à une inaccessibilité de la chromatine. La délétion de ADNP (élément clé du complexe ChAHP) ou de son site de fixation résultant en une augmentation de l'expression des gènes cibles [Ostapcuk et al., 2018]. Ces hypothèses représentent autant de pistes à explorer dans le rôle des ZNFP organisés sous formes de séquences répétées.

3.1.4 Approches à haut débit pour l'identification des éléments silencers

Mon projet de thèse était donc d'identifier et caractériser les éléments silencers du génome en se basant sur la technologie du CapSTARR-seq qui a déjà fait ses preuves pour l'identification des enhancers et des E-promoters [Vanhille et al., 2015, Dao et al., 2017, Santiago-Algarra et al., 2021]. Il n'existait pas d'approche fonctionnelle permettant l'identification des silencers à grande échelle. Au cours de ma thèse plusieurs articles ont été publiés pour l'étude des silencers utilisant des technologies à haut débit [Doni Jayavelu et al., 2020, Gisselbrecht et al., 2020, Pang and Snyder, 2020, Ngan et al., 2020] dont nous résumons les résultats dans le tableau 3.1.

Comparée aux autres études ayant pour but l'identification de silencers basées sur des techniques à haut débit, notre approche était basée sur peu d'apriori. La plupart des autres études partaient avec des assomptions fortes en incluant les régions possédant des cofacteurs de transcription répresseur [Gisselbrecht et al., 2020], ou excluaient les régions enrichies en marques enhancers, promoteurs ou insulateurs [Doni Jayavelu et al., 2020]. Il faut cependant souligner l'ingéniosité de toutes ces équipes, ayant permis une avancée considérable dans la caractérisation des silencers, proposant, adaptant, des technologies existantes afin d'identifier ces séquences régulatrices. Certaines de ces études ont exploré les variants présent dans les silencers associé à des pathologies [Doni Jayavelu et al., 2020], d'autres encore ont directement valider leurs silencers *In vivo* [Pang and Snyder, 2020, Gisselbrecht et al., 2020].

Table 3.1 – Études récentes sur les éléments silencers à grande échelle.

Study	Approach	Cell types	Validation Strategy	Comments
Doni Jayavelu et al., 2020	Simple Subtractive Analysis (SSA) approach to select region. MPRA using STARR-seq strategy with SCP1	K562	Luciferase assay, CRISPR/Cas9	They identified 3,001 silencers regions. Based on SSA approach, they select regions lacking promoters, enhancers, insulators marks. They start with the assumption that silencers have different marks than enhancers, promoters and insulators, but it's clear that some silencers share some histone marks with enhancers. This method will may have a low specificity because of incomplete annotations for other genomic elements. It is also questionable that their silencers have similar activity compare to the background regions.
Ngan et al., 2020	ChIA-PET following by ChIP-seq on PCR2 subunits	Mouse embryonic stem cells (mESCs)	CRISPR/Cas9, *In vivo* validation presenting homozygous KO mice	Ngan et al. focused on chromatin loops associated with polycomb repressive complex 2 (PRC2). They identified 13,629 regions showing both anchors PCR2 interactions. Function of PRC2-bound silencers was validated *In vivo* and a decrease of genes expression was observed in a region of 500kb to 1 Mb around the delete PRC2 anchors. Thanks to the *In vivo* validation, they were able to identify silencers essential in mouse development. Depending on the development stages and tissues, some anchors displayed H3K27ac marks specific to enhancers. It's a remarkable work that focused in one specific mechanism that use silencers to repress their targets, the role of the polycomb repressive complex.
Pang and Snyder, 2020	High-throughput ReSE lentiviral screen system. Regions are cloned upstream the EF-1α promoters driving the expression of an apoptotic gene. Silencers will repress the Cas9 and provide a selective advantage	K562, HepG2	Luciferase assay, CRISPR/Cas9	They select accessible chromatin enriched by FAIRE-seq to construct the ReSE screen library. The Cas9 approach was maybe too restrictive to identify silencer genome wide, the screen has considerable background cell survival and a small percentage of the potential fragments was consistently enriched between replicates when fold change was considered. 2,664 potential silencers were identified in K56 and 1,662 in HepG2.

Gisselbrecht et al., 2020	Silencer-FACS-Seq (sFS)	cells isolated from whole *Drosophila* embryos	CRISPR/Cas9	Gisselbrecht et al. used a parallelized reporter assay in whole *Drosophila* embryos. This led to an interesting *In vivo* approach that pinpoint silencer-enhancer bi-functional regulatory elements depending on cellular context. Library of 591 genomic elements based on several criteria (DHS, overlap H3K27me3 marks, enriched in co-transcriptional repressor, etc.). They identify 29 silencers some and find that they approximately act in the same distance range as transcriptional enhancers. This study is really important to pinpoint the tissue-specificity of regulatory elements.
Wei et al., 2022	Hi-C on accessible regulatory DNA(HiCAR) which use Tn5 transposase and chromatin proximity ligation	Human embryonic stem cells (hESCs), Human lymphoblast oid GM12878	Luciferase assay	The study was originally design to assess enhancer and silencer activity based on a new method: HiCAR. They show that HiCAR is a robust method able to identifies high-resolution chromatin contacts with high efficiency over all distance ranges and with a low cell input. They identified 6,662 interactions of H3K27me3 marks (poised interaction) in hESCs and 1,116 in GM12878. They identified promoter-promoter interactions acting as silencer.
Hussain S, Sadouni N, et al. In preparation	CapSTARR-seq	P5424	Luciferase assay, CRISPR-Cas9	28,055 DHS from P5424 cell line were captured and clone into 3 different STARR-seq vectors where promoters were changed to assess best silencer signal. 1 249 silencers were identified using SCP1 strategy, 672 using pPGK, 413 using pR-Eα. Subregion with main silencer activity was flagged as 'core-silencer'.

3.1.5 Améliorations et limites du STARR-seq

Cependant, même si le STARR-seq permet l'analyse de dizaines de milliers de séquences en parallèle, il présente des limites. En effet, la séquence testée est isolée de son contexte biologique, la privant d'éléments nécessaires à son bon fonctionnement. Ainsi, on peut supposer aisément que des faux positifs ainsi que des faux négatifs sont présents dans les silencers identifiés. En effet, on imagine que dans un contexte cellulaire *in vivo* d'autres éléments vont intervenir afin de moduler l'activité silencer de ces régions. Autres points importants, la chromatine n'est pas prise en compte. Or comme nous avons pu le voir, un des mécanismes essentiels de la régulation génique est l'accessibilité à la chromatine, les complexes répressifs HUSH [Tchasovnikarova et al., 2015] et KRAB-ZNF [O'Geen et al., 2007, Ecco et al., 2017] jouant sur l'accessibilité à la chromatine.

Comme nous l'avons présenté, nous avons utilisé une variante du STARR-seq afin de réduire la complexité de la librairie. Nous nous sommes basés sur les régions hypersensible à la DNase I

(DHS). Cette approche est discutable et peut présenter un biais dans le set de silencers obtenus. Cependant, nous sommes partis du postulat que les silencers actifs devaient être accessible aux facteurs de transcriptions répresseurs afin de pouvoir inhiber leurs gènes cibles. Cette approche a été validée par une équipe ayant cataloguée les DHS dans le génome humain et recherchaient simplement les régions qui chevauchaient la marque H3K27me3 obtenue par ChIP-seq à travers plusieurs lignées cellulaires [Huang et al., 2019]. Ils ont ensuite examiné la corrélation entre la présence ou l'absence de ce signal combiné et le niveau d'expression des gènes voisins. Les H3K27me3-DHS qui étaient corrélés négativement avec l'expression des gènes à proximités étaient significativement enrichis pour plusieurs caractéristiques compatibles avec les silencers telles qu'un enrichissement en facteurs de transcription répresseurs tels que CTCF, MECOM, SMAD4, SNAI3 et une déplétion en facteurs de transcription actifs ; de plus les auteurs ont validé expérimentalement quelques-uns de ces silencers par un test de gène rapporteur. Ainsi, on peut considérer la stratégie basée sur l'étude des DHS comme un très bon point de départ, mais limité et présentant un biais. On peut supposer aisément que de nombreux silencers inaccessibles n'ont pu être identifiés par cette approche. Une solution serait d'effectuer un STARR-seq complet du génome (whole genome STARR-seq) afin de ne pas sélectionner une fraction du génome. En dehors du coût prohibitif de cette technologie, un de ses inconvénients est que cette technologie présente beaucoup de bruit de fond, rendant l'identification d'enhancer déjà compliquée. Or, les silencers représentant une baisse du signal, on peut facilement imaginer que l'identification de silencer serait encore plus ardue. Une possibilité serait alors d'utiliser un promoteur fort tel que pPGK afin d'avoir un fort signal de bruit de fond et ainsi identifier les régions réduisant ce signal fort.

Une amélioration que l'on pourrait apporter au STARR-seq serait l'ajout des séquences UMI aux vecteurs. Ainsi chaque séquence posséderait son propre 'code barre', réduisant l'impact des artefacts PCR. Une approche alternative consisterait à développer un rapporteur intégratif spécifique basée sur une stratégie décrite par [Weingarten-Gabbay et al., 2019]. Brièvement, une librairie génomique telle que celle générée pour le STARR-seq pourrait être intégrée à un site spécifique en induisant une cassure double brin à l'aide de Zinc Finger Nucleases (ZFNs) suivie de l'intégration génomique d'une cassette reporter par recombinaison homologue.

L'idéal serait de tester directement des séquences silencers synthétiques. Pour cela, il faudrait suffisamment de données sur les silencers. Cela a été proposé chez les enhancers par une étude récente couplant Whole genome STARR-seq et deep learning [de Almeida et al., 2021]. On peut espérer que cela se produira prochainement, le nombre de données collectées sur les silencers ayant augmenté ces deux dernières années, comme en atteste la création d'une base de données sur les silencers : SilencerDB [Zeng et al., 2021].

3.2 Approches bio-informatiques

3.2.1 Méthode statistique pour l'identification des silencers

Concernant le pipeline d'analyse de données CapSTARR-seq : **STARR Track**, même si ce dernier a permis d'identifier des silencers à grande échelle dont un certains nombres ont pu être validé par des méthodes indépendantes (Luciferase assay, CRISPR), il comporte des limites. Dans notre approche, les régions étudiées sont des DHS. La partie statistique a été développée en étroite collaboration avec Dominic Van Essen à l'IRCAN, Nice. Un grand avantage de l'outil STARR Track est la visualisation de l'activité des régions ainsi que de l'activité par fragment. Ceci permet une visualisation rapide des résultats, de visualiser un chevauchement avec d'autres éléments, etc. C'est d'ailleurs grâce à cette approche que nous avons remarqué une présence abondante de motifs ZNF263 répétés en tandem. Afin d'optimiser l'identification des silencers, une analyse statistique est en cours de développement. Notre modèle tient compte des spécificités

du CapSTARR-seq. Dans notre approche, chaque fragment est considéré comme indépendant.

Pour notre modèle, nous posons d'emblée deux hypothèses, celle que l'activité du fragment est égale à l'activité de la région observée et que l'activité du fragment est égale à une valeur fixée pour la région. Par défaut nous avons pris comme seuil un fold change égal à 1 qui correspond à des régions sans activité.

Par la suite pour chaque hypothèse, nous avons calculé une p-valeur basée sur une négative binomiale pour chaque fragment. On fait cela pour les fragments qui chevauchent les régions d'intérêts mais aussi pour les régions capturées aléatoirement. En effet la capture n'est pas efficace à 100%, cela n'existe pas en biologie. Des régions aléatoires dans le génome sont capturées. Nous nous servirons de ces régions capturées aléatoirement dans notre approche statistique.

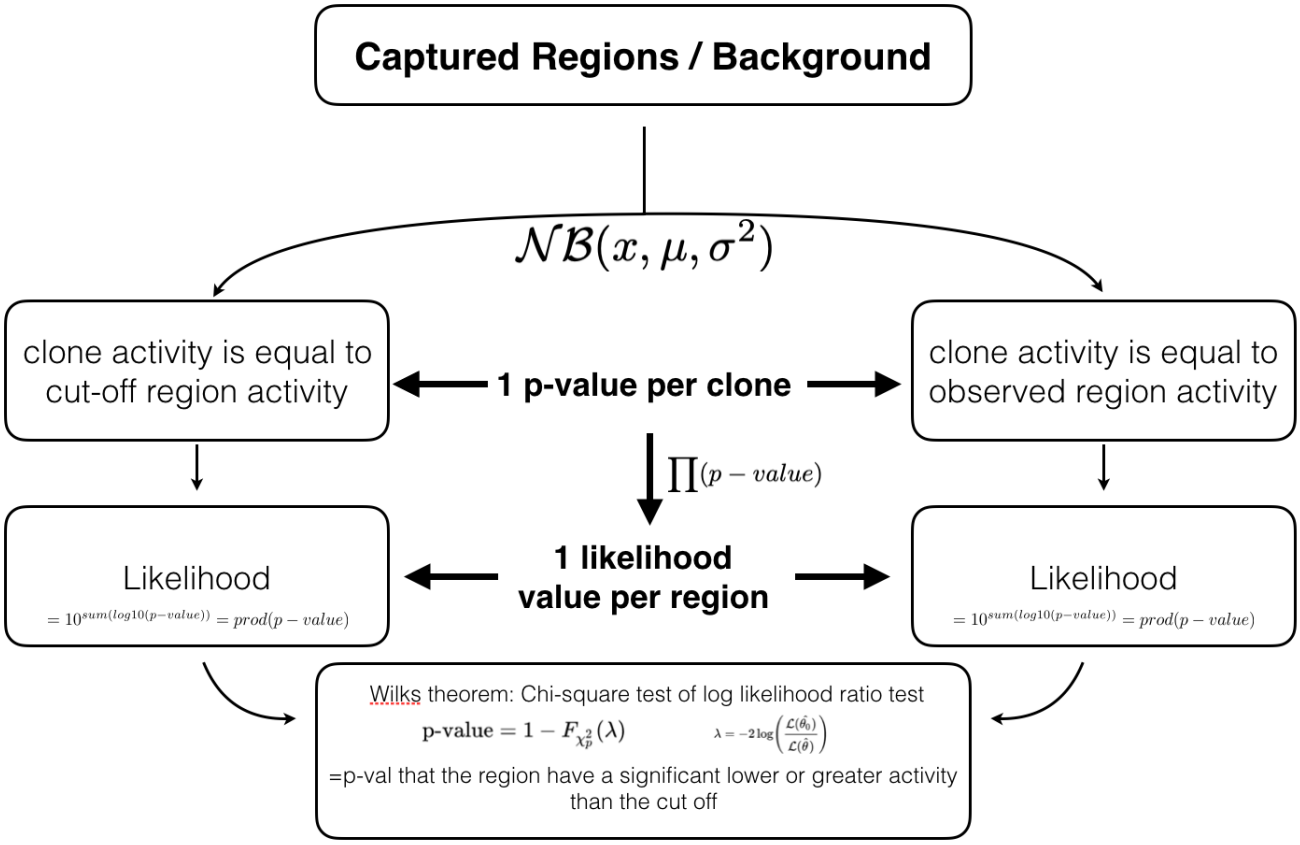


FIGURE 3.3 – Méthode statistique pour l'identification des silencers.

Pour chaque région, nous calculons ensuite la likelihood qui correspond à la somme des p-valeurs des fragments appartenant à la région. Théoriquement on ne peut pas obtenir un maximum likelihood supérieur à l'hypothèse basée sur l'activité observée de la région. Par la suite on effectue un log likelihood ratio test, où l'on divise le log(likelihood) de l'hypothèse nulle par le log(likelihood) de l'hypothèse alternative correspondant à la contrainte sur l'activité de la région. Le log likelihood ratio test est par la suite approximé en utilisant un test chi-square d'après le théorème de Wilk's. En résulte une p-valeur correspondant à la probabilité que la région ait une activité statistiquement différente du seuil fixé (Figure :3.3). L'avantage du seuil ici est de pouvoir déplacer le curseur, en fonction de la stringence qu'on souhaite appliquer. Par la suite, la FDR est calculée grâce aux régions capturées aléatoirement et hors de nos régions d'intérêts. Une méthode basée sur l'orientation du brin chromosomique est également en développement afin d'identifier si ce paramètre à un impact dans l'activité des éléments cis-régulateurs, notamment des silencers. Cette méthode statistique n'est pas totalement automatisée et pas encore incluse en routine dans le pipeline d'analyse de données CapSTARR-seq. Dans l'idéal, si l'outil STARR

Track est publié, il s'agirait d'uniformiser le code sous python et d'inclure cette méthode statistique.

3.2.2 Méthodes existantes pour le traitement de données STARR-seq

Très peu d'outils existent afin de traiter des données STARR-seq. Durant ma thèse, quelques outils ont été publiés afin de traiter ces données, ils sont comparés dans le tableau suivant (Tableau : 3.2). Seulement un de ces outils a été développé directement pour identifier les silencers. La plupart ont été conçue afin d'identifier les enhancers. Cela confirme la tendance évoquée sur la rareté des études menées sur les silencers. La plupart des méthodes permettant l'analyse des données STARR-seq sont basées sur une négative binomiale qui semble être un bon modèle pour ces données.

Table 3.2 – Méthode d'analyses des données STARR-seq

Study	Model	Comments
Vanhille et al. 2015	Lorenz curve (inflexion point)	This model was developed in our lab to process CapSTARR-seq data in order to identify enhancers. CapSTARR-seq is based on capture of regions of interest. Activity of the region is computed by comparing the coverage in cDNA vs library. This method doesn't take in account the random captured regions, restrict to captured regions, only consider region and not each fragment individually.
Lee et al. 2020	Negative binomial	They identified the negative binomial as the best model to fit STARR-seq data. They consider only coverage instead of each fragment individually. Also, their approach identifies only enhancers.
Kim et al. 2021	Generalized Linear regression Model (GLM) with covariates to model DNA structure	The model correct bias in STARR-seq signal. Biggest biases in PCR amplification had some of the strongest impacts on sequence biases in STARR-seq. Design for identification of enhancer but able to detect repressive elements.
He et al. 2022	cumulative density function (CDF) of negative binomial distribution (NBD)	Specifically design for identification of silencers. They look for nucleotides with very low fold change and compute p-value. Then they create a window of 601bp next to it. They compute p-value at each position in the window and filter window with low score. Then, they compare the similarity between the curves of reporter cDNA and the input insert DNA reads, and discard any window with a curve similarity score higher than the arbitrary threshold

Dans notre cas nous avions testé préalablement un modèle basé sur la loi de Poisson. Le test de Poisson suppose que nous connaissions le nombre moyen exact attendu de comptages d'ADNc pour chaque fragment. Toutefois, cette supposition n'est pas exacte car nous n'avons qu'une estimation basée sur les comptes observés dans la librairie (input). De plus, cette estimation non plus n'est pas totalement exacte, car elle suppose que l'expérience est parfaite. Si nous séquencions une seconde fois, nous devrions obtenir exactement la même valeur à chaque position. Or, il est attendu d'obtenir des comptages de la librairie légèrement différents à chaque séquençage. La conséquence est que la valeur de comptages d'ADNc que nous observons pour tous les fragments avec une valeur particulière dans la librairie aura une distribution plus "large" que celle prédite par Poisson - c'est un phénomène appelé "sur dispersion". Par conséquent, les p-valeurs calculées par Poisson pour chaques fragments seront systématiquement trop petites. Lee et al. [Lee et al., 2020] ont effectué une comparaison similaire où la négative binomiale est le modèle le plus adapté aux données STARR-seq.

3.3 Étude des silencers dans un contexte pathologique

3.3.1 Dérégulation des silencers et complexes répresseurs dans les mécanismes pathologiques

Une application directe aux découvertes effectuées en génomique est la compréhension des pathologies afin de proposer des pistes thérapeutiques. Ce n'est que récemment que les silencers ont réellement commencé à être documentés. Ainsi, il existe peu de pathologies directement liées à ces régions inhibitrices. Parmi celles connues, on peut notamment citer une région régulatrice du gène EDN1 [Gupta et al., 2017]. Un variant génétique associé aux maladies cardiovasculaire a été identifié dans cette étude. Il a été démontré que la suppression de la région génomique entourant le variant génétique a entraîné la surexpression du gène EDN1 situé à 600 kb du variant, suggérant que cette région pourrait fonctionner comme un élément silencer. On peut également citer la plateforme de recrutement du répresseur ZEB1 qui présente une configuration en STR [Balestrieri et al., 2018]. Cette étude a montré que l'organisation en tandem de motifs ZEB1 pouvaient fonctionner comme des éléments silencers. Il a été démontré également que la suppression d'un tandem ZEB1 pouvait conduire au développement d'un cancer quasi-mésenchymateux. En effet la séquence reconnue pour la fixation du facteur de transcription répresseur ZEB1 est impliquée dans le maintien des caractéristiques mésenchymateuses et est nécessaire au maintien de l'identité épithélial [Balestrieri et al., 2018].

Des études commencent à s'intéresser aux variants nucléotidiques (SNP) associés à des pathologies au sein des silencers. Huang et al. [Huang et al., 2019] ont identifié des silencers en régions ouvertes (DHS) portant des marques H3K27me3. Ils ont identifié que ces silencers étaient susceptibles de transporter des SNP GWAS et qu'environ 7% des silencers possédaient des SNP associés à des pathologies. Jayavelu et al. [Doni Jayavelu et al., 2020] ont effectué le même type d'analyse. Ils souhaitaient savoir si les silencers qu'ils ont identifié sont enrichis pour des variants associés à des pathologies. Ils ont découvert 57,961 SNP associés à 2,214 pathologies présents au niveau des éléments silencers sur tous les types cellulaires. Ainsi on comprend que l'implication des silencers dans les pathologies est très largement sous-estimés.

Par ailleurs, les dysfonctionnements des complexes répresseurs impliqués dans des processus pathologiques sont mieux connus. Par exemple, une expression accrue de REST dans le cerveau est associée à une incidence moindre de la maladie d'Alzheimer, inversement, une déficience en REST est associée à un risque accru de la maladie d'Alzheimer. REST protège puissamment les neurones du stress oxydatif et de la toxicité de la protéine β-amyloïde, la suppression conditionnelle de REST dans le cerveau de la souris entraînant une neurodégénérescence liée à l'âge [Lu et al., 2014]. On peut également citer le complexe répresseur HUSH, dont une hyper

activation via MORC2 va entraîner la Maladie de Charcot-Marie-Tooth [Sacoto et al., 2020]. Il a également été montré qu'une dérégulation de nombreux KRAB-ZFP est altérée dans plusieurs types de tumeurs, dans lesquelles ils peuvent agir comme oncogènes ou suppresseurs de tumeurs [Sobocińska et al., 2021].

3.3.2 Identification des silencers impliqués dans les LAL-T via crible CRISPRi

De façon générale il est possible d'envisager qu'une dérégulation de l'activité des silencers puisse être impliquée directement dans le développement de cellules cancéreuses avec des mécanismes opposés à ceux généralement proposés pour la dérégulation des enhancers (Figure : 3.4). L'inactivation d'un silencer pourrait conduire à l'activation d'un oncogène alors que le gain de fonction d'un silencer pourrait conduire à la répression d'un suppresseur de tumeurs. De façon similaire à la dérégulation des enhancers dans le cancer [Bradner et al., 2017], les causes moléculaires conduisant à la dérégulation des silencers pourraient être due à des anomalies génétiques (translocation, variations structurales ou mutations de l'élément régulateur) ou épigénétiques (surexpression de facteurs de transcription ou mutation de complexes régulateurs de la chromatine).

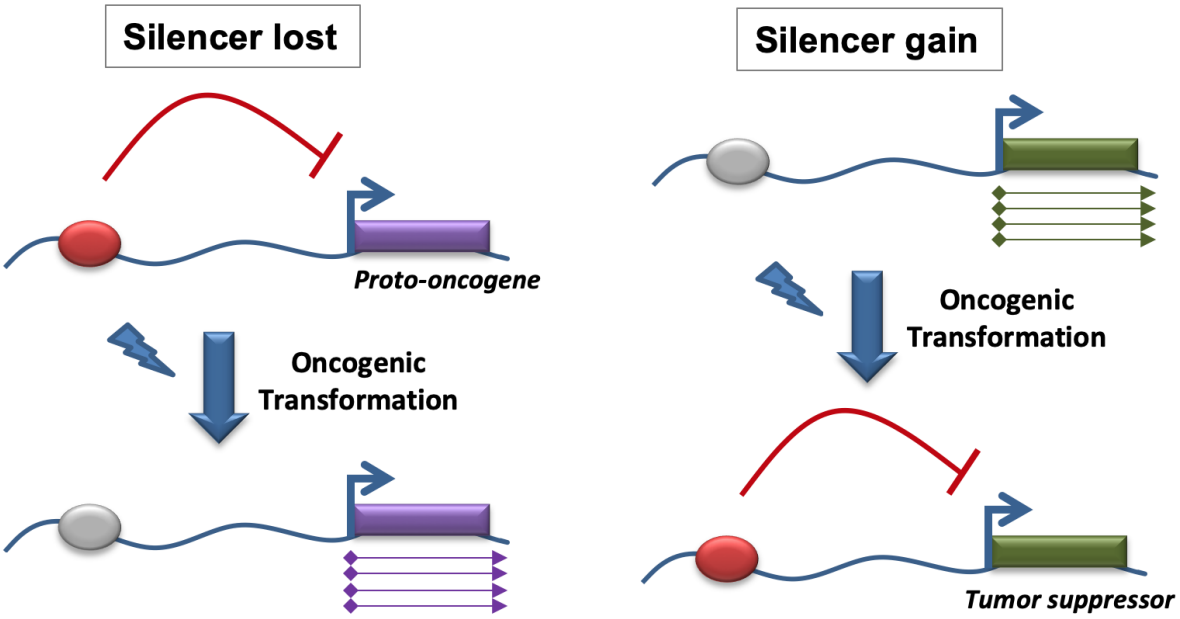


FIGURE 3.4 – On peut imaginer deux types de transformation oncogénique : une perte d'activité silencer qui résulterait en l'augmentation de l'activité des proto-oncogènes ou un gain d'activité silencer qui supprimerait l'activité d'un gène suppresseur de tumeur.

La technologie de crible CRISPR peut être utilisée afin de mettre en évidence des régions cis-régulatrices. Dans l'article publié par [Fulco et al., 2016], ces derniers ont pu étudier les régions régulatrices des locus des gènes MYC et GATA1 mettant en évidence 7 enhancers régulant ces gènes, mais également 2 silencers. Des études plus récentes, utilisant des approches CRISPRi, ont pu cribler des milliers d'éléments enhancers pour leur activité oncogénique [Lopes et al., 2021, Benbarche et al., 2022]. Il serait intéressant de mettre en place une stratégie couplant les différentes approches que nous avons pu développer lors de mon projet de thèse afin d'identifier les silencers dérégulés dans le cancer. L'objectif serait d'identifier des silencers potentiels dans des lignées cancéreuses en utilisant l'approche STARR-seq que nous avons développé et ensuite valider le rôle oncogénique de ces éléments par une approche CRISPRi.

Nous utiliserions comme modèle les leucémies aiguës lymphoblastiques de type T (LAL-T), car l'équipe dispose de données épigénomique et de modèles cellulaires adaptés. Ainsi il s'agirait :

1. Récupérer les régions ouvertes de la chromatine en passant par de l'ATAC-seq par exemple.
2. Effectuer des expériences de STARR-seq sur des lignées cellulaires modèle de la LAL-T (notamment CCRF-CEM et Jurkat) afin d'identifier les régions silencers. Une autre approche serait de coupler directement ATAC et STARR-seq [Hansen and Hodges, 2022].
3. Générer une librairie de sgRNA ciblant les régions silencers identifiées.
4. Effectuer un crible CRISPRi et identifier les guides ayant un impact sur la prolifération et la survie cellulaire.
5. Identifier les gènes cibles par une approche Single-cell CRISPR screening.

Cette analyse permettrait notamment de mettre en évidence la régulation des gènes suppresseurs de tumeurs par les silencers (Figure : 3.4).

Les silencers sont maintenant connus depuis une trentaine d'années, et pourtant ce n'est que très récemment qu'ils ont commencé à être caractérisés. D'après les études récentes et mon projet de thèse, de nombreuses questions restent en suspens. Existent-ils une signature épigénétique caractéristique des silencers ? Une réponse a été apportée, mais celle-ci ne représente très certainement qu'un échantillon de la réalité. Comment vont se comporter les silencers dans l'espace ? L'hypothèse proposée précédemment dans la discussion, sur la compétition du complexe ChAHP sur les sites de recrutement de CTCF au sein des éléments transposables agissant comme 'anti-boucle', mériterait une plus ample investigation. Comment ont évolué les silencers ? Certains sont issus d'éléments transposables. Il serait intéressant d'établir une phylogénie, ces séquences étant des éléments d'études phylogénétiques. Et bien-sûr il s'agirait également de creuser le rôle des silencers dans les pathologies, car leurs impacts sont très certainement sous-estimés.

4 Annexes

4.1 Identification d'enhancers impliqué dans la régulation du gène *ikzf1* via test de gène rapporteur à grande échelle

PLOS ONE

RESEARCH ARTICLE

Integration of high-throughput reporter assays identify a critical enhancer of the *Ikzf1* gene

Jaafar Alomairi[1,2], Anne M. Molitor[3,4,5,6], Nori Sadouni[1,2], Saadat Hussain[1,2], Magali Torres[1,2], Wiam Saadi[1,2], Lan T. M. Dao[1,2¤], Guillaume Charbonnier[1,2], David Santiago-Algarra[1,2], Jean Christophe Andrau[7], Denis Puthier[1,2], Tom Sexton[3,4,5,6], Salvatore Spicuglia[1,2]*

1 Aix-Marseille University, Inserm, TAGC, UMR1090, Marseille, France, **2** Equipe Labélisée Ligue Contre le Cancer, Marseille, France, **3** Institute of Genetics and Molecular and Cellular Biology (IGBMC), Illkirch, France, **4** CNRS UMR7104, Illkirch, France, **5** INSERM U1258, Illkirch, France, **6** University of Strasbourg, Illkirch, France, **7** Institut de Génétique Moléculaire de Montpellier, Univ Montpellier, CNRS, Montpellier, France

¤ Current address: Vinmec Research Institute of Stem cell and Gene technology (VRISG), Hà Nội, Vietnam
* salvatore.spicuglia@inserm.fr

Citation: Alomairi J, Molitor AM, Sadouni N, Hussain S, Torres M, Saadi W, et al. (2020) Integration of high-throughput reporter assays identify a critical enhancer of the *Ikzf1* gene. PLoS ONE 15(5): e0233191. https://doi.org/10.1371/journal.pone.0233191

Editor: Charalampos G Spilianakis, University of Crete & IMBB-FORTH, GREECE

Received: January 28, 2020

Accepted: April 29, 2020

Published: May 26, 2020

Data Availability Statement: ChIP-seq and 4C-seq data described in this study are available in GEO database under the accession number GSE147234 (http://www.ncbi.nlm.nih.gov/geo/).

Funding: We thank the Transcriptomics and Genomics Marseille-Luminy (TGML) platform for sequencing the ChIP-seq samples and the Marseille-Luminy cell biology platform for the management of cell culture. Sequencing of 4C samples was performed by the IGBMC GenomEast platform. TGML and GenomEast platforms are

Abstract

The *Ikzf1* locus encodes the lymphoid specific transcription factor Ikaros, which plays an essential role in both T and B cell differentiation, while deregulation or mutation of IKZF1/*Ikzf1* is involved in leukemia. Tissue-specific and cell identity genes are usually associated with clusters of enhancers, also called super-enhancers, which are believed to ensure proper regulation of gene expression throughout cell development and differentiation. Several potential regulatory regions have been identified in close proximity of *Ikzf1*, however, the full extent of the regulatory landscape of the *Ikzf1* locus is not yet established. In this study, we combined epigenomics and transcription factor binding along with high-throughput enhancer assay and 4C-seq to prioritize an enhancer element located 120 kb upstream of the *Ikzf1* gene. We found that deletion of the E120 enhancer resulted in a significant reduction of *Ikzf1* mRNA. However, the epigenetic landscape and 3D topology of the locus were only slightly affected, highlighting the complexity of the regulatory landscape regulating the *Ikzf1* locus.

Introduction

Cell-type specific regulation of gene expression requires the activation of promoters by distal genomic elements defined as enhancers. The classical view of enhancer function is that they contribute to increasing the overall level of gene expression by inducing transcription from associated promoters [1]. Complex gene regulation is mediated by the association of clusters of enhancers, also called super-enhancers [2]. Whether the individual components (i.e. single enhancers) synergistically contribute to transcription regulation of their target genes or have distinct specialized functions has been a matter of debate [2–5].

member of the France Genomique consortium (ANR-10-INBS-0009). Work in the laboratory of S. S. was supported by recurrent funding from INSERM and Aix-Marseille University and specific grants from the Ligue contre le Cancer (Equipe Labellisée LIGUE 2018), the Agence Nationale pour la Recherche, ANR (ANR-17-CE12-0035; ANR-18-CE12-0019), Cancéropôle PACA, Institut National contre le Cancer (PLBIO018-031 INCA_12619), the Excellence Initiative of Aix-Marseille University - A*Midex, a French "Investissements d'Avenir" programme (ANR-11-IDEX-0001-02). Work in the lab of TS is supported by funds from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Starting Grant 678624 - CHROMTOPOLOGY), the ATIP-Avenir program, and the grant ANR-10-LABX-0030-INRT, a French State fund managed by the Agence Nationale de la Recherche under the frame program Investissements d'Avenir ANR-10-IDEX-0002-02. AMM is supported by funds from IDEX (University of Strasbourg) and the Institut National du Cancer. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist

With the increasing awareness of the important role of enhancers in normal development as well as in disease, there is strong scientific interest in identifying and characterizing these elements. However, few predicted enhancer elements have been shown to affect transcription of their endogenous genes or to alter phenotypes when disrupted, highlighting the need to integrate different epigenomic resources and functional assays to identify critical distal regulatory elements [6]. Although putative enhancers can be identified genome-wide based on chromatin accessibility or histone modifications [7], these approaches do not provide direct proof of enhancer function. Recent developments of functional high-throughput assays have enabled quantitative measurements of enhancer activity of thousands of regulatory elements in parallel, providing a straightforward approach to prioritize *bona fide* enhancers [8]. In particular, a common observation of high-throughput assays based on massively paralleled reporter assays [9–14] or CRISPR-based screens [15, 16] is that many predicted enhancer regions do not show enhancer activity in reporter assays or after CRISPR deletion. Therefore, it is crucial to experimentally assess whether genomic regions function as *bona fide* enhancers in living cells.

Ikaros is a lymphoid specific transcription factor that plays a major role in both T and B cell differentiation [17, 18]. During T cell differentiation Ikaros is required for proper gene regulation during the $CD4^-CD8^-$ (double-negative; DN) to the $CD4^+CD8^+$ (double-positive; DP) transition (also called b-selection) mainly by recruiting chromatin repressors [19, 20] and silencing Notch1 target genes [20–23]. Importantly, Ikaros deregulation or mutation plays an important role in leukemia [24–33]. In mouse and human, Ikaros is encoded by the *Ikzf1* gene and is known to harbor several transcript isoforms playing different regulatory roles [34–38]. Several potential regulatory regions have been identified in the proximity of the *Ikzf1* locus, suggesting a complex network of regulatory elements are required to drive Ikaros expression during hematopoiesis and lymphocyte maturation [39–41]. To gain insight into the regulation of *Ikzf1* locus in T cell precursors, we have integrated data from high-throughput reporter assays, chromatin modifications, binding of key T cell transcription factors as well as genomic interactions. We prioritized an enhancer located 120 kb upstream of *Ikzf1* and studied the functional role of this regulatory element.

Material and methods

Cell culture

P5424 cell line [42] was kindly provided by Dr. Eugene Oltz, Washington, USA and was cultured as described previously [14]. Cells were passed every 2–3 days and routinely tested for mycoplasma contamination, and maintained in RPMI medium (Gibco) supplemented with 10% FBS (Gold, PAA) at 37 ˚C, 5% CO2. J1 mouse embryonic stem (ES) cells were grown on gamma-irradiated mouse embryonic fibroblast cells under standard conditions (4.5 g/L glucose-DMEN, 15% FCS, 0.1 mM non-essential amino acids, 0.1 mM beta-mercaptoethanol, 1 mM glutamine, 500 U/mL LIF, gentamicin), then passaged onto feeder-free 0.2% gelatin-coated plates for at least two passages to remove feeder cells before 4C.

Isolation of DN3 thymocytes

Thymuses from 4–5 weeks old c57/Bl6 mice were dissected and homogenized in cold PBE (PBS with 0.5% BSA and 2 mM EDTA), before incubation for 30 min at 4 ˚C with rat anti-CD4 and rat anti-CD8 anti-sera (gift from Susan Chan) and depletion of DP thymocytes with sheep anti-rat IgG magnetic beads (Invitrogen). $1x10^8$ cells were resuspended in 300 μl PBE plus 3 μl anti-mouse CD4-FITC, 3 μl anti-mouse CD8a-FITC, 3 μl anti-mouse CD3e-FITC, 3 μl anti-mouse B220-FITC, 3 μl anti-mouse CD11b-FITC, 3 μl anti-mouse Ly-6G(Gr-1)-FITC, 3 μl anti-mouse NK1.1-FITC, 6 μl anti-mouse CD44-PE, and 6 μl anti-mouse CD25-PE

(all eBioscience antibodies), and incubated on ice for 10 min before washing in PBE. DAPI was added to a final concentration of 100 ng/ml and live DN3 cells were purified by FACS (DAPI-negative, FITC-negative, APC-negative, PE-positive) before immediate fixation for 4C.

CRISPR/Cas9 genome editing

The targeted enhancer regions were defined by the peaks of CapStarr-seq and DNase-seq which bind the 6 TFs. Two gRNAs were designed at each end of the targeted region by CRISPR direct tool [43]. The gRNAs were cloned into the gRNA cloning vector (Addgene #41824) as previously described [44]. Two million cells were transfected with 3μg of each gRNA vector and 3μg of Cas9 vector (Addgene #41815) using the Neon transfection system (Life Technologies). After 3 days of transfection, the bulk transfected cells were plated in 96-well plates at the limiting dilution (0.5 cell per 100 μl per well) for clonal expansion. After 10–14 days, individual cell clones were screened for homozygous allele deletion by direct PCR using Phire Tissue Direct PCR Master Mix (Thermo Scientific) following the manufacturer's protocol. Forward and reverse primers were designed bracketing the targeted regions allowing the detection of knockout and wild-type alleles. Clones with homologous allele deletion were considered if having at least one expected deletion band and no wild-type band. The gRNAs and primers are listed in the S1 Table.

Gene expression analysis

Total RNA was isolated using the RNeasy kit (Qiagen). RNA samples (1 μg) were reverse-transcribed into cDNA using Superscript VILO Master Mix (Thermo Scientific). The quantitative PCR was performed using power SYBR Master Mix (Thermo Scientific) on a QuantStudio 6 Flex Real-Time PCR System. Primer sequences are listed in S2 Table. Gene expression was normalized to that of *Rpl32*. The relative expression was calculated by delta Ct method and all the shown data reported from the fold change over the control. For each cell clone, the Student's *t*-test was performed (unpaired, two-tailed, 95% confidence interval) from 3 biological replicates of independent cDNA preparations. Data are represented with standard deviation (s.d). For RT-qPCR, 1/20 of synthesized cDNA was used as template for one reaction; PCRs were performed with Phusion polymerase (Thermo Scientific), Tm = 60 ˚C, 35 cycles. RNA-seq from the P5424 cell line treated with either DMSO or PMA/ionomycin was published before [45] and was retrieved from GEO (GSE120655).

PMA/ionomycin induction

1 x 10^6 cell/ ml of P5424 cells (wt and ΔIkE120) were stimulated for 6 hours with 20 μg/ml of PMA plus 0.5 μg/ml of ionomycin in 6-well plate of 3 independent experiments. Then, total RNA was prepared from non-stimulated or stimulated cells using the RNeasy kit (Qiagen) as recommended by the manufacturer.

Chromatin immunoprecipitation-sequencing (ChIP-seq)

Total 40 x10^6 of wt and ΔIkE120P5424 cells were crosslinked in 1% formaldehyde for 10 min at 20 ˚C, followed by quenching with glycine at a final concentration of 250 mM. Pelleted cells were washed twice with ice-cold PBS, and then re-suspended in lysis buffer (20 mM Hepes PH 7.6, 1% SDS, 1X PIC) at final cell concentration of 15 x 10^6 cells/ml. Chromatin was sonicated with Bioruptor (Diagenode) to an average length of 200–400 bp (5 pulses of 30 sec ON and 30 sec OFF). An aliquot of sonicated cell lysate equivalent to 0.5 x 10^6 cells was diluted with SDS-free dilution buffer (1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris pH 8.0, 167 mM NaCl)

for single immunoprecipitation. Specific antibody to H3K27ac (C15410196; Diagenode) (1 μg per ChIP) and proteinase inhibitor cocktail were added to the lysate and rotated overnight at 4 ˚C. On the next day, Protein A magnetic beads (Invitrogen) were washed twice with dilution buffer (0.15% SDS, 1% Triton X-100,1.2 mM EDTA, 16.7 mM Tris pH 8.0, 167 mM NaCl and 0.1% BSA) and added to the lysate and rotated 1 hour at 4 ˚C. Then, beads were washed with each of the following buffers: once with Wash Buffer 1 (2 mM EDTA, 20 mM Tris pH 8.0, 1% Triton X-100, 01% SDS, 150 mM NaCl), twice with Wash Buffer 2 (2 mM EDTA, 20 mM Tris pH 8.0, 1% Triton X-100, 0.1% SDS, 500 mM NaCl), twice with Wash Buffer 3 (1 mM EDTA, 10 mM Tris pH 8.0). Finally, beads were eluted in Elution buffer (1% SDS, 0.1 M $NaHCo_3$) and rotated at RT for 20 min. Eluted materials were then added with 0.2 M NaCl, 0.1 mg/ml of proteinase K and incubated overnight at 65 ˚C reverse cross-linking, along with the untreated input (10% of the starting material). The next day, DNA was purified with QIAquick PCR Purification Kit (Qiagen) and eluted in 30 μl of water. At least 1 ng of ChIP was used for library preparation. Libraries for ChIPs against H3K27ac was prepared according to Illumina ChIP--Seq protocol and sequenced on a Nextseq500 (Illumina) according to the manufacturer's instructions. ChIP-seq was processed as described previously [14] and RPKM-normalized using deepTools bamCoverage [46] before visualization as heatmaps with deepTools plotHeatmap and as genome tracks with the IGV genome browser [47]. Differential H3K27ac signal between wt and ΔIkE120 cells was generated with the IGV genome browser.

CapStarr-seq

CapStarr-seq data in P5424 (two replicates) and NIH3T3 cell lines with a selected set of DHSs were previously published [14] and processed data retrieved from GEO (GSE60029). As described previously, the enhancer activity was computed by calculating the ratio (fold change; FC) of FPKMs (Fragment Per Kilobase per Million mapped reads) between the CapStarr-seq signal over the plasmid library (input). DHSs with a FC between 1.5 and 3 were labelled as 'weak enhancer' and DHSs with a FC equal or higher than 3 were labelled as 'strong enhancer'. To visualize the CapStarr-seq signal per individual cloned fragments we generated a bed file with an RGB color code proportional to the enhancer activity.

Conservation of *Ikzf1* enhancers

Mammalian conservation of *Ikzf1* enhancers and coordinates of human orthologues regions were assessed using the UCSC genome browser [48]. ChIP-seq data for H3K27ac in developing human thymocytes were obtained from the Blueprint consortium ([49]; http://dcc.blueprint-epigenome.eu; S3 Table).

Analyses of ImmGen dataset

ATAC-Seq normalized signals for cell types of hematopoietic lineages were retrieved from the ImmGen project [50, 51] as bigwig files (GSE100738). Their coverage within IkE120 locus (chr11:11,564,323–11,564,574, mm10) was extracted using deepTools multiBigwigSummary (http://doi.org/10.1093/nar/gkw257) and compared to *Ikzf1* DeSeq2 normalized RNA-Seq signal of corresponding cell types obtained from the ImmGen portail (www.immgen.org).

Hi-C and virtual 4C

Raw Hi-C data from primary DP thymocytes were taken from Hu et al. [52] and processed with FAN-C [53], entailing iterative mapping to the mm9 genome assembly with bowtie2, filtering self-ligation events and PCR duplicates, binning the data to 10 kb bins and balancing

the chromosome-wide matrices with the Knight-Ruiz method. TAD boundaries were identified by computing insulation scores [54] with windows of 100 kb (10 bins), normalizing to chromosome-wide averages of insulation scores, then filtering the local minima with the delta vector calculated for the three bins flanking the computed one, and with the difference of the minima and maxima of the delta vector being at least 0.7. Virtual 4C plots were made by plotting the values for one "row" of the normalized Hi-C matrix (corresponding to the interactions of different bins with one specific bin, set to either the *Ikzf1* promoter or the E120 enhancer) against the genomic coordinate of the interacting bin.

4C-seq

Cell preparations were fixed with 2% formaldehyde in their respective culture medium for 10 min at 23˚C. The fixation was quenched with cold glycine at a final concentration of 125 mM, then cells were washed with PBS and permeabilized on ice for 1 h with 10 mM Tris-HCl, pH 8, 100 mM NaCl, 0.1% NP-40 and protease inhibitors. Nuclei were resuspended in DpnII restriction buffer at 10 million nuclei/mL concentration, and 5 million nuclei aliquots were further permeabilized by treatment for either 1 h with 0.4% SDS at 37˚C (ES cells), or for 20 min with 0.7% SDS at 65˚C, then for 40 min at 37˚C (DN3 and P5424 cells). The SDS was then neutralized by incubating for a further 1h with either 2.6% (ES) or 3.3% (DN3 and P5424) Triton-X100 at 37˚C. Nuclei were digested overnight with 1000 U DpnII at 37˚C, then washed twice by centrifuging and resuspending in T4 DNA ligase buffer. *In situ* ligation was performed in 400 μL T4 DNA ligase buffer with 20,000 U T4 DNA ligase overnight at 16˚C. DNA was purified by reverse cross-linking with an overnight incubation at 65˚C with proteinase K, followed by RNase A digestion, phenol/chloroform extraction and isopropanol precipitation. The DNA was digested with 5 U/μg Csp6I at 37˚C overnight, then re-purified by phenol/chloroform extraction and isopropanol precipitation. The DNA was then circularized by ligation with 200 U/μg T4 DNA ligase under dilute conditions (5 ng/μL DNA), and purified by phenol/chloroform extraction and isopropanol precipitation. 50 ng aliquots of this DNA were used as template for PCR with a bait-specific primer located 1.2 kb upstream of E1L (chr11: 11,583,929–11,583,952) and containing Illumina adapter termini (optimal PCR conditions available on request). PCR reactions were pooled, primers removed by washing with 1.8x AMPure XP beads, then quantified on a Bioanalyzer (Agilent) before sequencing with a HiSeq 4000 (Illumina). All bait sequence (including and downstream of the primer sequence, up to but not including the GATC DpnII site) are trimmed by the demultiplexing Sabre tool (https://github.com/najoshi/sabre), allowing two mismatches, before mapping to the mm9 genome with Bowtie [55]. Mapped reads were processed and visualized by 4See [56]. Interactions were called by peakC [57] with a window size of 21.

Results and discussion

Prioritization of an *Ikzf1* enhancer

In primary DP thymocytes, *Ikzf1* is associated with two clusters of enhancers or super-enhancers (Fig 1A). We identified thirteen DNAse I hypersensitive sites (DHSs) within the two super-enhancers in DP thymocytes (Fig 1A; S4 Table). To prioritize functional *Ikzf1* enhancers, we used our previously generated data from a CapStarr-seq high-throughput reporter assay [14] performed in the P5424 cell line [42], which roughly reflects T cell precursors [45]. In this assay, DHSs from primary DP thymocytes were assessed for enhancer activity [14]. Enhancer activity was calculated by the fold change (FC) of CapStarr-seq over the input signals for each DHS. As defined previously [14], active enhancers were classified as weak ($1.5 < FC < 3$) or strong ($FC > 3$). We found that 6 out of 13 *Ikzf1* associated DHSs displayed significant

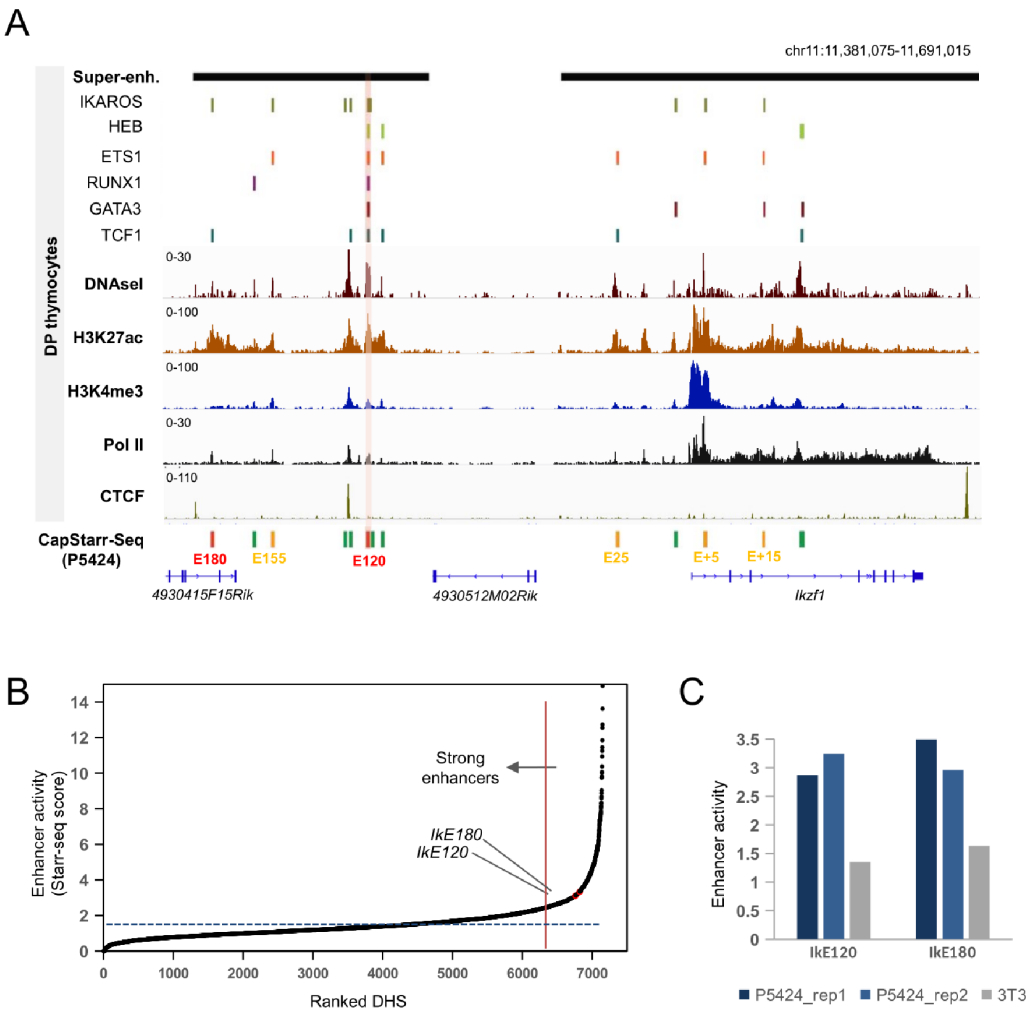


Fig 1. Prioritization of *Ikzf1* enhancers. (A) Epigenomic profiles of the *Ikzf1* locus showing ChIP-Seq signals for H3K4me1, H3K27ac and Pol II, DNAseI-seq, Super-enhancers and peaks of the indicated lymphoid transcription factors in mouse primary DP thymocytes. The enhancer activity of DHS regions as assessed by CapStarr-seq in P5424 cells (green: inactive; orange: weak; red: strong; merged from two replicates) is also shown. A strong enhancer associated with six transcription factors is highlighted. Coordinates of DHS and datasets are listed in S3 and S4 Tables, respectively. (B) Ranked DHSs from primary DP thymocytes in the function of enhancer activity assessed by CapStarr-seq in the P5424 cell line (merge of two replicates). The vertical line indicates the top 5% of the most active enhancer. The IkE120 enhancer is highlighted. (C) Enhancer activity of two *Ikzf1* enhancers assessed by CapStarr-seq in the P5424 and NIH3T3 cell lines.

https://doi.org/10.1371/journal.pone.0233191.g001

enhancer activity (Fig 1A). Of these, the weak enhancer located 15 kb downstream of the *Ikzf1* promoter (IkE+15) overlapped with a previously described enhancer [39, 41]. The two strongest enhancers (average FC higher than 3) were located 180 kb (hereafter IkE180) and 120 kb (hereafter IkE120) upstream of *Ikzf1* and have not been previously identified. These two strong enhancers overlapped the *Ikzf1* upstream super-enhancer and were classified within the top 5% of the most active DHSs present in primary DP thymocytes (ranked 360 and 405 out of 7,152 tested DHSs) (Fig 1B). Neither of the two enhancers was active in the fibroblast-derived NIH-3T3 cell line (Fig 1C). Analyses of H3K27ac ChIP-seq data from human developing

thymocytes [58], suggested that the orthologous regions of IkE180 and IkE120 enhancers are also actives at certain stages of thymic T cell differentiation (S1 Fig).

We next explored whether the putative enhancers were bound by lymphoid specific transcription factors in primary DP thymocytes, using previously published ChIP-seq data for six transcription factors [14, 22, 59–61](S3 Table). Strikingly, the IkE120 enhancer was the only one to be bound by all tested lymphoid transcription factors, including Ikaros itself (Fig 1A). Interestingly, it was also located between two CTCF sites flanking the *Ikzf1* locus (Fig 1A), a known hallmark of "architectural" chromatin loops [62].

To assess whether the identified enhancers directly interact with the *Ikzf1* promoter, we initially analyzed published Hi-C data from primary DP thymocytes [52]. Identified enhancers were all embedded in the same Topological Associated Domain (TAD) as the *Ikzf1* locus, and flanked by convergent CTCF sites (Fig 2A). Virtual circularized chromosome conformation capture (4C) plots suggested that IkE120 and *Ikzf1* promoter preferentially interact together (Fig 2A, bottom panels). To directly demonstrate the interaction between IkE120 and *Ikzf1* promoter we analyzed published 4C-sequencing (4C-seq) experiments in primary DP thymocytes [56] along with newly generated 4C-seq data in primary $CD44^{+}CD25^{+}$ DN thymocytes (DN3) and non-expressing embryonic stem cells (ES), using the *Ikzf1* promoter as the viewpoint (Fig 2B). The *Ikzf1* promoter specifically interacted with the upstream super-enhancer in thymic cells but not ES cells, and displayed a strong interaction with the IkE120 enhancer.

To further interrogate how IkE120 enhancer relates to *Ikzf1* expression, we compared *Ikzf1* expression (RNA-seq) and chromatin accessibility (ATAC) at IkE120 enhancer using a comprehensive resource of hematopoietic cells from the ImmGen consortium [50, 51] (S2A Fig; representative examples are shown in S2B Fig). The IkE120 enhancer displayed the highest ATAC-seq signal in a subset of hematopoietic cells expressing moderated levels of *Ikzf1* expression, including T and B cell precursors and hematopoietic stem cells (HSC). In contrast, hematopoietic cells expressing high levels of *Ikzf1*, such as NK, γδ and mature CD4+ T cells, displayed a weak ATAC-seq signal at IkE120. Stroma cells that did not express *Ikzf1* were not associated with ATAC-seq peak at IkE120. These observations suggest that the IkE120 enhancer might play a preferential role in the expression of the *Ikzf1* gene in lymphoid precursors.

In conclusion, the IkE120 enhancer, displayed one of the strongest enhancer activities within the *Ikzf1* locus, was found to be associated with key lymphoid transcription factors and directly interacted with the *Ikzf1* promoter. While the regulatory elements within the *Ikzf1*-overlapping super-enhancer have been extensively studied [39, 41], the upstream super-enhancer harboring the IkE120 enhancer has remained unexplored. We, therefore, decided to further explore the functional role of this enhancer within its endogenous context.

Deletion of the *Ikzf1* enhancer IkE120

We used CRISPR/Cas9 technology to delete the IkE120 genomic region in the P5424 cell line, encompassing 305 bp covering the DHS site and the six transcription factor binding sites (ΔIkE120) (Fig 3A). Homozygous deletion of IkE120 was assessed by qualitative PCR and Sanger sequencing (Fig 3B and 3C). Note that the P5424 cell line was a valid model to study the endogenous IkE120 enhancer as the enhancer was highly enriched in H3K27ac in these cells and was associated with enhancer RNA (eRNA) expression (Fig 3D). In particular, expression of both *Ikzf1* and associated eRNA can be induced by the treatment of P5424 cells with PMA/ionomycin, which partially mimics T cell differentiation and β-selection [45](Fig 3D and S3 Fig).

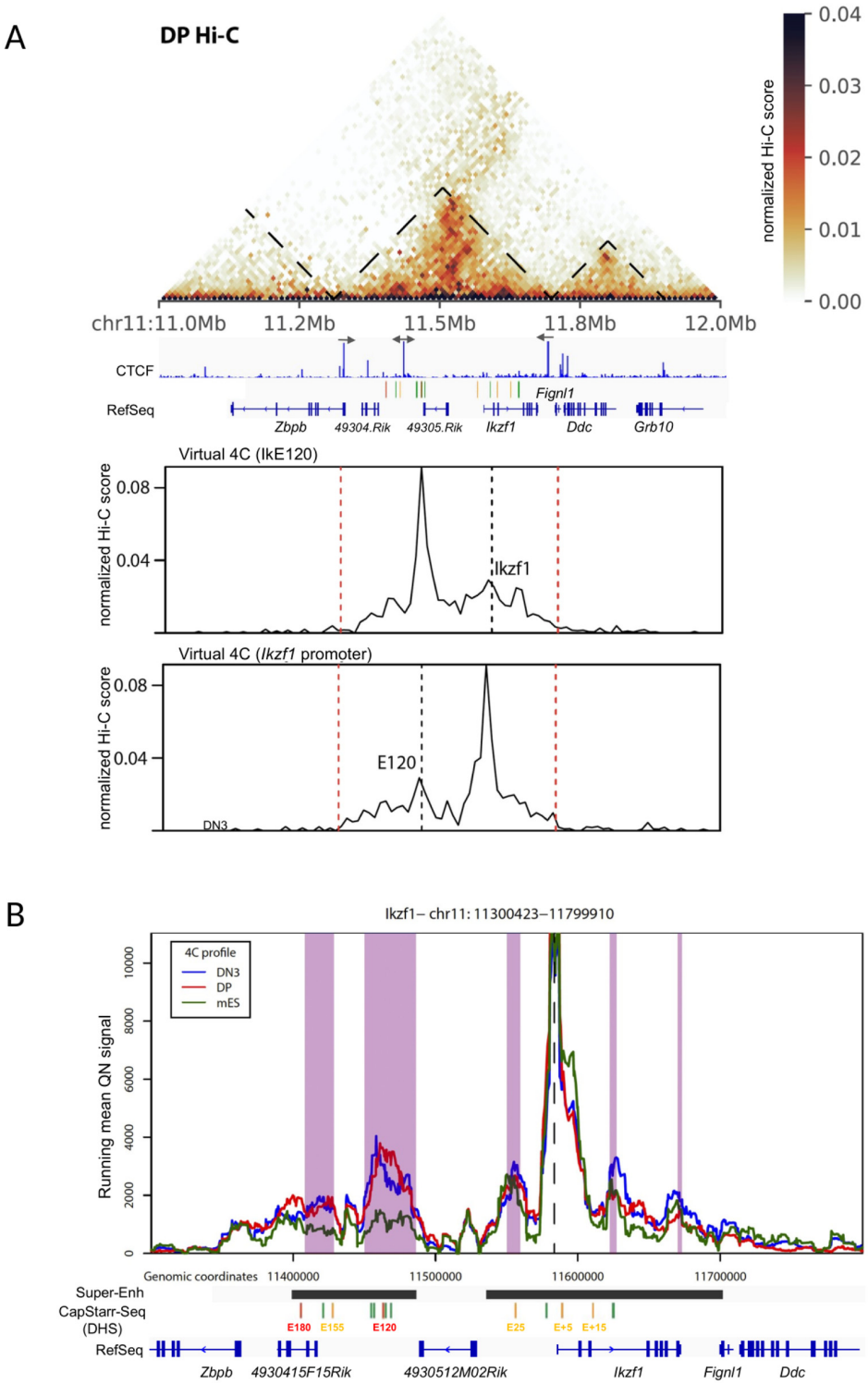


Fig 2. 3D topology of the *Ikzf1* locus. (A) Hi-C view of primary DP thymocytes around the *Ikzf1* locus (top panel). TAD boundaries are shown. The orientation of the main CTCF peaks in primary DP thymocytes is displayed. Virtual 4C plots corresponding to the Hi-C interactions with the *Ikzf1* promoter or the E120 enhancer are shown in the bottom panels. (B) 4C-seq analysis of *Ikzf1* promoter interactions. Running mean (window of 21 fragments), quantile normalized 4C-seq profiles are shown from the *Ikzf1* promoter bait (dotted line) for primary DN3 (blue) and DP (red) thymocytes and mouse ES cells (green). Locations of genes and the six regions with enhancer activity in CapStarr-seq are shown below the plot. Conserved called interactions with thymic cells are highlighted in purple.

https://doi.org/10.1371/journal.pone.0233191.g002

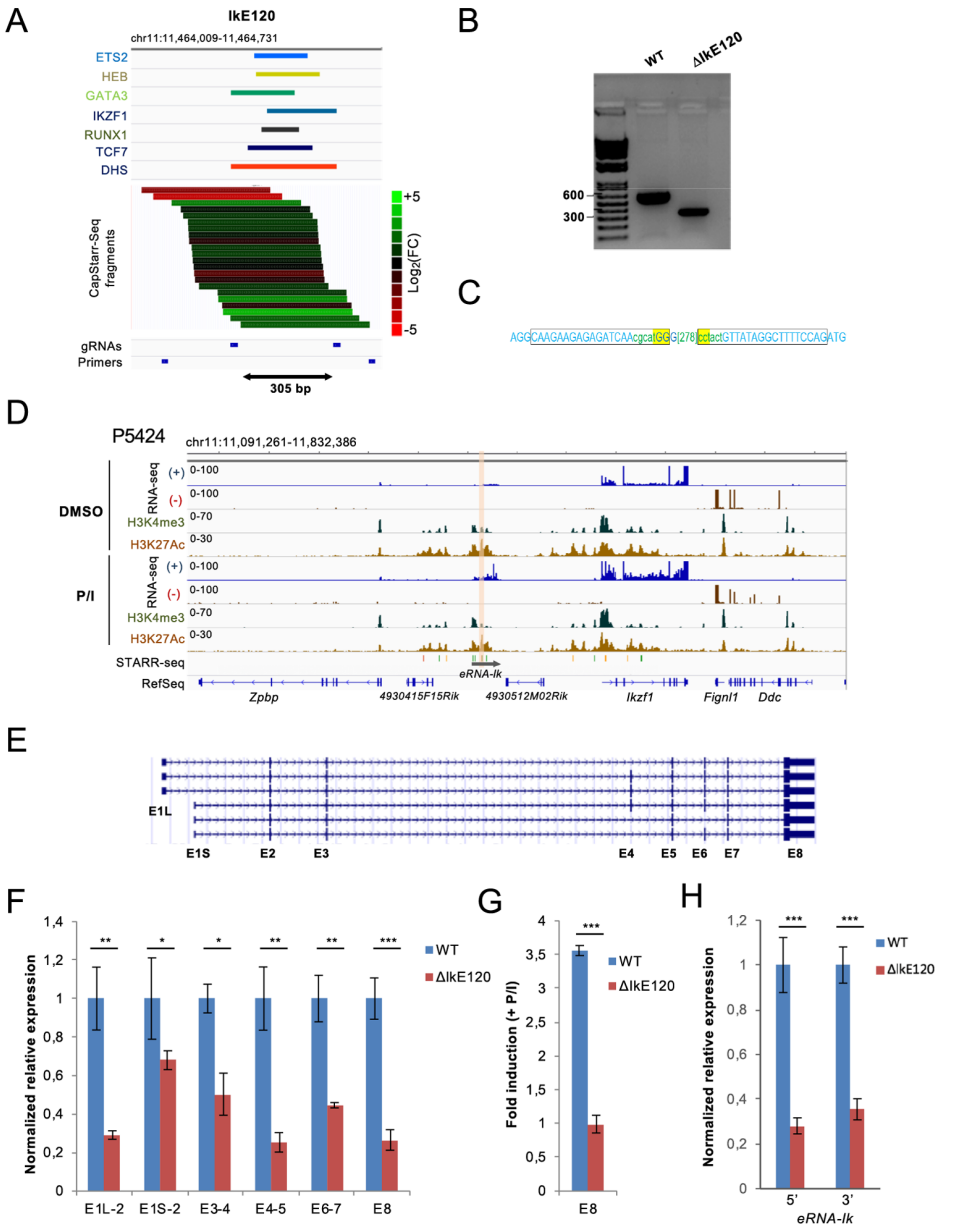


Fig 3. Deletion of the IkE120 enhancer. (A) Genomic tracks showing the binding peaks of the indicated transcription factors overlapping the IkE120 enhancer in primary DP thymocytes as well as the enhancer activity of individual clones assessed by the CapStarr-seq assay in P5424 cells. The color scale indicates the enhancer activity as a Log_2 fold change of the CapStarr-seq signal over the input. The two sgRNAs used to delete the enhancer and primers to detect the deletion are also shown. (B) PCR analyses of IkE120 deletion in the P5424 cell line. (C) Sanger sequencing results from deletion junctions amplified from the genomic DNA of the targeted ΔIkE120 clone. The rectangles represent the position of the sgRNA. The deleted region is indicated in the bracket. (D) Genomic tracks for RNA-seq and ChIP-seq around the *Ikzf1* locus in P5424 cells stimulated or not with PMA/ionomycin (P/I). The IkE120 enhancer is highlighted. The scale of the RNA-seq tracks has been adjusted to visualize the non-coding transcripts overlapping the IkE120 enhancer (a screenshot with unmodified scales for the *Ikzf1* gene is shown in S3 Fig) (E) UCSC genome browser showing the transcripts isoforms of the *Ikzf1* gene found in RefSeq. (F) RT-qPCR analyses of *Ikzf1* expression at the indicated exon-exon junctions in wt and ΔIkE120 P5424 cells. Values represent relative expression as compared with wt samples. (G) Fold induction of *Ikzf1* expression (exon 8) after treatment with PMA and ionomycin of P5424 cells. (H) Relative expression of the non-coding transcripts (eRNA) overlapping the IkE120 enhancer in ΔIkE120 cells as compared with wt P5424 samples. Two sets of primers surrounding the IkE120 deleted region were used. In panels F-H, each point represents the means of three independent experiments normalized by the *Rpl32* housekeeping gene. Statistical significance was assessed by Student's t-test (unpaired, two-tailed) from 3 biological replicates ($^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.1$). Error bars represent standard deviation.

https://doi.org/10.1371/journal.pone.0233191.g003

Based on RefSeq annotation, the *Ikzf1* locus harbors 6 transcript isoforms (Fig 3E), which might play different regulatory functions [34–38]. We assessed the effect of Ik120 deletion on different exon-exon junctions encompassing all annotated *Ikzf1* transcripts by reverse transcription quantitative PCR (RT-qPCR) analyses in wild-type (wt) and ΔIkE120 P5424 cells (Fig 3F). The expression of the common 3' UTR Exon 8 (E8) was decreased four-fold in the ΔIkE120 clone with respect to wt cells. Same results were observed for transcripts encompassing exons E4-E5, while those encompassing E3-E4 and E6-E7 were decreased only two-fold in the ΔIkE120 cells (Fig 3F). We also assessed promoter usages by quantifying the transcripts initiating from either E1L or E1S (Fig 3F). Transcripts originating from both promoters were significantly reduced, although the most upstream promoter appeared to be more affected (Fig 3F).

The deletion of the Ik120 enhancer completely inhibited the upregulation of *Ikzf1* by PMA/ionomycin treatment (Fig 3G). The Ik120 enhancer is associated with an eRNA transcript (hereafter eRNA-Ik), whose expression is correlated with *Ikzf1* induction in P5424 cells and during the DN to DP transition [45](Fig 3D). As expected, the expression of the eRNA-Ik transcript was strongly reduced in ΔIkE120 cells (Fig 3H).

In conclusion, the Ik120 enhancer appears to similarly regulate the different *Ikzf1* isoforms and is particularly required for the induction of the *Ikzf1* gene after cell stimulation.

Deletion of IkE120 affects local epigenomic profiles

To assess whether IkE120 deletion affects the epigenomic profile of the *Ikzf1* locus we performed ChIP-seq experiments to assess H3K27ac profiles. As shown in Fig 4A, the deletion of IkE120 resulted in decreased levels of H3K27ac around the deleted enhancer region and to a lesser extent around the *Ikzf1* promoter, while H3K27ac at the promoter of the neighbor *Zpbp* gene was not affected. Besides, IkE120 deletion did not result in global changes of H3K27ac at gene promoters (S4 Fig). We next performed 4C-seq experiments using the *Ikzf1* promoter as a viewpoint in wt and ΔIkE120 P5424 cells in normal and stimulated conditions (Fig 4B). The genomic interactions observed in wt and mutant P5424 cells were very similar to the interactions observed in primary thymocytes (see Fig 2). Furthermore, no differences were observed between the wt and mutant cells, suggesting that IkE120 is not absolutely required for the establishment of the genomic interaction between the 5' super-enhancer and the *Ikzf1* promoter (Fig 4B). Curiously, the promoter-IkE120 interaction is equivalent in normal and stimulated P5424 cells, whereas the interaction with IkE180 appears to increase slightly on *Ikzf1* upregulation during stimulation. Such findings are consistent with previous studies suggesting that some promoter-enhancer interactions are concomitant with transcriptional induction whereas others are formed prior to gene expression regulation [63]. Despite an overall reduction in *Ikzf1* expression and a complete loss of response to stimulation on IkE120 deletion, these topology dynamics are unchanged by the deletion. Overall, the IkE120 enhancer does not have a widespread influence on H3K27 acetylation and 3D topology of the locus, but rather contribute to localized epigenetic marking. This suggest that others regulatory elements within the 5' super-enhancer are required to ensure the interaction with the *Ikzf1* promoters. Such role might be played by the DHS bound by CTCF upstream of the IkE120 enhancer (Figs 1A and 2A), which also corresponds to the 5' border of influence of IkE120 on H3K27ac (Fig 4A).

Conclusion

Integrative analyses of high-throughput reporter assays, chromatin structure and, 3D topology identified a strong enhancer (IkE120) associated with the *Ikzf1* gene. The deletion of the IkE120 enhancer using CRISPR/Cas9 technology demonstrated a critical role of this enhancer

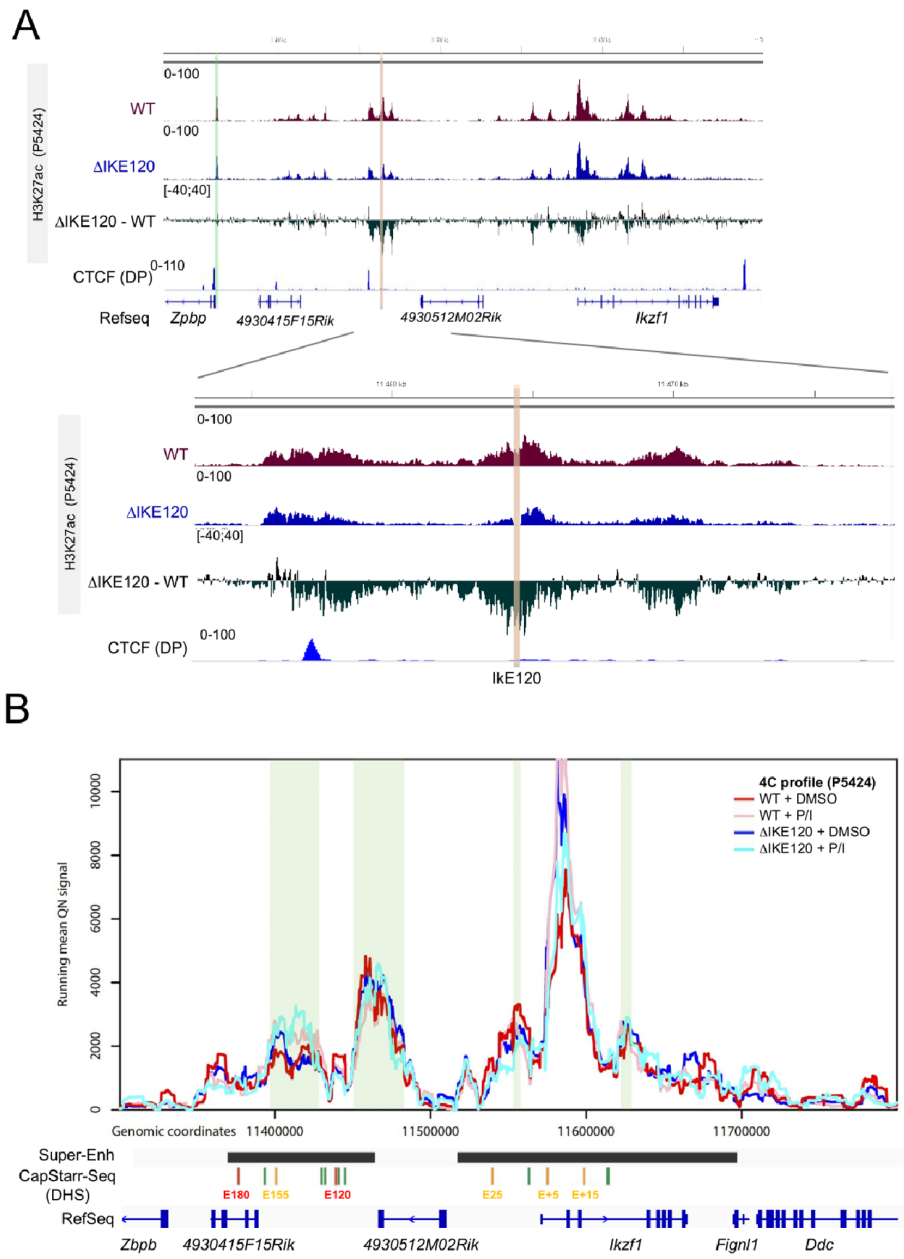


Fig 4. Epigenomic impact of IkE120 deletion. (A) The H3K27ac ChIP-seq at the *Ikzf1* locus (top) and around the IkE120 enhancer (bottom) in wt and ΔIkE120 P5424 cells are shown as individual tracks and as the differential signal between wt and ΔIkE120 cells. The genomic track of CTCF ChIP-seq in primary DP thymocytes is also shown. H3K27ac The IkE120 deleted region is highlighted in red and the promoter region of the neighbor *Zpbp* gene is highlighted in green. (B) Running mean (window of 21 fragments), quantile normalized 4C-seq profiles are shown from the *Ikzf1* promoter bait for wt unstimulated (red), wt stimulated (pink), ΔIkE120 unstimulated (blue) and ΔIkE120 stimulated P5424 cells. Locations of genes and the six regions with enhancer activity in CapStarr-seq are shown below the plot. Conserved called interactions are highlighted in green.

https://doi.org/10.1371/journal.pone.0233191.g004

in controlling the expression of the *Ikzf1* gene. However, the IkE120 enhancer has a modest impact on the chromatin structure and 3D topology of the locus, highlighting the complexity of the regulatory landscape regulating the *Ikzf1* locus.

Supporting information

S1 Fig. A) Conservation of Ikzf1 enhancers across mammalian species. Detailed view of the IkE120 enhancer conservation is indicated at the bottom panel. B) H3K27ac tracks at the indicated human T cell precursors and Hematopoietic Stem Cells (HSC). The position of the human orthologous regions of the *Ikzf1* enhancers are indicated.
(PDF)

S2 Fig. A) Comparison between *Ikzf1* expression and IkE120 chromatin opening in the hematopoietic lineages as indicated. Normalized RNA-seq and ATAC-seq data was retrieved from the ImmGen portal (http://www.immgen.org). B) ATAC-seq signal around the IkE120 enhancer (Left panel; signal scale was set to 5) and Ikezf1 expression (right panel) at selected hematopoietic samples.
(PDF)

S3 Fig. Genomic tracks at the *Ikzf1* gene for the RNA-seq in P5424 cells stimulated or not with PMA/ionomycin(P/I).
(PDF)

S4 Fig. Average profiles and heatmaps of H3K27ac centered on the TSS of coding genes in wt and ΔIkE120 P5424 cells.
(PDF)

S1 Table. Primer sequences for CRISPR.
(PDF)

S2 Table. Primer sequences for RT-qPCR.
(PDF)

S3 Table. Information about published datasets used in this study and downloaded from the NCBI Gene Expression Omnibus.
(PDF)

S4 Table. List of DHSs associated with Ikzf1. The enhancer activity as assessed by CapStarr-seq in the P5424 cell line is indicated.
(PDF)

S1 Raw images. Original gel image corresponding to Fig 3B. Lanes not included in the final figure were marked with an "X". TrackIt 1 Kb Plus DNA Ladder (Thermo Fisher) was used as DNA ladder.
(PDF)

Acknowledgments

We thank the Transcriptomics and Genomics Marseille-Luminy (TGML) platform for sequencing the ChIP-seq samples and the Marseille-Luminy cell biology platform for the management of cell culture. Sequencing of 4C samples was performed by the IGBMC GenomEast platform. TGML and GenomEast platforms are member of the France Genomique consortium (ANR-10-INBS-0009).

Author Contributions

Methodology: Anne M. Molitor, Nori Sadouni, Saadat Hussain, Magali Torres, Wiam Saadi, Lan T. M. Dao, Guillaume Charbonnier, David Santiago-Algarra, Denis Puthier, Tom Sexton.

Supervision: Jean Christophe Andrau, Salvatore Spicuglia.

Validation: Anne M. Molitor.

Writing – original draft: Jaafar Alomairi, Salvatore Spicuglia.

Writing – review & editing: Salvatore Spicuglia.

References

1. Plank JL, Dean A. Enhancer Function: Mechanistic and Genome-Wide Insights Come Together. Molecular cell. 2014; 55(1):5–14. Epub 2014/07/06. https://doi.org/10.1016/j.molcel.2014.06.015 PMID: 24996062.
2. Pott S, Lieb JD. What are super-enhancers? Nature genetics. 2015; 47(1):8–12. Epub 2014/12/31. https://doi.org/10.1038/ng.3167 PMID: 25547603.
3. Hnisz D, Shrinivas K, Young RA, Chakraborty AK, Sharp PA. A Phase Separation Model for Transcriptional Control. Cell. 2017; 169(1):13–23. Epub 2017/03/25. https://doi.org/10.1016/j.cell.2017.02.007 PMID: 28340338
4. Suzuki HI, Young RA, Sharp PA. Super-Enhancer-Mediated RNA Processing Revealed by Integrative MicroRNA Network Analysis. Cell. 2017; 168(6):1000–14 e15. Epub 2017/03/12. https://doi.org/10.1016/j.cell.2017.02.015 PMID: 28283057
5. Bevington SL, Cauchy P, Cockerill PN. Chromatin priming elements establish immunological memory in T cells without activating transcription: T cell memory is maintained by DNA elements which stably prime inducible genes without activating steady state transcription. BioEssays: news and reviews in molecular, cellular and developmental biology. 2017; 39(2). Epub 2016/12/28. https://doi.org/10.1002/bies.201600184 PMID: 28026028.
6. Chatterjee S, Ahituv N. Gene Regulatory Elements, Major Drivers of Human Disease. Annual review of genomics and human genetics. 2017. Epub 2017/04/13. https://doi.org/10.1146/annurev-genom-091416-035537 PMID: 28399667.
7. Natoli G, Andrau JC. Noncoding transcription at enhancers: general principles and functional models. Annu Rev Genet. 2012; 46:1–19. Epub 2012/08/22. https://doi.org/10.1146/annurev-genet-110711-155459 PMID: 22905871.
8. Santiago-Algarra D, Dao LTM, Pradel L, Espana A, Spicuglia S. Recent advances in high-throughput approaches to dissect enhancer function. F1000Res. 2017; 6:939. https://doi.org/10.12688/f1000research.11581.1 PMID: 28690838
9. Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. High-throughput functional testing of ENCODE segmentation predictions. Genome research. 2014; 24(10):1595–602. https://doi.org/10.1101/gr.173518.114 PMID: 25035418
10. Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome research. 2013; 23(5):800–11. Epub 2013/03/21. https://doi.org/10.1101/gr.144899.112 PMID: 23512712
11. Ernst J, Melnikov A, Zhang X, Wang L, Rogov P, Mikkelsen TS, et al. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nat Biotechnol. 2016; 34(11):1180–90. https://doi.org/10.1038/nbt.3678 PMID: 27701403
12. Wang X, He L, Goggin SM, Saadat A, Wang L, Sinnott-Armstrong N, et al. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. Nat Commun. 2018; 9(1):5380. Epub 2018/12/21. https://doi.org/10.1038/s41467-018-07746-1 PMID: 30568279
13. Li QL, Wang DY, Ju LG, Yao J, Gao C, Lei PJ, et al. The hyper-activation of transcriptional enhancers in breast cancer. Clin Epigenetics. 2019; 11(1):48. Epub 2019/03/15. https://doi.org/10.1186/s13148-019-0645-x PMID: 30867030
14. Vanhille L, Griffon A, Maqbool MA, Zacarias-Cabeza J, Dao LTM, Fernandez N, et al. High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. Nat Commun. 2015; 6:6905. https://doi.org/10.1038/ncomms7905 PMID: 25872643

15. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. Nature genetics. 2019; 51(12):1664–9. Epub 2019/12/01. https://doi.org/10.1038/s41588-019-0538-0 PMID: 31784727

16. Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, et al. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell. 2019; 176(1–2):377–90 e19. Epub 2019/01/08. https://doi.org/10.1016/j.cell.2018.11.029 PMID: 30612741

17. Georgopoulos K. The making of a lymphocyte: the choice among disparate cell fates and the IKAROS enigma. Genes Dev. 2017; 31(5):439–50. Epub 2017/04/08. https://doi.org/10.1101/gad.297002.117 PMID: 28385788

18. Heizmann B, Kastner P, Chan S. The Ikaros family in lymphocyte development. Curr Opin Immunol. 2018; 51:14–23. Epub 2017/12/27. https://doi.org/10.1016/j.coi.2017.11.005 PMID: 29278858.

19. Kim J, Sif S, Jones B, Jackson A, Koipally J, Heller E, et al. Ikaros DNA-binding proteins direct formation of chromatin remodeling complexes in lymphocytes. Immunity. 1999; 10:345–55. KIM99B. https://doi.org/10.1016/s1074-7613(00)80034-5 PMID: 10204490

20. Kleinmann E, Geimer Le Lay AS, Sellars M, Kastner P, Chan S. Ikaros represses the transcriptional response to Notch signaling in T-cell development. Molecular and cellular biology. 2008; 28(24):7465–75. KLEINMANN2008. https://doi.org/10.1128/MCB.00715-08 PMID: 18852286

21. Sridharan R, Smale ST. Predominant Interaction of Both Ikaros and Helios with the NuRD Complex in Immature Thymocytes. Journal of Biological Chemistry. 2007; 282(41):30227–38. SRI07. https://doi.org/10.1074/jbc.M702541200 PMID: 17681952

22. Oravecz A, Apostolov A, Polak K, Jost B, Le Gras S, Chan S, et al. Ikaros mediates gene silencing in T cells through Polycomb repressive complex 2. Nat Commun. 2015; 6:8823. Epub 2015/11/10. https://doi.org/10.1038/ncomms9823 PMID: 26549758

23. Gomez-del Arco P, Kashiwagi M, Jackson AF, Naito T, Zhang J, Liu F, et al. Alternative promoter usage at the Notch1 locus supports ligand-independent signaling in T cell development and leukemogenesis. Immunity. 2010; 33(5):685–98. Epub 2010/11/26. https://doi.org/10.1016/j.immuni.2010.11.008 PMID: 21093322

24. Morel G, Deau MC, Simand C, Caye-Eude A, Arfeuille C, Ittel A, et al. Large deletions of the 5' region of IKZF1 lead to haploinsufficiency in B-cell precursor acute lymphoblastic leukaemia. Br J Haematol. 2019; 186(5):e155–e9. Epub 2019/05/31. https://doi.org/10.1111/bjh.15994 PMID: 31148164.

25. Winandy S, Wu P, Georgopoulos K. A dominant mutation in the Ikaros gene leads to rapid development of leukemia and lymphoma. Cell. 1995; 83:289–99. WIN95. https://doi.org/10.1016/0092-8674(95)90170-1 PMID: 7585946

26. Schjerven H, McLaughlin J, Arenzana TL, Frietze S, Cheng D, Wadsworth SE, et al. Selective regulation of lymphopoiesis and leukemogenesis by individual zinc fingers of Ikaros. Nature immunology. 2013; 14(10):1073–83. Epub 2013/09/10. https://doi.org/10.1038/ni.2707 PMID: 24013668

27. Olsson L, Johansson B. Ikaros and leukaemia. Br J Haematol. 2015; 169(4):479–91. https://doi.org/10.1111/bjh.13342 PMID: 25753742.

28. Kastner P, Chan S. Role of Ikaros in T-cell acute lymphoblastic leukemia. World J Biol Chem. 2011; 2(6):108–14. https://doi.org/10.4331/wjbc.v2.i6.108 PMID: 21765975

29. Zhang J, Jackson AF, Naito T, Dose M, Seavitt J, Liu F, et al. Harnessing of the nucleosome-remodeling-deacetylase complex controls lymphocyte development and prevents leukemogenesis. Nature immunology. 2012; 13(1):86–94. Epub 2011/11/15. https://doi.org/10.1038/ni.2150 PMID: 22080921.

30. Joshi I, Yoshida T, Jena N, Qi X, Zhang J, Van Etten RA, et al. Loss of Ikaros DNA-binding function confers integrin-dependent survival on pre-B cells and progression to acute lymphoblastic leukemia. Nature immunology. 2014; 15(3):294–304. Epub 2014/02/11. https://doi.org/10.1038/ni.2821 PMID: 24509510

31. Churchman ML, Low J, Qu C, Paietta EM, Kasper LH, Chang Y, et al. Efficacy of Retinoids in IKZF1-Mutated BCR-ABL1 Acute Lymphoblastic Leukemia. Cancer cell. 2015; 28(3):343–56. Epub 2015/09/01. https://doi.org/10.1016/j.ccell.2015.07.016 PMID: 26321221

32. Mullighan CG, Miller CB, Radtke I, Phillips LA, Dalton J, Ma J, et al. BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. Nature. 2008; 453(7191):110–4. Epub 2008/04/15. https://doi.org/10.1038/nature06866 PMID: 18408710.

33. Churchman ML, Qian M, Te Kronnie G, Zhang R, Yang W, Zhang H, et al. Germline Genetic IKZF1 Variation and Predisposition to Childhood Acute Lymphoblastic Leukemia. Cancer cell. 2018; 33(5):937–48 e8. Epub 2018/04/24. https://doi.org/10.1016/j.ccell.2018.03.021 PMID: 29681510

34. Bellavia D, Mecarozzi M, Campese AF, Grazioli P, Talora C, Frati L, et al. Notch3 and the Notch3-upregulated RNA-binding protein HuD regulate Ikaros alternative splicing. The EMBO journal. 2007; 26(6):1670–80. BELLAVIA2007A. https://doi.org/10.1038/sj.emboj.7601626 PMID: 17332745

35. Molnar A, Georgopoulos K. The Ikaros gene encodes a family of functionally diverse zinc finger DNA-binding proteins. Molecular and cellular biology. 1994; 14:8292–303. MOL94. https://doi.org/10.1128/mcb.14.12.8292 PMID: 7969165

36. Molnar A, Wu P, Largespada DA, Vortkamp A, Scherer S, Copeland NG, et al. The Ikaros gene encodes a family of lymphocyte-restricted zinc finger DNA binding proteins, highly conserved in human and mouse. J Immunol. 1996; 156:585–92. MOL96. PMID: 8543809

37. Klug CA, Morrison SJ, Masek M, Hahm K, Smale ST, Weissman IL. Hematopoietic stem cells and lymphoid progenitors express different Ikaros isoforms, and Ikaros is localized to heterochromatin in immature lymphocytes. Proc Nat Acad Sci USA. 1998; 95:657–62. KLU98. https://doi.org/10.1073/pnas.95.2.657 PMID: 9435248

38. Sun L, Liu A, Georgopoulos K. Zing finger-mediated protein interactions modulate Ikaros activity, a molecular control of lymphocyte development. The EMBO journal. 1996; 15:5358–69. SUN96. PMID: 8895580

39. Kaufmann C, Yoshida T, Perotti EA, Landhuis E, Wu P, Georgopoulos K. A complex network of regulatory elements in Ikaros and their activity during hemo-lymphopoiesis. The EMBO journal. 2003; 22 (9):2211–23. Epub 2003/05/03. https://doi.org/10.1093/emboj/cdg186 PMID: 12727887

40. Perotti EA, Georgopoulos K, Yoshida T. An Ikaros Promoter Element with Dual Epigenetic and Transcriptional Activities. PloS one. 2015; 10(7):e0131568. Epub 2015/07/03. https://doi.org/10.1371/journal.pone.0131568 PMID: 26135129

41. Yoshida T, Landhuis E, Dose M, Hazan I, Zhang J, Naito T, et al. Transcriptional regulation of the Ikzf1 locus. Blood. 2013; 122(18):3149–59. Epub 2013/09/05. https://doi.org/10.1182/blood-2013-01-474916 PMID: 24002445

42. Mombaerts P, Terhorst C, Jacks T, Tonegawa S, Sancho J. Characterization of immature thymocyte lines derived from T-cell receptor or recombination activating gene 1 and p53 double mutant mice. Proc Natl Acad Sci U S A. 1995; 92(16):7420–4. Epub 1995/08/01. https://doi.org/10.1073/pnas.92.16.7420 PMID: 7638208

43. Naito Y, Hino K, Bono H, Ui-Tei K. CRISPRdirect: software for designing CRISPR/Cas guide RNA with reduced off-target sites. Bioinformatics. 2015; 31(7):1120–3. Epub 2014/11/22. https://doi.org/10.1093/bioinformatics/btu743 PMID: 25414360

44. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, et al. RNA-guided human genome engineering via Cas9. Science. 2013; 339(6121):823–6. Epub 2013/01/05. https://doi.org/10.1126/science.1232033 PMID: 23287722

45. Saadi W, Kermezli Y, Dao LTM, Mathieu E, Santiago-Algarra D, Manosalva I, et al. A critical regulator of Bcl2 revealed by systematic transcript discovery of lncRNAs associated with T-cell differentiation. Scientific Reports. 2019; 9(1):4707. https://doi.org/10.1038/s41598-019-41247-5 PMID: 30886319

46. Ramirez F, Dundar F, Diehl S, Gruning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic acids research. 2014; 42(Web Server issue):W187–91. Epub 2014/05/07. https://doi.org/10.1093/nar/gku365 PMID: 24799436

47. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013; 14(2):178–92. Epub 2012/04/21. https://doi.org/10.1093/bib/bbs017 PMID: 22517427

48. Lee CM, Barber GP, Casper J, Clawson H, Diekhans M, Gonzalez JN, et al. UCSC Genome Browser enters 20th year. Nucleic acids research. 2020; 48(D1):D756–D61. Epub 2019/11/07. https://doi.org/10.1093/nar/gkz1012 PMID: 31691824.

49. Stunnenberg HG, Hirst M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell. 2016; 167(5):1145–9. Epub 2016/11/20. https://doi.org/10.1016/j.cell.2016.11.007 PMID: 27863232.

50. Heng TS, Painter MW. The Immunological Genome Project: networks of gene expression in immune cells. Nature immunology. 2008; 9(10):1091–4. Epub 2008/09/19. https://doi.org/10.1038/ni1008-1091 PMID: 18800157.

51. Yoshida H, Lareau CA, Ramirez RN, Rose SA, Maier B, Wroblewska A, et al. The cis-Regulatory Atlas of the Mouse Immune System. Cell. 2019; 176(4):897–912 e20. Epub 2019/01/29. https://doi.org/10.1016/j.cell.2018.12.036 PMID: 30686579

52. Hu G, Cui K, Fang D, Hirose S, Wang X, Wangsa D, et al. Transformation of Accessible Chromatin and 3D Nucleome Underlies Lineage Commitment of Early T Cells. Immunity. 2018; 48(2):227–42 e8. Epub 2018/02/22. https://doi.org/10.1016/j.immuni.2018.01.013 PMID: 29466755

53. Kruse K, Hug CB, Vaquerizas JM. FAN-C: A Feature-rich Framework for the Analysis and Visualisation of C data. bioRxiv. 2020:2020.02.03.932517. https://doi.org/10.1101/2020.02.03.932517

54. Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature. 2015; 523(7559):240–4. Epub 2015/06/02. https://doi.org/10.1038/nature14450 PMID: 26030525

55. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10(3):R25. Epub 2009/03/06. https://doi.org/10.1186/gb-2009-10-3-r25 PMID: 19261174

56. Ben Zouari Y, Platania A, Molitor AM, Sexton T. 4See: A Flexible Browser to Explore 4C Data. Front Genet. 2019; 10:1372. Epub 2020/02/11. https://doi.org/10.3389/fgene.2019.01372 PMID: 32038719

57. Geeven G, Teunissen H, de Laat W, de Wit E. peakC: a flexible, non-parametric peak calling package for 4C and Capture-C data. Nucleic acids research. 2018; 46(15):e91. Epub 2018/05/26. https://doi.org/10.1093/nar/gky443 PMID: 29800273

58. Adams D, Altucci L, Antonarakis SE, Ballesteros J, Beck S, Bird A, et al. BLUEPRINT to decode the epigenetic signature written in blood. Nat Biotechnol. 2012; 30(3):224–6. Epub 2012/03/09. https://doi.org/10.1038/nbt.2153 PMID: 22398613.

59. Koch F, Fenouil R, Gut M, Cauchy P, Albert TK, Zacarias-Cabeza J, et al. Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. Nature structural & molecular biology. 2011; 18(8):956–63. Epub 2011/07/19. https://doi.org/10.1038/nsmb.2085 PMID: 21765417.

60. Lepoivre C, Belhocine M, Bergon A, Griffon A, Yammine M, Vanhille L, et al. Divergent transcription is associated with promoters of transcriptional regulators. BMC genomics. 2013; 14:914. Epub 2013/12/25. https://doi.org/10.1186/1471-2164-14-914 PMID: 24365181

61. Wei G, Abraham BJ, Yagi R, Jothi R, Cui K, Sharma S, et al. Genome-wide analyses of transcription factor GATA3-mediated gene regulation in distinct T cell types. Immunity. 2011; 35(2):299–311. Epub 2011/08/27. https://doi.org/10.1016/j.immuni.2011.08.007 PMID: 21867929

62. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014; 159(7):1665–80. Epub 2014/12/17. https://doi.org/10.1016/j.cell.2014.11.021 PMID: 25497547

63. Schoenfelder S, Fraser P. Long-range enhancer-promoter contacts in gene expression control. Nature reviews Genetics. 2019; 20(8):437–55. Epub 2019/05/16. https://doi.org/10.1038/s41576-019-0128-0 PMID: 31086298.

4.2 Les Epromoteurs : des plateformes pour le recrutement des facteurs de transcription nécessaire à la réponse inflammatoire

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-26861-0 OPEN

Epromoters function as a hub to recruit key transcription factors required for the inflammatory response

David Santiago-Algarra[1,2,6], Charbel Souaid[1,2,6], Himanshu Singh[1,2,6], Lan T. M. Dao[1,2,3], Saadat Hussain[1,2], Alejandra Medina-Rivera[4], Lucia Ramirez-Navarro[4], Jaime A. Castro-Mondragon[1,5], Nori Sadouni[1,2], Guillaume Charbonnier[1,2] & Salvatore Spicuglia[1,2]✉

Gene expression is controlled by the involvement of gene-proximal (promoters) and distal (enhancers) regulatory elements. Our previous results demonstrated that a subset of gene promoters, termed Epromoters, work as bona fide enhancers and regulate distal gene expression. Here, we hypothesized that Epromoters play a key role in the coordination of rapid gene induction during the inflammatory response. Using a high-throughput reporter assay we explored the function of Epromoters in response to type I interferon. We find that clusters of IFNa-induced genes are frequently associated with Epromoters and that these regulatory elements preferentially recruit the STAT1/2 and IRF transcription factors and distally regulate the activation of interferon-response genes. Consistently, we identified and validated the involvement of Epromoter-containing clusters in the regulation of LPS-stimulated macrophages. Our findings suggest that Epromoters function as a local hub recruiting the key TFs required for coordinated regulation of gene clusters during the inflammatory response.

[1] Aix-Marseille University, INSERM, TAGC, UMR 1090 Marseille, France. [2] Equipe Labellisée Ligue Contre le Cancer, Paris, France. [3] Vinmec Research Institute of Stem cell and Gene technology, Vinmec Healthcare System, Hanoi, Vietnam. [4] Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Juriquilla, Mexico. [5] Present address: Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway. [6] These authors contributed equally: David Santiago-Algarra, Charbel Souaid, Himanshu Singh. ✉email: salvatore.spicuglia@inserm.fr

Regulation of gene transcription in higher eukaryotes is accomplished through the involvement of transcription start site (TSS)-proximal (promoters) and -distal (enhancers) regulatory elements. It is now well acknowledged that enhancer elements play an essential role during development and cell differentiation, while genetic alterations in these elements are a major cause of human diseases[1]. The classical definition of enhancers implies the property to activate gene expression at a distance, while promoters induce local gene expression. However, this basic dichotomy has been challenged by broad similarities between promoters and enhancers[2,3].

While epigenomic studies allow genome-wide identification of putative enhancers, they do not provide direct proof of enhancer function or activity. To tackle this problem, several high-throughput reporter assays have been developed and implemented to study enhancer activity in different cellular contexts[4]. Using high-throughput reporter assays in different cellular contexts from drosophila to humans, it was found that a subset of gene-promoters, also termed Epromoters, displays enhancer activity when tested in vitro[5–9]. Importantly, several concomitant studies provided evidence that some core promoters indeed control distal gene expression in their natural context[5,10–12], and that human genetic variation within Epromoters influences distal gene expression, as assessed by expression Quantitative Trait Loci (eQTL)[5,13]. Overall, the finding that a subset of promoters works also as bona fide enhancers have significant implications for the understanding of complex gene regulation in normal development and open the intriguing possibility that sequence variation lying within a subset of (E)promoters might impact on physiological traits or diseases by directly regulating distal gene expression[3].

Previous studies have suggested a link between the Epromoter function and the stress responses, particularly the regulation of interferon-response genes[5,14,15]. The inflammatory response requires the activation of a complex transcriptional program that is both cell-type- and stimulus-specific and involves the dynamic regulation of hundreds of genes[16]. In the context of inflamed tissue, extensive changes in gene expression occur in both parenchymal cells and infiltrating cells of the immune system. Transcriptional regulation of interferon (IFN)-response genes is one of the most intensively studied regulatory processes. In general, type I (IFNa/IFNb) and II (IFNg) interferons are responsible for regulating and activating the immune response. Expression of type I interferons can be induced in virtually all cell types upon recognition of viral components, especially nucleic acids, whereas type II interferon is induced by cytokines such as IL-12, and its expression is restricted to immune cells. While most transcription factors (TFs) involved in type I and II response are well characterized (e.g., STAT1/2 and IRFs for type I), there are still open questions concerning the precise epigenetic mechanisms involved in the rapid and precise activation of IFNa-stimulated genes. For instance, not all promoters of IFNa-stimulated genes are bound by the aforementioned factors and underlying mechanisms still await further clarification[17].

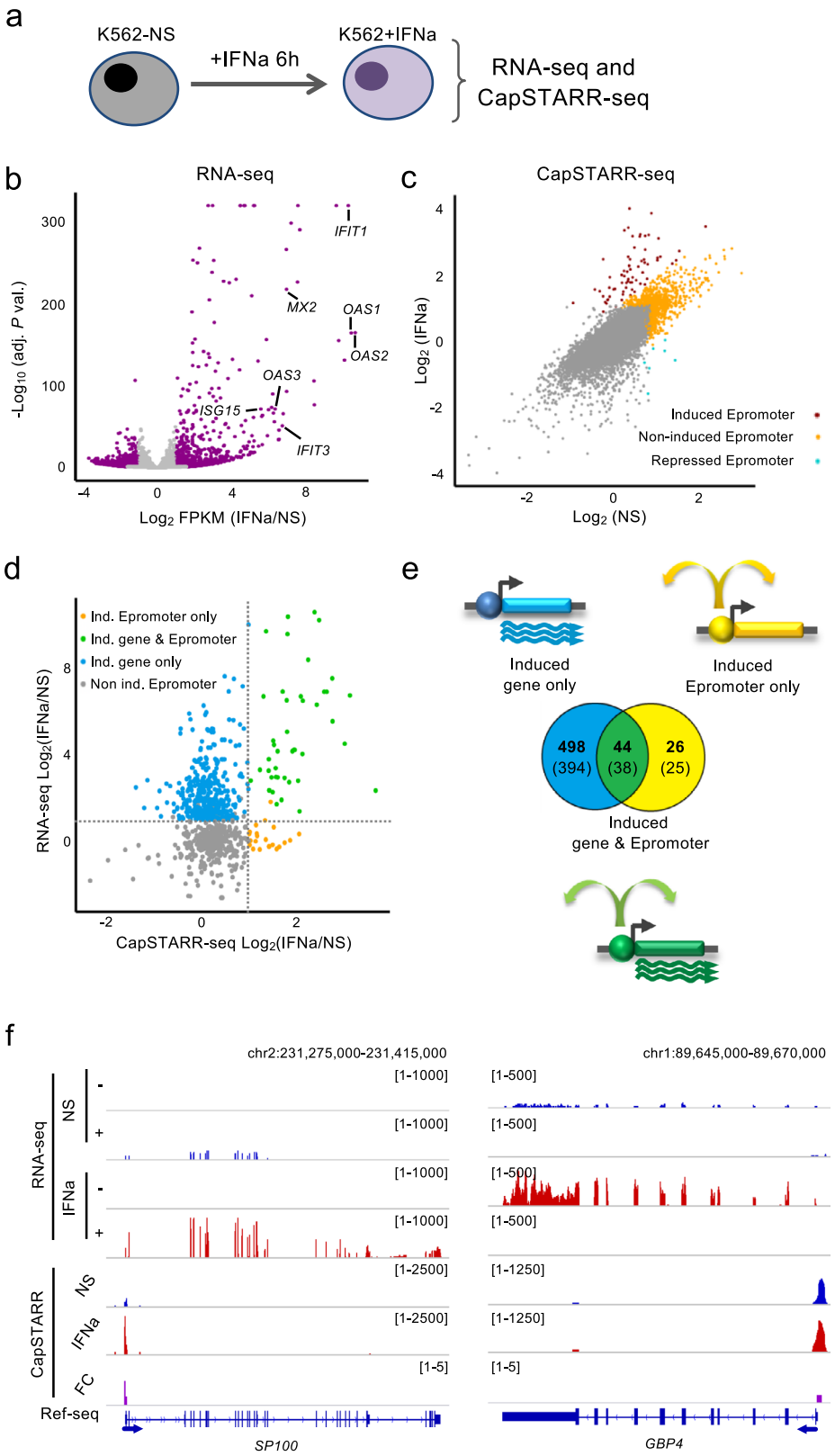
We hypothesize that Epromoters might play an important role in the coordination of rapid gene induction after IFNa stimulation, and more generally during the inflammatory response. By combining high-throughput reporter assays and gene expression analyses, we find that a significant subset of IFNa-induced genes is associated with Epromoters and regulates the induction of other neighbor genes. Consistently, we found that IFNa-induced Epromoters have the ability to specifically and efficiently recruit the interferon-response transcription factors. In fact, within a typical cluster of IFNa-response genes, the interferon response factors are found to exclusively bind to the Epromoter, which is required for the proper induction of other genes within the cluster. Predictions based on gene expression dynamics and TF-binding profiles allow us to identify and validate Epromoter-regulated clusters in primary LPS-stimulated macrophages. Thus, Epromoters have a broad role in the induction of co-regulated genes during the mammalian inflammatory response.

Results

Epromoters are involved in type I interferon response. To initially explore the link between Epromoters and type I interferon response, we analyzed active enhancers in HeLa cells in the presence or absence of Interferon type I response inhibitors, using published whole-genome enhancer screen from Self-Transcribing Active Regulatory Region Sequencing (STARR–seq) experiments[15]. Noted that, HeLa cells have active interferon signaling in the absence of exogenous stimulation[5,15]. The genomic distribution of active enhancers shows that the interferon-dependent enhancers (active without inhibitors) are significantly closer to the TSS (P-val. $= 2.2 \times 10^{-16}$; KS test) as compared to interferon-independent enhancers (actives in the presence of inhibitors) (Supplementary Fig. 1a, b). Indeed, we found that the number of proximal enhancers (<1 kb from the closest TSS; Epromoters) decreased after treatment with interferon inhibitors, while distal enhancers (>1 kb from any TSS) increased (Supplementary Fig. 1c). Interferon-dependent TSS-proximal enhancers (Epromoters) were enriched in Interferon-Stimulated Response Elements (ISRE), including binding sites for TFs of the IRF family (Supplementary Fig. 1d). In contrast, TSS-proximal enhancers in the presence of the interferon inhibitors, as well as, the TSS-distal enhancers were enriched in binding sites for developmental TFs (Supplementary Fig. 1e). This suggests that TSS-proximal enhancers (i.e., Epromoters) are preferentially activated by the type I interferon signaling in HeLa cells.

To directly assess the contribution of Epromoters to the regulation of type I interferon response, we performed paralleled experiments to assess gene expression (RNA-seq) and enhancer activity of gene promoters (CapSTARR-seq) (Fig. 1a and Supplementary Data 1) in the K562 cell line with and without IFNa stimulation. CapSTARR-seq, is a high-throughput reporter assay, coupling capture of regions of interest to STARR-seq[18,19] and was previously used to identify Epromoters in unstimulated K562 cells[5]. Notably, the K562 cells do not express type I interferon response genes in the absence of stimulation and do not induce an interferon response after DNA transfection[5,15], making this cell line an appropriate model to study type I interferon stimulation. To quantify the enhancer activity of gene promoters, we used a capture-based library containing 17,941 promoters (−200 bp to +50 bp from the TSS), corresponding to 14,188 RefSeq-defined coding genes as previously described[5]. Analyses of RNA-seq after IFNa stimulation of K562 cells resulted in 426 induced and 436 repressed genes (P-val. < 0.001; Fig. 1b). However, induced genes generally reached higher significance as compared to repressed genes, consistent with previous findings[20]. Highly IFNa-Induced genes included classical type I interferon response genes such as the *MX*, *OAS*, and *IFIT* family of genes (Supplementary Data 1).

The enhancer activity was calculated as the fold-change of the CapSTARR-seq signal over the input library and was highly reproducible between replicates (Pearson's $R^2 > 0.9$; Supplementary Fig. 2a). Gene promoters were defined as Epromoter when the enhancer activity was above the inflection point of the ranked signal of all captured regions, as previously described[5]. Induced Epromoters were defined as Epromoters that gained enhancer activity by a fold-change of 2 between IFNa-stimulated and non-stimulated K562 cells. We identified 429 constitutive Epromoters, as well as 70 induced and 7 repressed Epromoters (Fig. 1c,



Supplementary Data 1). Combining the RNA-seq and CapSTARR-seq experiments allowed us to define three sets of IFNa-response loci in K562 cells (Fig. 1d, e and Supplementary Data 2): (i) 498 promoters associated with 394 IFNa-induced genes but without induced Epromoter activity (induced gene only), (ii) 44 promoters with induced Epromoter activity and associated with 38 IFNa-induced genes (induced gene & Epromoter), and (iii) 26 promoters associated with IFNa-induced Epromoter activity and associated with 25 non-induced genes (induced Epromoter only). Moreover, from the 394 induced gene only, 16 were associated with a constitutive Epromoter (Supplementary Data 1). Of note, the set of "induced gene & Epromoter" displayed stronger induction than the set of "induced gene only" (Fig. 1d and Supplementary Fig. 2b).

Fig. 1 Comparison of RNA-seq and CapSTARR-seq in non-stimulated and IFNa-stimulated K562 cells. a Experimental approach to assess gene expression and enhancer activity of human promoters in non-stimulated (K562-NS) and IFNa-stimulated (K562 + IFNa) K562 cells. **b** Volcano plot of RNA-seq data in K562. The data plotted is the fold-change (FC) K562 + IFNa over K562-NS (log_2 scale) of the Fragment *per* Kilobase *per* Million (FPKM) and the adjusted (adj.) *P*-value ($-log_{10}$ scale). Genes significantly regulated (adj. *P*-val. < 0.001) are highlighted in purple. Some typical ISGs are indicated. **c** Scatter plot showing the CapSTARR-seq signal in K562-NS and K562 + IFNa. The data plotted is the CapSTARR-seq signal over the input (log_2 scale). IFNa-induced Epromoters (brown), non-induced Epromoters (yellow), and repressed Epromoters (cyan) are highlighted. **d** Scatter-plot of RNA-seq and CapSTARR-seq signals, displaying only IFNa-induced genes or genes harboring an Epromoter. The data plotted is the RNA-seq and the CapSTARR-seq fold-change of K562 + IFNa over K562-NS (log_2 scale). The different categories of genes in function of the IFNa response are indicated. Dot lines indicated a fold-change of 2. **e** Venn diagram showing the overlap between significantly induced genes and induced Epromoters. The number of promoters (bold) and the associated genes (under brackets) is shown. Note that some genes can be associated with multiple categories due to the presence of alternative promoters. **f** Examples of IFNa-induced genes with induced and non-induced Epromoters. The genomic tracks show the RNA-seq and CapSTARR-seq signal in non-stimulated (blue), IFNa-stimulated (red) K562 cells and the fold-change (purple). Positive and negative signs indicate the RNA-seq signal on the sense and anti-sense strands, respectively. Arrows indicate the orientation of the gene transcripts.

Examples of "induced gene & Epromoter" (*SP100*) and "induced gene only" (*GBP4*) loci are shown in Fig. 1f. Taken together these results suggest a potential involvement of Epromoters in the type I interferon response.

IFNa-induced Epromoters specifically recruit the interferon response factors. To better understand the relevance of the three sets of IFNa-stimulated loci, we analyzed the functional enrichment of the associated genes. As expected, the set of "induced gene only" was primarily linked to type I interferon response and related pathways. However, the set of "induced gene & Epromoter" achieved higher enrichment in those pathways even though they represent a minority (9%) of IFNa-response genes (Fig. 2a and Supplementary Data 3). The set of "induced Epromoter only" was not associated with any biological process, likely due to the low number of genes. However, manual inspection of the GO biological processes reveals three genes associated with the interferon response (*HLA-G*, *MVB12A*, *TRIM34*; Supplementary Data 3). Besides, analyses of a dataset of interferon-stimulated genes (ISGs) across different animal species[21], shows that 58% (22) of the genes from the set of "induced gene & Epromoter" overlapped with a conserved set of 98 ISGs, as compared with 12% (47 genes) of the "induced gene only" set (Fig. 2b; *P*-val. = 0.00001, Chi-Square test), suggesting that induced Epromoters are preferentially associated with a conserved interferon response.

Classical type I interferon response requires binding to the ISRE by the regulatory complex formed by phosphorylated STAT1 and STAT2 TFs with the constitutive IRF9 TF also referred to as the ISGF3 complex[16,22]. Subsequent activation can be also mediated by the inducible IRF1 TF. To assess the binding of these TFs to the three sets of IFNa-response promoters, we analyzed previously generated ChIP-seq data in IFNa-stimulated K562 cells for STAT1, STAT2, and IRF1 (Supplementary Data 4) and generated ChIP-seq for IRF9. Strikingly, all four TFs preferentially bound to the set of "induced gene & Epromoter" as compared to the other two sets. In particular, the set of promoters associated with "induced gene only" displayed binding levels close to the background (Fig. 2c). Recruitment of these TFs was specific to IFNa-induced Epromoters as no binding was observed in non-induced Epromoters (Supplementary Fig. 2c). As expected, the majority of induced Epromoters were bound by the ISGF3 complex and were more likely to be associated with a conserved ISGs (Supplementary Fig. 2d; *P*-val. = 0.0002; Chi-Square test).

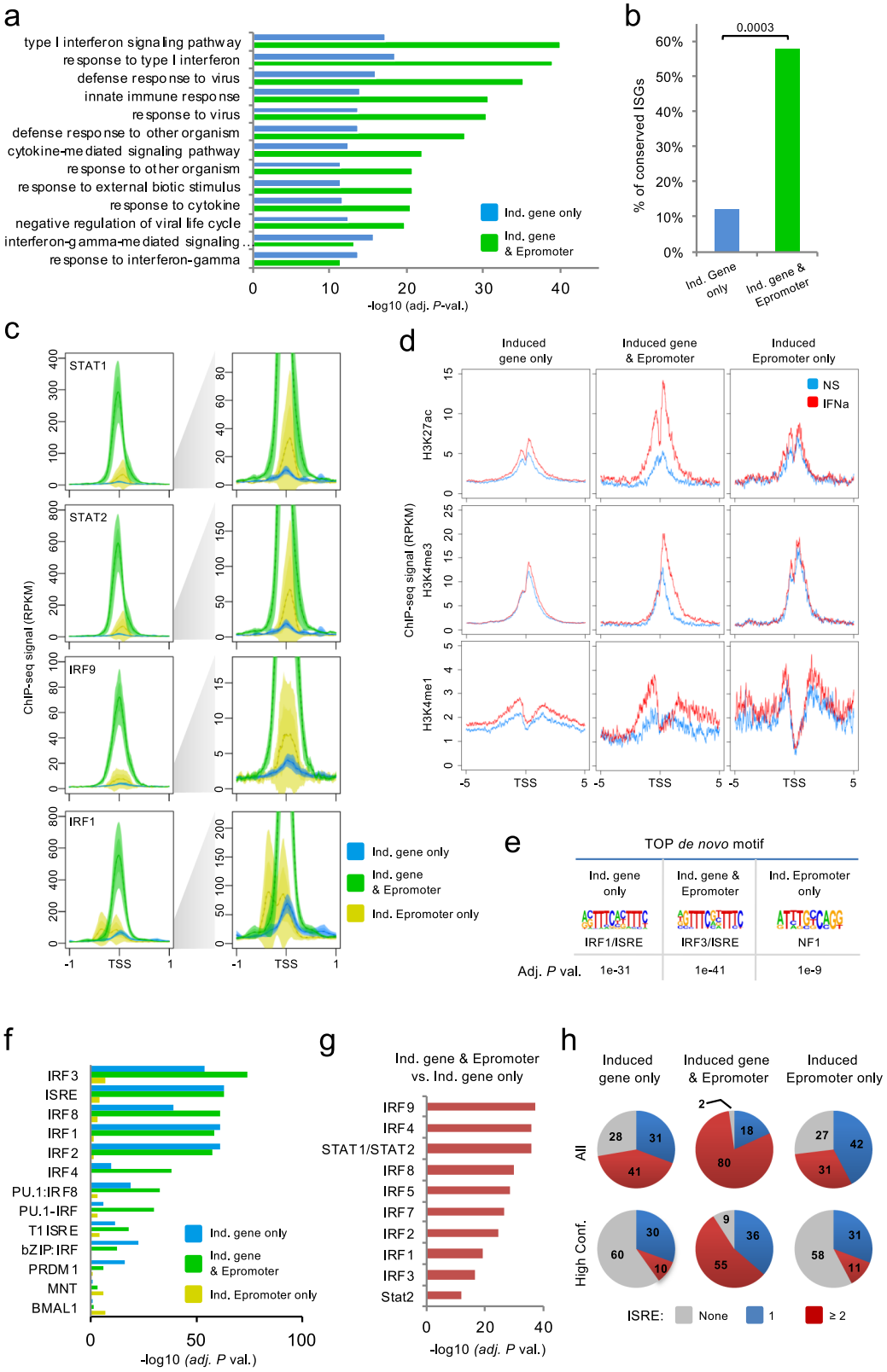
To gain insight into the epigenetic dynamic of these promoters, we performed ChIP-seq for the histone modifications H3K4me3, H3K27ac and H3K4me1 in non-stimulated and IFNa-stimulated K562 cells (Fig. 2d and Supplementary Data 4). For all three histone modifications, the strongest gain of the signal was observed for the set of "induced gene & Epromoter". Thus, the set of "induced gene & Epromoter" has the ability to recruit the key TFs associated with the interferon response and acquires a highly active chromatin state, likely associated with their enhancer activity.

To determine whether the preferential binding of IFN-response TFs to the induced Epromoters was due to an intrinsic feature of the DNA sequence we analyzed the enrichment in TFBS in the three groups of promoters. De novo discovery of DNA motifs (Fig. 2e) as well as global enrichment analyses of known motifs (Fig. 2f), demonstrated that promoters associated with the "induced gene only" and "induced gene & Epromoter" sets, but not with the "induced Epromoter only" set, were similarly enriched in STAT- and IRF-binding sites and generally contain the consensus ISRE. Therefore, the presence of STAT/IRF-binding sites did not explain per se the higher efficiency of TF recruitment observed at induced Epromoters.

Differential enrichment analysis of ISRE-associated binding sites shows that the STAT and IRF motifs were significantly enriched in the "induced gene & Epromoter" set as compared with the "induced gene only" set (Fig. 2g). To assess whether the difference in the density of TFs could be different between the three promoter sets, we first computed the number of sites *per* promoter found for each IRF factor (Supplementary Fig. 3). We observed that the density of IRF sites per promoter was significantly higher in the "induced gene & Epromoter" set as compared with the two other sets.

Next, we computed the number of non-redundant ISRE sites *per* promoter by combining all STAT1-2 and IRF1-9 sites (Fig. 2h, top panels). The "induced gene & Epromoter" set harbor a majority of promoters with 2 or more ISRE sites (80%; 35). In contrast, only 41% (207) and 31% (8) of "induced gene only" and "induced Epromoter only" sets, respectively, displayed 2 or more ISRE sites. The same analysis performed with high confidence sites shows that the majority of the sites found in the promoters of the "induced gene only" and "induced Epromoter only" sets are of lower affinity (Fig. 2h, bottom panels). While 55% (24) of promoters of the "induced gene & Epromoter" set are associated with 2 or more high confidence ISRE sites, only 10% (48) and 11% (3) of the other two sets of promoters contains 2 or more high confidence ISRE sites. This suggests that both the high density and the quality of ISRE sites likely contribute to the efficient recruitment of the STAT1/2 and IRFs complexes and that this is a specific property of the IFNa-stimulated Epromoters.

Clusters of IFNa-induced genes are regulated by Epromoters. The above results showed that the majority of IFNa-responding genes do not efficiently recruit the key IFNa-response TFs. One potential explanation is that the induction of these genes requires the action of distal regulatory elements, which might involve



either typical enhancers or Epromoters. We reasoned that for induction to take place, a given locus needs to be associated with a regulatory element recruiting the ISGF3 complex (i.e., STAT1/STAT2/IRF9). The IFN-response element might be located at the promoter of the same gene or at another regulatory element, which might overlap (Epromoter-like) or not (TSS-distal or "typical" enhancers) the promoter of another gene. To address this issue, we determined the distance of the closest binding of the ISGF3 complex (i.e., overlapping STAT1/STAT2/IRF9 ChIP-seq peaks) to the promoter of all "induced genes only" set and classified it as located within the same promoter, in a TSS-distal region or another promoter (Fig. 3a). As expected only a minority of the promoters associated with the "induced genes only" loci (17.30 %) recruited the ISGF3 complex. For 48.60 %, the closest

Fig. 2 Genomic and epigenomic characteristics of IFNa-induced genes and Epromoters. a Top10 Gene Ontology enrichment for the biological process of genes associated with "induced gene only" (blue) and "induced Epromoter & gene expression" (green). No enrichment was found in the "induced Epromoter only" category. Data plotted are the adjusted *P*-value ($-\log_{10}$ scale) of the enrichment. **b** Percentage of conserved ISGs[21] found in the "induced gene only" and "induced gene & Epromoter" sets. Chi-square test was performed and *P*-values are annotated. **c** Average profiles of ChIP-seq signals for the TFs STAT1, STAT2, IRF9, and IRF1 in K562 cells stimulated with IFNa for 6 h. The groups shown were defined in Fig. 1e. The solid line represents the mean of the signal while the colored area represents the 95% confidence interval. The left panels display the average profiles set at the maximum scale, while the right panels display the same profiles set at a lower scale. **d** Average profiles of ChIP-seq signals for the histone modifications, H3K27ac, H3K4me3 and H3K4me1 in K562 cells non-stimulated (blue) or stimulated with IFNa for 6 h (red). The solid line represents the mean of the signal. **e** Top de novo motifs found enriched in "induced gene only", "induced gene & Epromoter" and "induced Epromoter only" sets, using the HOMER tool[73]. The enrichment adjusted *P*-value is shown. **f** Motif enrichment analysis in promoter regions from the groups defined in Fig. 1E using the HOMER tool. Only the top 10 motifs for each promoter set are shown. The data shown is the adjusted *P*-value ($-\log_{10}$). **g** Comparison of the relative number of binding sites per promoter between the "induced Epromoter and induced gene" versus the "induced genes only" sets. The data shown is the adjusted *P*-value ($-\log_{10}$) of the comparison. **h** Pie-charts showing the percentage of promoters from the indicated datasets that contain either none (gray), one (blue) or two or more ISRE-binding sites (red) (merged sites for IRF1-9 and STA1-2 motifs) based on the Jaspar 2020 annotation[74] by default (All ISRE sites) or with a *P*-value < 1e-4 (High confidence).

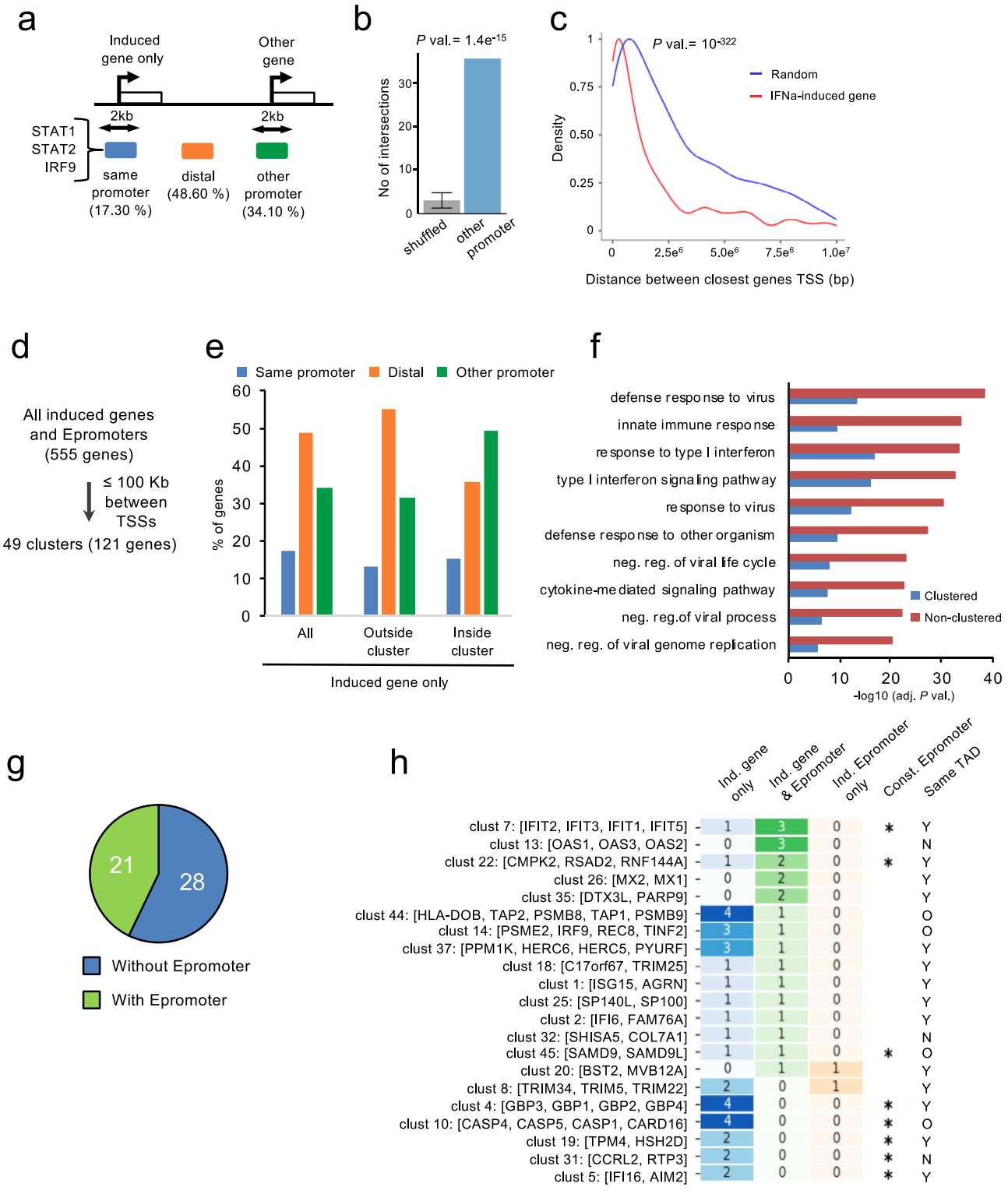
ISGF3 binding was located in a TSS-distal region, indicating that half of the induced genes might be regulated by typical enhancers. However, for more than one-third of induced genes loci (34.10 %), the closest ISGF3 peak was found in another promoter. Importantly, the binding of the ISGF3 complex to a distal promoter was much more frequent than expected by chance (Fig. 3b). To further compare the potential regulation by intergenic and TSS-proximal (either same o other promoters) regions, we analyzed the ChIP-seq profiles of TFs and histone modifications. All three sets of ISGF3-binding regions bound similar levels of IFN-responsive factors (Supplementary Fig. 4a), consistent with the presence of a composite ISGF3 peak. The ISGF3-binding regions displayed increased histone modification levels after IFNa stimulation (Supplementary Fig. 4b). As expected, intergenic ISGF3-bound regions were associated with an enhancer-like chromatin signature (high H3K27ac and H3K4me1), while TSS-proximal ISGF3-bound regions were associated with a promoter-like signature (high H3K27ac and H3K4me3). As described above for the induced Epromoters, the majority of both distal and proximal ISGF3-bound regions contains more than one ISRE site (Supplementary Fig. 4c).

Given the significant frequency of induced genes associated with an ISGF3 complex located in the promoter of another gene, we hypothesized that many induced genes might be located in clusters, likely sharing the same IFNa-stimulated regulatory elements. Therefore, we analyzed the genomic distance between all IFNa-induced genes. We found that induced genes were located closer to each other than expected by chance (Fig. 3c; *P*-val. $= 10^{-322}$; KS test), and frequently to less than 100 kb from a constitutive or induced Epromoter (Supplementary Fig. 5a), suggesting that a subset of IFNa-induced genes are located in clusters and might be co-regulated by Epromoters.

Based on the above observations, we aimed to identify clusters of induced genes and Epromoters for which the corresponding TSS are located less than 100 kb from each other (Fig. 3d). We identified a total of 49 clusters encompassing 121 IFNa-induced loci (Supplementary Data 5). Of 286 "induced gene only" outside clusters, 158 (55.3%) were associated with an ISGF3 complex located in an intergenic region and 90 (31.5%) in another promoter (Fig. 3e). In contrast, of 73 "induced gene only" within clusters, 26 (35.6%) were associated with an ISGF3 complex located in an intergenic region and 36 (49.3%) in another promoter (Fig. 3e). Thus, induced genes within clusters appeared to be preferentially regulated by Epromoters, as compared to induced genes outside clusters (*P*-val. = 0.001, Chi-Square test). Moreover, clustered genes were more enriched in interferon-related biological processes as compared to the unclustered genes (Fig. 3f).

We further explored the regulation of clusters of induced genes by Epromoters. We found that 21 clusters (42%) contain either constitutive or induced Epromoters (Fig. 3g) and were generally located within the same Topological Associated Domain (TAD) in K562 cells (Fig. 3h). Among these clusters, we found typical IFNa-response gene families, such as *OAS*, *IFIT*, *MX*, and *TRIM* (Fig. 3h). The lack of association with Epromoters for the remaining 28 clusters might indicate either a regulation by a typical enhancer or the presence of an Epromoter that was not detected based on our relatively stringent criteria to select induced genes and Epromoters. For instance, the APOBEC3 locus was found to contain three induced genes (*APOBEC3D*, *APOBEC3F*, and *APOBEC3G*), however, it also contains the *APOBEC3C* gene, which has an Epromoter, but the differential expression was slightly under the applied threshold (Supplementary Fig. 5b).

Overall, our finding raises the possibility that a subset of induced genes might be co-regulated by Epromoters located in the same cluster. Consistent with the average patterns of TF binding, visual inspection of several clusters indicated that IFN-response TFs bind preferentially to the induced Epromoters (Fig. 4 and Supplementary Figs. 6 and 7). This is the case of the *ISG15/AGRN/HES4* cluster (Fig. 4a). In this cluster, all three genes are induced by IFNa stimulation, but only *ISG15* is associated with an inducible Epromoter. *ISG15* encodes for an IFN-induced ubiquitin-like protein and plays a central role in the host antiviral response[23]. *HES4* and *AGRN* are also classified as interferon-stimulated genes (ISG) in the Interferome database[24], while *HES4*, encoding for a bHLH (basic helix loop helix) TF, have been suggested to play a role in the IFN response[25,26]. As shown in Fig. 4a, the induced *ISG15* Epromoter, recruits the TFs STAT1, STAT2, IRF1, and IRF9 in IFNa-stimulated K562 cells, while the other induced genes located in the same cluster, *AGRN*, and *HES4*, do not. The three genes are located in the same TAD and no other region, apart from the *ISG15* Epromoter, was found to bind the TFs within the TAD (Supplementary Fig. 6, upper panel). This suggested that the *ISG15* Epromoter might harbor the necessary TFs to induce the transcription of all three genes in an Epromoter-dependent manner. To test this hypothesis, we generated three K562 clones in which the *ISG15* Epromoter has been deleted from all alleles (Supplementary Fig. 7a) and analyzed the gene expression of associated genes after different times of IFNa stimulation (Fig. 4b). As expected, the expression of the *ISG5* gene was completely abolished. Strikingly, the induction of *HES4* was completely abolished while *AGRN* expression was mostly affected at the induction peak (6 h). Although the *AGRN* and *HES4* genes have distinct stimulatory kinetics, the induction of both genes was significantly impaired in ΔEp*ISG15* clones as compared to WT K562 cells. Thus, the *ISG15* Epromoter is required for the accurate induction of the two neighbors' IFNa-responsive genes.



We observed that some of the IFNa-induced clusters contained more than one gene associated with an Epromoter (Fig. 3d). One example is provided by the *IFIT* locus, harboring a family of genes encoding for IFN-induced proteins with tetratricopeptide repeats and have broad-spectrum activity against replication, spread, and disease pathogenesis of a range of human viruses[27]. The cluster contains four IFNa-induced genes; *IFIT3*, *IFIT1*, and *IFIT5* have inducible Epromoters that recruit all four TFs, while *IFIT2* does not have an Epromoter and only recruit the IRF1 factor (Supplementary Fig. 7c; note that *IFIT3* contains two inducible Epromoters, with the most upstream *IFIT3* promoter strongly binding all four IFN-response TFs). All five genes were located in the same TAD (Supplementary Fig. 6, middle panel). We initially generated two K562 clones where the most upstream Epromoter was deleted from all alleles (Supplementary Fig. 7b). The deletion of the upstream *IFIT3* Epromoter resulted in a moderated reduction of *IFIT3* induction. We also observed a significant impairment of *IFIT2* induction at the earliest (2 h) time point (Supplementary Fig. 7d), while the induction of *IFIT1* and *IFIT5* slightly increased. Thus, proper induction of *IFIT2*

Fig. 3 Clustering of IFNa-induced Epromoters and genes. a Schematic diagram showing the percentage of the closest ISGF3 binding with respect to the TSS of the "induced gene only" set. ISGF3-binding peaks in the same promoter (blue; ≤1 Kb from the TSS of "induced gene only"; another promoter (green; ≤1 Kb from the TSS of other genes), or intergenic region (orange; >1 Kb from any TSS of a coding gene). **b** Assessment of the significance of ISGF3 binding in other gene promoters (from Fig. 3a, green) by the OLOGRAM tool[76], the shuffled plot represents the mean and s.d. of 100 random iterations. The observed number of intersections is shown in blue and the mean of 100 shuffled regions are shown in gray. Error bars represent the standard deviation of the shuffled distribution. Statistical significance was calculated against a negative binomial model. **c** The density distribution of the distance between pairs of closest genes using the list of IFNa-induced genes (red) or a set of the same number of randomly selected genes (blue) from the human genome. **d** Summarized diagram of the clustering pipeline of all induced genes and Epromoter-associated genes from which the distance between the TSS was less than 100 Kb was considered to belong to the same cluster. **e** Percentage of the closest ISGF3 binding with respect to the TSS of the "induced gene only" set, inside and outside a cluster (<100 Kb). **f** Top 10 enriched Gene Ontology biological process associated with the "induced gene only" sets that are clustered (blue) or non-clustered (red). Data plotted are the adjusted *P*-value (-$\log_{10}$ scale) of the enrichment. **g** Pie-chart with the number of clusters that contain (green) or not (blue) at least one Epromoter. The detailed list of clusters is provided in Supplementary Data 5. **h** Gene clusters harboring at least one Epromoter. The data shown contain the cluster number and the genes within the cluster. The heatmap shows the number of genes in each category. The induced gene (blue), induced gene & Epromoter (green), and induced Epromoter only (orange) groups are illustrated. The number of genes for each category is indicated and their value is represented by color intensity. The presence of constitutive Epromoters in the clusters is shown by a star (*). Whether the clustered genes were found within the same Topological Associated Domain (TAD) (Y), different TADs (N), or outside any TAD (O) is indicated.

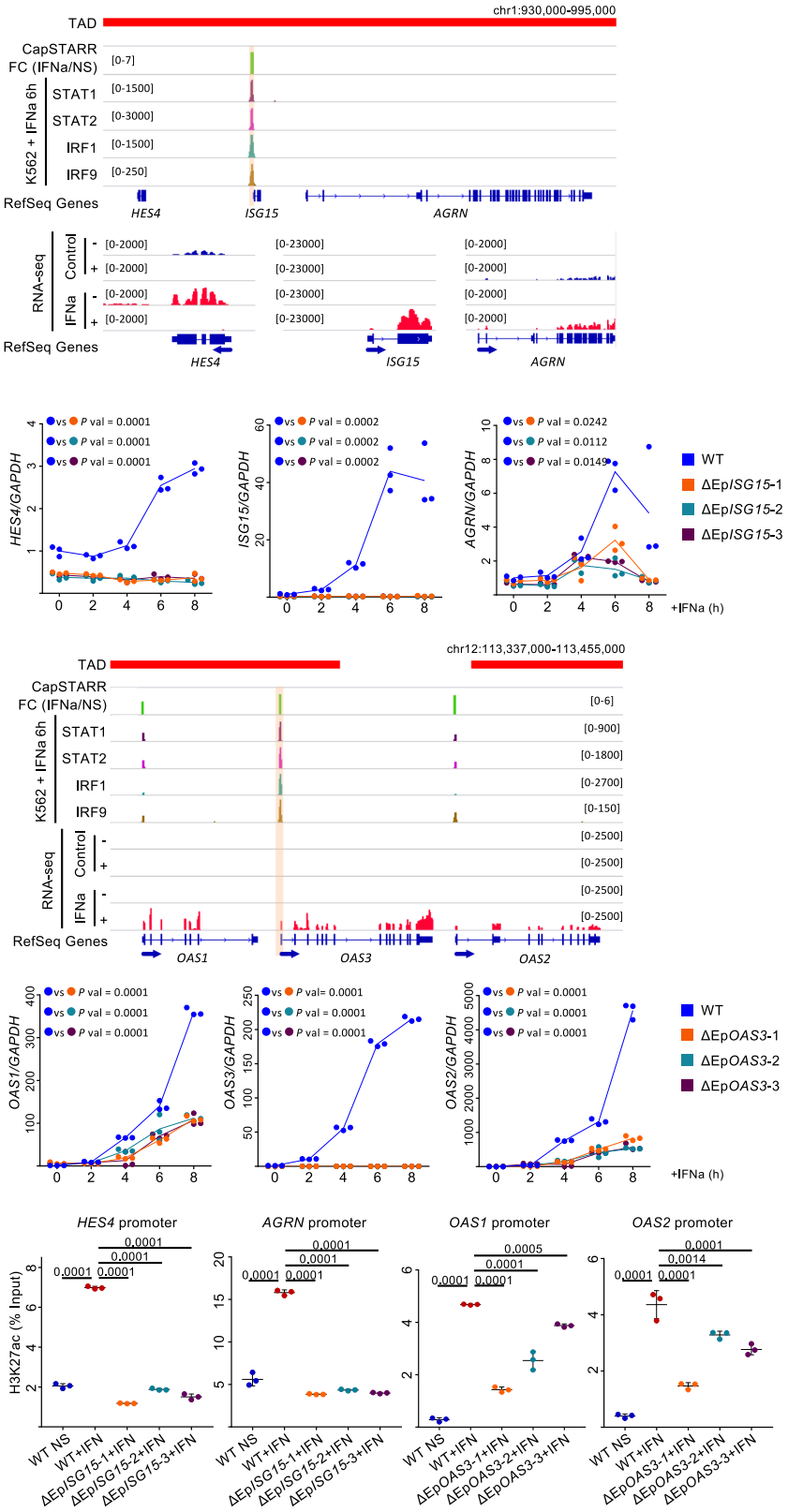
requires the function of at least one induced Epromoter within the cluster. One reason to explain the moderate impact of *IFIT3* Epromoter on *IFIT2* expression would be the synergistic interaction with the internal *IFIT3* Epromoters. Concomitant deletion of the two *IFIT3* Epromoters in two independent K562 clones, completely abolished *IFIT3* expression (Supplementary Fig. 7d). The severity of the effect on *IFIT2* induction varied between the two clones with the double *IFIT3* Epromoter deletion, but in both cases the strongest effect was observed at 2 h after IFNa treatment, consistent with the results obtained with the deletion of the upstream *IFIT3* Epromoter. Also consistent with the single deletions, the induction of *IFIT1* and *IFIT5* significantly increased in the double mutant clones. Therefore, our results suggest that *IFIT3* Epromoter(s) are required for the timely induction of *IFIT2*.

Another noticeable example is seen in the *OAS* locus, harboring three related genes encoding oligoadenylate synthase (OAS) family proteins, and playing a crucial antiviral function[28]. In this cluster, all three IFNa-induced genes were associated with inducible Epromoters (Fig. 4c). The cluster was divided into two TADs in non-stimulated K562 cells (Supplementary Fig. 6, lower panel). However, visual inspection of the Hi-C matrix suggested that the insulation at the TAD boundary within the *OAS* cluster in non-stimulated cells is very weak. We noticed that the *OAS3*-associated Epromoter displayed a higher binding of the four interferon-response TFs, as compared with the *OAS1* and *OAS2* Epromoters. Therefore, we asked whether the specific binding of IFN-response factors was a predictor of the distal regulatory activity of the Epromoter. To explore the specific contribution of each Epromoter towards the regulation of the entire cluster, we generated K562 clones where each of the *OAS* Epromoters have been deleted from all alleles (Supplementary Fig. 8a). The deletion of *OAS1* and *OAS2* Epromoters did not affect the induction of *OAS3* expression (Supplementary Fig. 8b). However, the deletion of the *OAS3* Epromoter resulted in a dramatic reduction of *OAS1* and *OAS2* induction after IFNa stimulation (Fig. 4d). To control for potential bias that might be induced by CRISPR-mediated homologous recombination, we analyzed the induction of *OAS1* and *OAS2* genes after IFNa stimulation from the pool of K562 cells individually transfected with each of the two sgRNAs used for Ep*OAS3* deletion or a clone with an inert insertion within the first intron of *OAS3*. In all three situations, the induction of the three *OAS* genes was not affected (Supplemental Fig. 8c). Finally, the rescue of *OAS3* expression in ΔEp*OAS3* clones did not affect *OAS1* nor *OAS2* induction levels (Supplemental Fig. 8d), indicating direct regulation of neighboring gene expression by the *OAS3* Epromoter. Overall, these results suggested that within a cluster of IFN-induced genes only the promoters that efficiently recruit the key IFN-response factors have the capability to provide a distal regulatory function.

Finally, we explored whether the deletion of the Epromoters also have an impact on the neighbor promoters at the chromatin level. To this end, we analyzed the level of H3K27ac at the *ISG15* and *OAS3* Epromoter-associated promoters by ChIP (Fig. 4e). Deletion of the *ISG15* Epromoter completely abolished the gain of H3K27ac at the *HES4* and *AGRN* promoters observed in wild-type K562 cells after 6 h of IFNa treatment. Deletion of the *OAS3* Epromoter resulted in a less severe, although significant, impairment of H3K27ac after IFNa stimulation. Thus, Epromoters might be required for the chromatin remodeling required for the induction of the IFNa-response associated genes.

Promoter versus enhancer activity of the *OAS3* Epromoter. A key question is whether enhancer and promoter activities are dictated by the same or distinct regulatory sequences. To start addressing this issue we analyzed the TF-binding sites IFNa-inducible *OAS3* Epromoter. The *OAS3* Epromoter harbors two ISRE-binding sites (hereafter $ISRE^1$ and $ISRE^2$), as well as a RELA-binding site (NFkb) in proximity to $ISRE^1$ (Fig. 5a). Notably, $ISRE^1$ corresponds to a canonical IRF9 motif and is closer to the edge of the STAT1/2 and IRF1/9-binding peaks (Fig. 5a, b). We set up a luciferase reporter assay to assess the contribution of each binding site to the enhancer and promoter activities. As expected, the *OAS3* Epromoter displayed IFNa-inducible promoter and enhancer activities in a luciferase assay (Fig. 5c). Mutation of the $ISRE^1$ site, or replacement of $ISRE^1$ by $ISRE^2$, impaired enhancer activity by five-fold and promoter activity by two-fold, while mutation of the $ISRE^2$ or the NFkb sites had no significant effect (Fig. 5c and Supplementary Fig. 8e), suggesting that enhancer activity of the *OAS3* Epromoter is more dependent on the presence of a "consensus" ISRE. The presence of the two ISRE-binding sites, however, was absolutely required for the IFNa-dependent activity. In conclusion, the enhancer and promoter activities of the *OAS3* Epromoter rely on the same regulatory motifs but display different sensitivities with respect to the similarity of the sequences to the consensus ISRE motif.

Epromoters are involved in the regulation of an LPS-induced cluster in macrophages. Based on the above results, we reasoned that within a cluster of induced genes in response to extracellular signaling, the promoter that preferentially recruits the key TFs



can be predicted as having an Epromoter-like function. We used this prediction to explore the function of Epromoters in other inflammatory responses, namely the primary immune response of macrophages. For this purpose, we analyzed a dataset including stimulation of mouse primary macrophages by Lipopolysaccharides (LPS)[29]. The dataset consisted of RNA-seq as well as ChIP-seq for the IRF1, IRF8, and STAT2 TFs before and after LPS stimulation for 4 h. Briefly, we identified clusters of co-induced genes (for which the TSS are separated by less than 100 kb). Then, determined how many promoters from the cluster recruit a given TF (Fig. 6a; see Methods). Clusters harboring only one promoter binding at least one of the three aforementioned TFs, while the other promoters were devoid of binding of any of the TFs, were considered as potentially regulated by Epromoters

Fig. 4 Genomic visualization and kinetic analysis of IFNa-induced genes clustered with induced Epromoters. a, **c** Genomic tracks centered on the *HES4/ISG15/AGRN* (**a**) and *OAS1/OAS2/*OAS3 (**c**) loci. Top panels show the TAD, CapSTARR-seq signal fold-change (IFNa over non-stimulated), and the ChIP-seq signal for the indicated TFs after IFNa induction. The bottom panels show the RNA-seq signals of the induced genes. Positive and negative signs indicate the RNA-seq signal on the sense and anti-sense strands, respectively. Arrows indicate the orientation of the gene transcripts. The induced Epromoter of the *ISG15* and *OAS3* genes are highlighted. **b**, **d** qPCR analysis of gene expression of the indicated genes in K562 *wild-type* ($n = 3$) and ΔEp*ISG15* (**b**, $n = 3$) or ΔEp*OAS3* (**d**, $n = 3$) mutants in non-stimulated conditions or after the indicated times of IFNa stimulation. Values represent the relative expression levels as compared to the unstimulated conditions and normalized by the *GAPDH* housekeeping gene in independent experiments. Two-side ANOVA test was performed between the *wild-type* and each of the mutant clones. **e** H3K27ac ChIP-qPCR analyses in unstimulated and IFNa-stimulated (6 h) *wild-type* K562 cells as well as stimulated ΔEp*ISG15* and ΔEp*OAS3* clones. The promoter regions of the Epromoter-regulated genes were analyzed. Values represent the percentage of input measured in ChIP replicates ($n = 3$). The error bars represent the s.d.. Two-sided Student's *t*-test was performed between the unstimulated and stimulated wild-type cells and between stimulated wild-type cells and each of the mutant clones.

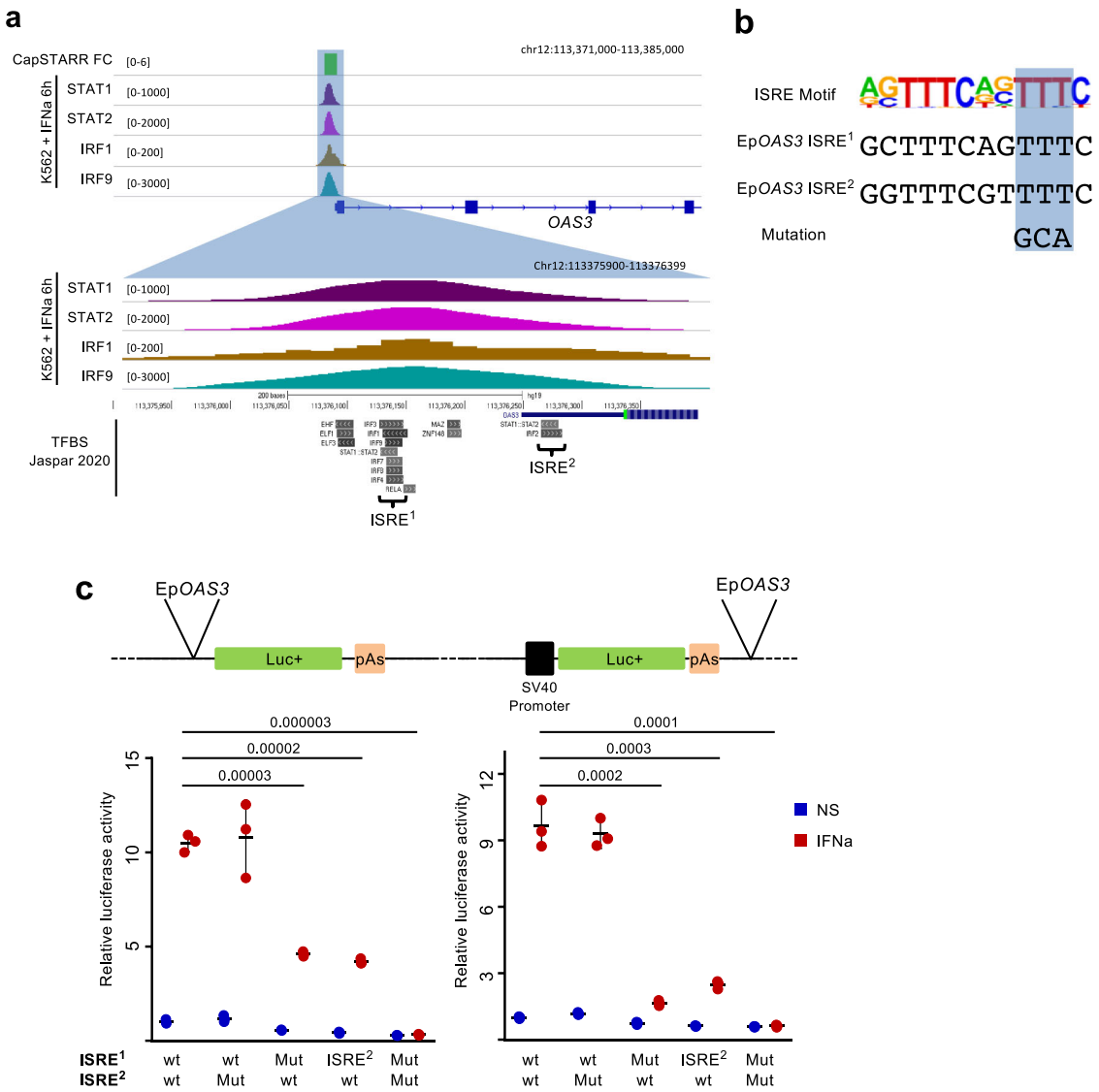
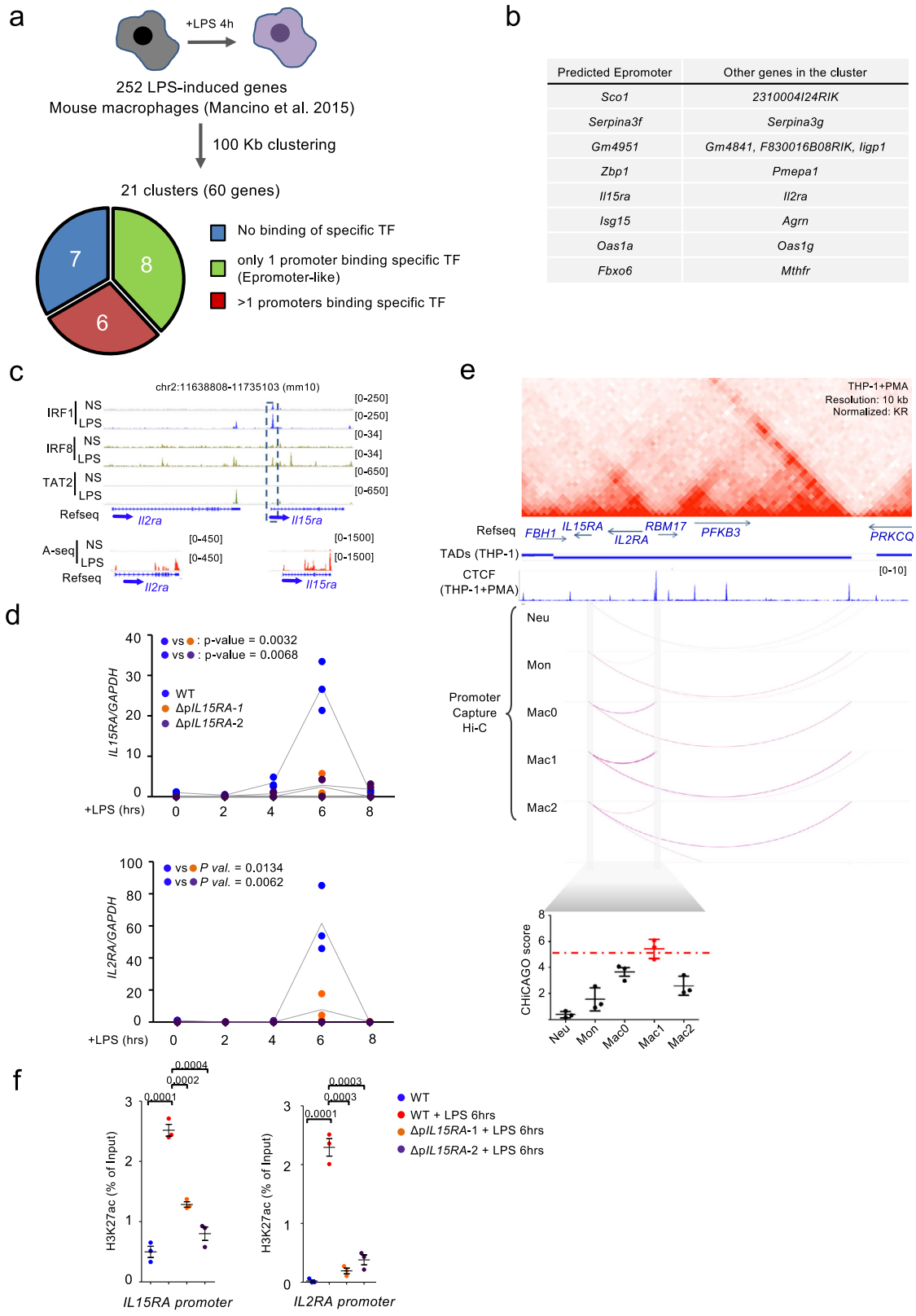


Fig. 5 Luciferase assay of promoter and enhancer activity in wild-type and *OAS3* promoter mutants. a Genome browser tracks showing the fold-change of CapSTARR-seq activity and the ChIP-seq signal of the indicated TFs surrounding the *OAS3* Epromoter. The lower panel shows the location of significant TF-binding sites (Jaspar 2020) and the relative position of the two Interferon-Stimulated Response Element (ISRE) motifs in comparison with the ChIP-seq peaks. **b** Consensus sequence of the ISRE motif, as well as, the genomic sequence of the two identified ISREs in Ep*OAS3*, and the mutation made for the luciferase test is highlighted. **c** Luciferase assays to quantify the promoter (left) and the enhancer (right) activity of the wild-type (wt) *OAS3* Epromoter or with the indicated mutations (Mut) in non-stimulated (blue) or IFNa-stimulated (red) in K562 cells. Central values represent the median of the signal and the error bars represent the s.d. of three independent replicates. *P*-values were calculated by two-sided Student's *t*-test.



Predicted Epromoter	Other genes in the cluster
Sco1	*2310004I24RIK*
Serpina3f	*Serpina3g*
Gm4951	*Gm4841, F830016B08RIK, Iigp1*
Zbp1	*Pmepa1*
Il15ra	*Il2ra*
Isg15	*Agrn*
Oas1a	*Oas1g*
Fbxo6	*Mthfr*

(i.e., Epromoter-clusters). From 252 LPS-induced genes, we identified a total of 21 induced clusters (consisting of 60 genes) (Fig. 6a and Supplementary Data 6). Of these, 8 clusters were classified as Epromoter-like clusters (Fig. 6b). For instance, we identified the *ISG15-AGRN* cluster (Supplementary Fig. 9a), which was also validated as an Epromoter-dependent cluster in the IFNa response.

To further validate the approach, we selected the *Il15ra/Il2ra* cluster (Fig. 6c) coding for the receptors of the IL15 and IL2 cytokines, respectively. IL15RA (also known as CD215) plays a major role in the modulation of the pro-inflammatory response of LPS-stimulated macrophages[30] and in supporting the homeostatic proliferation of CD8 + T lymphocytes[31]. IL2RA (also known as CD25) is mainly known for its role in T cell

Fig. 6 Epromoter-like clusters identification and validation of enhancer activity in LPS stimulated macrophages. a Schematic description of the clustering of the Lipopolysaccharide (LPS) induced genes in the mouse macrophages and pie-chart with the proportion of promoters that are bound by none (blue), 1 (green) or more than one (red) of the transcription factors (TFs) IRF1, IRF8 and STAT2, using data from Mancino et al.[29]. LPS induced genes in the mouse macrophages were clustered together if at least two of their promoters were located within less than 100 kilobases (kb). **b** Table showing the list of the 8 LPS-induced clusters with putative Epromoter elements as defined in **a**. **c** Genomic tracks centered on the mouse *Il15ra/Il2ra* cluster and showing the IRF1 (blue), IRF8 (bright olive) and STAT2 (green) ChIP-seq signal as well as the RNA-seq in unstimulated (NS) and after 4 h of LPS stimulation (LPS) of mouse macrophages. **d** qPCR analysis of gene expression of *IL15RA* and *IL2RA* genes in wild-type (blue) and Δp*IL15RA* mutants 1 (orange) and 2 (purple) of in vitro differentiated THP-1 macrophages in unstimulated and at different intervals of time of LPS stimulation. Values represent the relative expression levels as compared to the unstimulated conditions and normalized by the *GAPDH* housekeeping gene. Points indicate values of $n = 3$ independent experiments and the line in gray present the mean of the values. *P*-values were calculated using Two-Way ANOVA test. **e** Top panel: genomic tracks showing Hi-C triangular matrix (resolution at 10 kb, Knight-Ruiz (KR) normalized) and ChIP-seq CTCF signal in in vitro differentiated human THP-1 macrophages. Topological Associated Domains (TADs) from THP-1 cells[82]. Middle panel: DNA interactions detected by CHiCAGO at the *IL15RA* from Promoter Capture Hi-C from Javierre et al.[36], in human primary neutrophils (Neu), monocytes (Mon), in vitro differentiated macrophages unstimulated (Mac0), stimulated with LPS (Mac1) or with IL-13 (Mac2). Bottom panel: scatter plot showing the mean +/− s.d. of the CHiCAGO score of the three interactions between *IL15RA* to the three bins surrounding the CTCF site downstream *IL2RA* in addition to a red line indicating the threshold. The data was taken from Phanstiel et al.[35]. **f** H3K27ac ChIP-qPCR on *IL15RA* and *IL2RA* promoters in the wild-type NS (in blue) and in the wild-type and Δp*IL15RA* mutants 1 and 2 of in vitro differentiated THP-1 macrophages upon 6 h of stimulation with LPS in red, orange and purple, respectively. Points indicate values of $n = 3$ independent experiments and the scatter plots shows the mean of values +/− s.d. *P*-values were calculated by two-sided Student's *t*-test.

differentiation and activation[32]. Previous studies, however, have observed induction of *IL2RA* and responsiveness to IL2 in LPS-stimulated macrophages[33,34]. This suggests a role of IL2RA in the macrophage-mediated immune response. We observed that the IRF1 and IRF8 factors were bound to the *Il15ra* promoter upon LPS stimulation, but no binding was observed at the *Il2ra* promoter (Fig. 6c). We predicted that the *Il15ra* promoter might work as an Epromoter to coordinate the induction of both *Il15ra* and *Il2ra* genes after the LPS-stimulation of macrophages.

To experimentally validate our hypothesis, we used PMA-induced in vitro differentiated macrophage from the human THP-1 monocyte cell line, a classical model to study macrophage function (Supplementary Fig. 9b, c). We observed that both *IL15RA* and *IL2RA* genes were induced by LPS stimulation of in vitro differentiated THP-1 macrophages with similar kinetics and reached their maximum induction after 6 h (Fig. 6d), in line with the co-regulation observed in mouse macrophages. Next, we examined the 3D chromatin organization of the human IL15RA/IL2RA locus. CTCF binding and Hi-C data in macrophage-differentiated THP-1 cells[35] revealed that both genes are in the same TAD within a regulatory subdomain flanked by CTCF (Fig. 6e). Moreover, analysis of Promoter Capture Hi-C interactions[36] centered on the *IL15RA* promoter showed that the contact frequency (ChICAGO score) with the *IL2RA*-proximal CTCF site increase from monocyte to macrophage differentiation (M0) to reach the highest contact frequency in LPS stimulated macrophages (M1), but not in Il13-stimulated macrophages (M2) (Fig. 6e, bottom panels). Deletion of the *IL15RA* promoter in THP-1 cells (Supplementary Fig. 9d) resulted in significant impairment of both *IL15RA* and *IL2RA* induction after 6 h of LPS stimulation of in vitro differentiated THP-1 macrophages (Fig. 6d). Consistently, the gain of H3K27ac observed at both *IL15RA* and *IL2RA* promoters after 6 h of LPS stimulation of wild-type in vitro differentiated THP-1 macrophages was significantly reduced in *IL15RA* promoter deleted clones (Fig. 6f). Therefore, the *IL15RA* promoter function as a bona fide Epromoter to regulate both *IL15RA* and *IL2RA* induction in LPS stimulated macrophages as predicted by the binding profiles of involved TFs and the dynamic of 3D interactions. More generally, we show that it is possible to predict inducible Epromoter elements based on the binding profile of key TFs and that Epromoter-like regulatory elements play a role not only in type I interferon response, but also in other inflammatory responses, such as macrophage activation.

Discussion

By systematically assessing gene expression and enhancer activity of coding gene promoters in response to type I interferon stimulation we found that a subset of IFNa response genes was associated with Epromoters. IFNa-induced Epromoters associated with an induced gene were found to preferentially recruit the key interferon response factors (STAT1/2, IRF1/9) and to be required for the efficient induction of neighbor genes within the same cluster. Functional studies in LPS-stimulated macrophages suggested that Epromoters play an important role in other inflammatory responses as well.

Previous studies suggested that Epromoters might be involved in stress response and, in particular, the interferon response[5,14,15]. In our previous study, we found that active Epromoters in HeLa cells, which have a constitutive interferon response, are significantly associated with stress and interferon response[5]. Indeed, inhibition of type I interferon in HeLa cells preferentially affects the activity of Epromoters as compared with distal enhancers[14,15] (Supplementary Fig. 1). Moreover, distal genes interacting with Epromoters in both Hela and K562 cells are significantly enriched in the interferon response[5]. Induced Epromoters are associated with highly induced genes and correlate with a higher representation of bona fide interferon-response genes (i.e., conserved ISGs). IFNa-response Epromoters can be found in the proximity of other induced genes and ensure coordinated regulation of interferon response clusters. Notably, IFNa-induced genes with no Epromoters were significantly enriched for interferon-related GO terms arguing that they are also bona fide interferon response genes. We observed that IFNa-induced clusters were generally contained within a single TAD (Fig. 3e), consistent with their potential co-regulation. While unclustered genes are more likely to be regulated by typical enhancers, our results suggest that clustered genes are preferentially regulated by Epromoters. Overall, we suggest that distal regulation by interferon response Epromoters might play an important role in the coordinated response to IFNa-mediated signaling.

A general question about the Epromoter function is related to the mechanistic bases leading to their enhancer activity[2,37]. In a previous study, we showed that Epromoters bind a higher number of TFs and harbor a more complex combination of TF-binding sites as compared with classical promoters[5], suggesting that one of the potential mechanisms mediating enhancer function might be the efficient recruitment of key TFs. Here, we made the striking observation that essential type I interferon-response factors (STAT1/2, IRF1/9) are preferentially recruited by induced

Epromoters, which represents a minority (~8%) of the IFNa-induced promoters in K562 cells. In other words, the vast majority of promoters associated with IFNa-induced genes do not recruit the TFs that are essential for their activation. This is surprising given that it is commonly acknowledged that ISG promoters harbor the ISRE sites and that binding of the ISGF3 complex is required for the induction of ISGs[16]. Moreover, IFNa-induced activation of Epromoters was also associated with a preferential gain of activating histone modifications, H3K4me1, H3K4me3 and H3K27ac, suggesting that they represent the main IFNa-response promoters. Of note, we observed that several clusters of induced genes were associated with a constitutive Epromoter (Fig. 3H). However, constitutive Epromoters were not found to recruit the ISGF3 complex, thus the functional role of these Epromoters in the IFNa response will require further investigation.

Our study suggests that the majority of ISGs are regulated by distal regulatory elements, which could be either typical enhancer or Epromoter associated with another gene. Unbiased analysis of ISGF3 complex binding demonstrated that for a significant subset of IFNa-induced genes, the closest binding of ISGF3 was observed within a promoter of another gene. Indeed, we were able to identify several IFNa-response clusters that are associated with at least one Epromoter. Visual inspection of the epigenetic and TF-binding profiles at typical type I IFN response clusters, such as *IFIT* and *OAS*, suggests that no potential distal "intergenic" enhancers are found within the TADs containing these clusters (Supplementary Fig. 6). Experimental validation by CRISPR/Cas9 genome editing at three independent clusters demonstrated that the Epromoters were required for accurate induction of the ISGs within the same cluster.

Based on the binding of the ISGF3 complex, we suggest that ISGs within clusters are more likely to be regulated by Epromoters while ISGs outside clusters are more likely to be regulated by typical enhancers. Indeed, we found 21 clusters included at least one Epromoter as defined by the STARR-seq. However, the number of genes regulated by Epromoters might be underestimated by several reasons. On the one hand, we have used a stringent criterion of clustering based on a maximum of 100 kb between the TSS of induced genes. Indeed, analyses of TADs identify three additional clusters containing either induced or constitutive Epromoters (DHX58, TMEM140 and SIGLEC14; Supplemental Dataset 5) that were missed by the 100 kb clustering. Yet, it is plausible that clusters of induced genes might be found in a poorly interacting region before stimulation (e.g., clusters found as "outside" TADs in Fig. 3h) or that TAD borders might change upon stimulation. For instances, the OAS1/2/3 cluster was separated by to TADs in unstimulated K562 cells but were found to functionally interact after IFNa stimulation. On the other hand, the number of Epromoters might be underestimated by the STARR-seq approach. It is well accepted that STARR-seq as other similar reporter approaches have a low rate of sensibility[4].

Besides the quantitative aspects about how many ISGs might regulated by Epromoters, our study points to a key role of Epromoters in the interferon response. First, induced genes associated with and induced Epromoter are more significantly associated with conserved ISGs, as compared to other induced genes. This included key interferon response genes such as *OAS1-3*, *IFIT1-5* and *ISG15*. Second, the analyses of distal and proximal enhancers in HeLa cells showed that the activity of Epromoters was significantly more affected than distal enhancers to inhibition of the IFN signal. Third, the binding of key IFN-response TFs and the density of associated binding sites is higher at induced Epromoters as compared to the promoters of other induced genes, but similar to the predicted distal enhancers. Fourth, the increase of histone modifications and gene expression associated with induced Epromoters are significantly higher than at the "induced gene only" set.

What can explain the favored recruitment of STAT-IRF complexes to the induced Epromoters? Although promoters of induced genes are generally enriched in ISRE, consistent with previous results[16], the induced Epromoters were found to contain a higher number of ISRE-binding sites. This can lead to increased efficiency of TF recruitment and/or stabilization required for enhancer function. Features defining the enhancer versus promoter activity of regulatory elements are a fundamental question in the gene regulation field and a focus of extensive research[8,37–42]. Our study suggests that it is both the higher density and the better quality of key TF-binding sites that allow the efficient activation of the Epromoters and mediate its enhancer function. Indeed, mutation of one out of two ISRE in the *OAS3* Epromoter resulted in a more marked effect on its enhancer as compared to its promoter activity. Several studies have also suggested that the degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription[38–43]. We have previously shown that Epromoter activity is associated with the presence of unstable anti-sense transcripts, which are reminiscent of eRNA transcription, as found at typical enhancers[5]. It will be of interest to assess the contribution of eRNA in the function of IFNa-response Epromoters. Overall, our results show that for a subset of IFNa-regulated clusters, the Epromoter functions as a local hub efficiently recruiting the key TFs required for the coordinated regulation of the neighbor ISGs.

The use of high-throughput reporter assays provides the ability to test thousands of sequences in a certain nuclear environment, and pinpoint candidates with enhancer/promoter activity with or without any specific cellular stimulation. However, these activities do not necessarily reflect the actual function of any given sequence in their natural context, which might be influenced by several additional factors, including chromatin composition, histone modifications, DNA modifications, and the presence of other cis-regulatory elements[4]. As highlighted by the results obtained with the *OAS1-3* cluster, it is the preferential in vivo binding of the IFNa-response factors to one of the promoters of the co-regulated genes, within a given cluster, that better predict the enhancer activity in the endogenous locus. Indeed, while all three promoters of the *OAS1-3* cluster displayed enhancer activity when assessed by CapSTARR-seq, only the *OAS3* promoter, which efficiently recruits the STAT1/2 and IRF1/9 factors, was required for the proper induction of the other *OAS* genes. Importantly, analyses of non-deleted clones, as well as *OAS3* rescue experiments further support a direct role of *OAS3* Epromoter in the activation of the other two *OAS* genes. Based on the analyses of RNA-seq and ChIP-seq of STAT2, IRF1, and IRF8 in primary mouse macrophages stimulated by LPS, we identified several LPS-response clusters potentially regulated by Epromoters, including one also identified in K562 cells after IFNa stimulation. Experimental validation of one of these loci (*IL15RA-IL2RA*) demonstrated the predictive value of this approach and suggests that the involvement of inducible Epromoters might also apply to other types of inflammatory responses. This is also consistent with previous findings that promoters highly induced during the immune challenge of macrophages are characterized by enhancer-like features[38,43].

Aside from our current evidence for the role of Epromoters in the interferon and immune response, it is likely that this type of regulatory element could be more generally involved in the rapid response of genes to cellular stress. Noteworthy, many of the early characterized enhancers are located close to, or overlapping with, the promoter region of inducible genes, such as metallothioneins,

histones of early cleavage stages, viral immediate-early genes (from some papovaviruses, cytomegaloviruses, and retroviruses), heat-shock genes and the antiviral interferon genes[3,44]. A common characteristic of most of the aforementioned promoters is that they are associated with inducible genes that have to quickly respond to environmental stress. We hypothesize that, within clusters of co-regulated genes, enhancer-like promoters play an essential role in the coordination of rapid gene induction during the inflammatory response, and likely upon other cellular responses to intra- and extracellular stress signaling. This idea fits well with the notion of inducible transcription factories defining discrete membrane-less sub-nuclear compartments containing a high concentration of RNA polymerase and key TFs and where efficient transcription can be triggered[45,46]. A striking example is provided by NFkB-regulated genes in response to TNF alpha stimulation. Experimental removal of a gene from the NFkB-dependent multi-gene complex was shown to directly affect the transcription of its interacting genes, suggesting that co-association of co-regulated genes might contribute to a hierarchy of gene expression control[47]. The enhancer-like promoters could either facilitate the assembly or maintenance of the transcription factories by tightening promoter-promoter interactions or bringing specific transcriptional regulators required for the regulation of the neighbor genes. This would be particularly relevant in the case of rapid and coordinated regulation of gene expression in response to environmental or intrinsic cellular stimuli. In support of this hypothesis, we found that the *IL15RA* and *IL2RA* co-regulated genes are within the same interacting loop before macrophage stimulation, but the interaction is increased upon LPS stimulation, in agreement with previous results suggesting that a pre-establish loop between cis-regulatory elements can facilitate rapid gene induction[48,49]. Our work provides a framework for future studies aiming to address the contribution of enhancer-like promoters in the formation or stabilization of transcription factories in response to different signaling. In particular, the development of a bioinformatic pipeline based on the analyses of RNA-seq and ChIP-seq data in different stimulatory conditions might allow us to identify clusters of stimulus-regulated genes where only one promoter recruits the key TFs and potentially plays enhancer-like functions.

Our work has general implications for the understanding of genome organization and gene regulation in normal and pathological contexts. First, the coordinated regulation by Epromoters has physiological relevance for the co-regulation of inducible genes as a single regulatory element might ensure the synchronized gene expression and fine-tune the response to extracellular signaling. For instance, in the case of the interferon response, complexity can arise from differences in the endogenous levels of signaling pathway components and IFN properties[50]. These differences may be integrated by Epromoters to coordinate the variations in the nature and number of genes that are transcriptionally up and down-regulated and as a result, lead to distinct biological outcomes.

Secondly, as Epromoters potentially regulate several genes at the same time and have the ability to efficiently recruit essential TFs, mutations in these regulatory elements are expected to have a pleiotropic impact in disease, namely inflammatory diseases. Indeed, several examples point toward the relation between disease-associated variants and disrupted Epromoters[3]. Previous studies suggested that Single Nucleotide Polymorphisms (SNPs) affecting distal gene expression are significantly enriched within Epromoters[5,13,51–53] and that human genetic variation within Epromoters influences distal gene expression[5,54]. Based on our current results, it is expected that Epromoters might have a role in the etiology of inflammatory diseases. Inflammation is viewed as the driving factor in many diseases, including atherosclerosis, cancer, autoimmunity, and infections, and it is a major contributor to age-related conditions[55]. Although IFN signatures are induced by environmental cues, their amplitudes, time courses, and patterns of gene expression are modulated by genetic factors. Allelic variants that regulate gene expression have been associated with complex multigenic autoimmune diseases as well as monogenic interferonopathy, characterized by a type I IFN signature[56]. We forecast that dissecting the consequences of the genetic contributions to the risk of developing a systemic autoimmune disease can be more complex than previously anticipated, given that a variant associated with an Epromoter can have a complex contribution to the interferon response by influencing the expression of different genes involved in the interferon signaling pathway. For instance, a recent study integrating 3D interactions and GWAS found an association between the interferon gene *IFNA2* and a genetic variant lying within the promoter of *CDKN2B* in coronary artery disease[52]. More generally, it is plausible that genetic variants within Epromoters might differentially impact enhancer versus promoter activity. For instance, two studies working on a promoter-overlapping SNP associated with prostate cancer demonstrated that the alternative variant increases the enhancer activity of the promoter leading to increased expression of two distal transcripts directly involved in cancer progression[57,58].

Finally, our findings have significant implications for the understanding of the evolution of regulatory elements and interferon response loci. On the one hand, recent works suggested that repurposing of promoters and enhancers facilitated regulatory innovation and the origination of new genes during evolution[39,59–64]. On the other hand, gene expansion has been shown to significantly contribute to the evolution of the IFN system and suggested to confer a selective advantage to the host species[21]. It is possible that during duplication of ISG, such as in the case of IFIT or OAS genes, ancestral promoters' elements have acquired enhancer functions to coordinate the interferon response of the newly appeared ISG within the cluster. Similarly, the proximity to an Epromoter-associated ISG locus might provide a rapid co-option for neighbor genes to acquire new functions in the interferon response. This is consistent with our results indicating that Epromoters are preferentially associated with conserved ISGs. In this context, it will be interesting to re-examine the dynamic evolution of the interferon responses across different species[21].

Overall, we show that Epromoters work as a local hub for recruiting key TFs required for the activation of type I interferon response genes and are required for the accurate induction of interferon-regulated clusters. Our study suggests that Epromoters play an essential role in the coordination of rapid gene induction in the inflammatory response, and likely upon other cellular responses to intra- and extracellular signals, and to establish connections with other distal response genes.

Methods

Cell culture. Cell line K562 (CCL-243), a chronic myelogenous leukemia cell line, was obtained from the ATCC (American Type Culture Collection) and maintained in RPMI 1640 media (Thermo Fischer Scientific) supplemented with 10% FBS (Thermo Fischer Scientific) at 37 °C and 5% CO_2. Cells were passaged every 3 days at 2×10^5 cells/mL and routinely tested for mycoplasma contamination. Interferon-alpha (Sigma Aldrich, SRP4594) was used to induce the IFN type I response. In all, 1×10^6 wild-type and mutant cells were induced using a concentration of 50 ng/mL of IFNα in 2 mL culture for 6 h. For each group, three independent stimulations were made.

THP-1, a human acute monocytic leukemia cell line, was obtained from DSMZ (ACC 16). Cells were grown in RPMI 1640 media (Thermo Fischer Scientific) supplemented with 10% FBS (Thermo Fischer Scientific) at 37 °C and 5% CO_2. Cells were passaged every 3 days at 10^6 cells/mL and routinely tested for mycoplasma contamination. To induce macrophage differentiation, THP-1 cells were firstly plated on 6-well plates (2×10^6 cells/well), in media containing 10 ng/

mL phorbol 12-myristate 13-acetate (PMA; Sigma-Aldrich, P1585) for 48 h. After 48 h of incubation, the PMA containing media was replaced with fresh media, and cells were incubated for an additional 24 h. THP-1 in vitro differentiated macrophages were then stimulated by replacing media with 2 mL of fresh media containing 100 ng/mL of LPS (Sigma-Aldrich, L9143) for 0, 2, 4, 6 and 8 h, upon which the extraction of RNA was performed. For each group, three independent stimulations were made.

Human promoter CapSTARR-seq. The CapSTARR-seq promoter library used in this study has been generated previously[5] and was based on the capture of the regions −200 to +50 bp relative to the TSS of 20,719 human protein-coding genes as well as 370 random genomic regions using a human genome library with an average size of 300 bp. We transfected 50 millions of K562 cells with 1.25 mg of the CapSTARR-seq promoter library using the Neon Transfection System (Thermo Fisher Scientific; pulse voltage 1450 V, pulse width 10, pulse number 3) and after 18 h we treated the cells with IFNa (50 ng/mL) for 6 h in duplicate. After 24 h, we processed the cells according to the Starr-seq protocol[5,19]. Briefly, transfected (cDNA) and non-transfected (input) libraries were single-end sequenced on an Illumina NextSeq 500 sequencer (Supplementary Data 4). FastQ files were trimmed using sickle with -q 20 option and mapped to the hg19 reference genome using Bowtie2 with default parameters. Sam files were converted using SamTools and bed files were generated with BedTools (v2.17.0) "BamToBed" command. Fragment reads were extended to 314 nt, corresponding to the average size of the captured fragments. Coverage of captured regions was computed using BedTools "coverage" command for both transfected and non-transfected libraries. The coverage was normalized by Fragments per kilobase per million reads mapped (FPKM). The ratio of the Starr-seq coverage over the input (fold-change) was computed for each sample. Promoter regions with enhancer activity were defined using the inflection point of the ranked fold-change as a threshold. Promoter regions with an FPKM < 1 in the input library were removed (421 promoters). Finally, we obtained the enhancer activity of 17,941 promoters regulating the transcription of 14,188 genes (Supplementary Data 1). IFNa-induced Epromoters were defined as those Epromoters with an average ratio between IFNa-stimulated and non-stimulated conditions greater than two and considering that at least one replicate was active. Results are summarized in Supplementary Data 1 and 2.

RNA-seq. Poly(A) RNA was isolated from three replicates of K562 cells non-stimulated or stimulated with IFNa for 6 h and used for RNA-seq library preparation, following the TruSeq RNA Library Prep Kit v2 (Illumina). Libraries were paired-end sequenced on an Illumina NextSeq 500 sequencer (Supplementary Data 4). Reads were aligned using STAR aligner (v2.4.2a) with arguments "outFilterMismatchNoverLmax" and "outFilterMultimapNmax" set to 0.08 and 1, respectively. Differential gene expression was performed using DESeq2 (v1.6.3) Bioconductor package[65] implemented in the R statistical environment, and transcripts associated with promoters analyzed by CapSTARR-seq were retrieved. Genes with changes in expression with a $\log_2$ fold-change >1 and adjusted *P*-value <0.001 were considered as significant Supplementary Data 1 and 2). To create bigwig files reads from Watson and Crick strands were selected using SAMtools (v0.1.9) and provided to the bam2wig.py script from the RSeQC program suite (v2.6.4).

ChIP and ChIP-seq. Cells were cross-linked with 1 % formaldehyde at room temperature for 10 mins followed by quenching with Glycine at 250 mM. Chromatin was sonicated using Bioruptor® Pico from 10 to 20 cycles (30 s ON, 30 s OFF) at 4 °C. The ChIP for the IRF9 TF was performed with 5×10^6 K562 cells stimulated with IFNa for 6 h and 5 µL of IRF9 Rabbit mAB (Cell signal, D2T8M) in 300 µL of ChIP dilution buffer. ChIP for the histone modifications was performed with 5×10^5 of K562 cells non-stimulated and stimulated with IFNa for 6 h, or 5×10^5 of in vitro differentiated THP-1 cells non-stimulated and stimulated with LPS for 6 h, and 1 µg of mAB against H3K27ac, H3K4me3 or H3K4me1 (Diagenode, C15210016, C15410003 and C15410194, respectively) in 300 µL of ChIP dilution buffer. Chromatin was incubated with the antibody for overnight at 4 °C followed by the addition of 20 µL of Dynabeads protein G (Life technologies, 10004D). Bound immune complexes was washed three times and finally incubated in ChIP elution buffer (1% SDS, 0.1 M NaHCO3) and 2 µL of proteinase K (20 mg/mL, Life technologies 25530049) at 65 °C for overnight. DNA was then purified using QIAquick PCR Purification Kit (Quiagen, 28104). qPCR analyses were performed using the SYBR green Master Mix (Thermo Fisher Scientific) on the QuantStudio 6 instrument (Thermo Fisher Scientific). The relative expression was analyzed using the relative standard curve method and values were assessed as percentage of the input. Primers used are listed in Supplementary Data 7. ChIP-seq libraries were generated with the MicroPlex Library Preparation Kit (Diagenode), according to the manufacturer's instructions. The libraries were sequenced in paired-end 50/30nt mode using the NextSeq® 500/550 (Illumina), according to the manufacturer's instructions (Supplementary Data 4). Reads were trimmed with Sickle v1.33 and aligned to the hg19 genome assembly using Bowtie v2.3.4.3 and default parameters. ChIP-Seq coverage tracks (bigwig) were generated with deepTools[66] v3.2.1 using -normalize RPKM. Peaks for ChIP-Seq were called using Macs2. ChIP-seq data for the IRF1, STAT1, and STAT2 TFs in K562 cells stimulated with IFNa were obtained from the ENCODE Consortium (Supplementary Data 4). Median average profiles were generated using the SeqPlots tool[67] for the 2 Kb and 10 Kb regions centered at the TSS of the selected genes for the TFs (IRF9, IRF1, STAT1, and STAT2) and the histone modifications (H3K27ac and H3K4me3), respectively. ChIP-seq profiles were visualized using the IGV genome browser[68].

Functional enrichment. The GO enrichment for biological functions was performed using the webtool GREAT[69] using the hg19 genome and default settings. The background regions used were all the captured promoters. Only the top 10 most significant terms are shown. The full list is provided in Supplementary Data 3.

CRISPR/Cas9 genome editing. For the deletion of *OAS3*, *IFIT3*-P1, *OAS1*, and *OAS2* Epromoters in K562 cells we used the webtool CRISPRdirect[70] to design two guide RNAs flanking the Epromoter region and clone them into the gRNA vector (Addgene #41824)[71]. Two primers were designed flanking the target region that allows us to identify the wild-type and the mutant alleles. For the deletion of Ep*OAS3*, Ep*ISG15*, and Ep*IFIT3* the transfection was made with 1×10^6 K562 cells mixed with 2 µg of hCas9 vector (Addgene #41815) and 1 µg of each gRNA vector using the Neon Transfection System (Thermo Fisher Scientific; pulse voltage 1450 V, pulse width 10, pulse number 3) and cultured in 5 mL. For the deletion of Ep*OAS2* and Ep*OAS1* 1×10^6 K562 cells were transfected with 0.5 µg of each gRNA and 1 µg of Cas9 (Addgene, #41815) in 20 µL transfection solution using the 4D-Nucleofactor X Unit (Lonza), P1 Primary Cell 4D-Nucleofactor X kit S, program FF-120. For the deletion of *ISG15*, and *IFIT3*-P2 Epromoter in K562 cells and for the *IL15RA* promoter in THP-1 cells, gRNAs were designed using CrispRGold[72]. We used the Alt-R CRISPR-Cas9 System from IDT (Integrated DNA Technology) where Ribonucleoprotein (RNP) complexes were assembled in vitro and then transfected by electroporation to the cells, accordingly to the manufactured instructions. Three days after transfection, cells were cultured in 3 × 96-well plates at limit dilution (0.5 cells/100 µL/well). After 2–4 weeks the clones were screened for homologous deletion using the kit Phire Tissue Direct PCR Master Mix (Thermo Fisher Scientific, F170L). Clones with homologous deletion were those showing a mutant band of the expected size and no wild-type band. For the CRISPR controls of the deletion of the *OAS3* Epromoter, we transfected 1×10^6 K562 *wild-type* cells with 1 µg of each of the gRNA used for the *OAS3* Epromoter deletion. Sanger sequencing identified one clone with an inert mutation (single-base insertion) in the first intron of the *OAS3* gene. Primers and gRNAs are listed in Supplementary Data 7.

***OAS3* gene expression recovery**. To recover the expression of the *OAS3* gene, we transfected 1×10^6 K562 Δ*OAS3* clones 1 and 2 with the human ORF clone of *OAS3* (ORIGENE, ref. RC222722) using the Neon Transfection System (Thermo Fisher Scientific; pulse voltage 1450 V, pulse width 10, pulse number 3). We used the expression plasmid of GFP (addgene, #46956) as a transfection negative control. After 2 h, wild-type and transfected cells were induced using a concentration of 50 ng/mL of IFNα in 2 mL culture for 6 h. For each group, three independent stimulations were made.

Gene expression analysis. RNA was extracted using the RNeasy Plus Mini Kit (Qiagen) according to the manufacturer's instructions with DNase treatment. One microgram of total RNA was reverse transcribed using the SuperScript II (Thermo Fisher Scientific). qPCR reactions were made using the SYBR green Master Mix (Thermo Fisher Scientific) on the QuantStudio 6 instrument (Thermo Fisher Scientific). 1:10 dilution of cDNA was used for the qPCR. The relative expression was analyzed using the relative standard curve method and the *GAPDH* gene was used for normalizing samples. For each group, three independent RNA and cDNA preparations were made. For the analyses of *Il15RA* and *IL2RA* genes in the THP-1 cell line, reverse transcription of 2,5 ug of RNA was done using Master Mix SuperScript™ VILO™ (Thermo Fisher Scientific, 11755250). 1:50 dilution of cDNA was used for the qPCR. qPCR reactions were made using the PowerUp SYBR Green Master Mix (Thermo Fisher Scientific, A25742) and the measurement was made using the Applied Biosystems QuantStudio 6 Flex Real-Time PCR System. The relative expression was calculated using the relative standard curve method from the mean of two technical replicates values according to the *GAPDH* gene expression. The mean value and standard deviations were calculated between the relative expression values of the three biological replicates, and values were afterward normalized by the value of the unstimulated WT. Primers used are listed in Supplementary Data 7.

Luciferase reporter assay. The promoter of human *OAS3* (500 bp, Chr12:113375900–113376399) was cloned into the pGL3-Basic vector (Promega, E1751) upstream the luciferase gene at the BglII and HindIII restriction sites (p*OAS3*-luc), and into the pGL3-Promoter vector (Promega, E176; containing the SV40 promoter) downstream the luciferase gene at the BamHI and SalI restriction sites (pSV40-luc-p*OAS3*). Site-specific mutagenesis of the pOAS3 was done using the Q5 site-directed Mutagenesis Kit (NEB, E0554S) using a set of primers listed in Supplementary Data 7. For cell transfection, 1×10^6 K562 cells were mixed with 1 µg of each construct and 200 ng of Renilla

vector using the Neon Transfection System (Thermo Fisher Scientific; pulse voltage 1,450 V, pulse width 10, pulse number 3) and cultured in 2 mL in 12-well plates. After 18 h, half of the cell population was treated with 100 ng of IFNα (Sigma Aldrich, 50 ng/mL) for 6 h. Data were normalized to Renilla values and represented as the fold-change of relative light units over the wild-type pOAS3-luc or pSV40-luc-pOAS3 vector from non-stimulated K652 cells. Experiments were performed in triplicate.

Genomic analyses of interferon-dependent enhancers in Hela cells. Interferon-dependent and independent enhancers in Hela-S3 cells were retrieved from the Muerdter et al. study[15]. The dataset comprises a genome-wide STARR-seq enhancer screening in HeLa-S3 cells treated or not with a combination of two inhibitors of type I interferon response. The annotation of enhancer regions with respect to the closest RefSeq TSS was performed using HOMER's annotatePeaks.pl tool[73]. Based on the distance of active enhancers from the TSS, the genomic density distribution of active enhancers in inhibitor-treated and untreated conditions was calculated. Kolmogorov–Smirnov (KS) test was performed to compare the genomic distribution of active enhancers in the two conditions. Based on the distance to the nearest TSS, the enhancer regions were further divided into two categories: (I) TSS-proximal (≤1 kb from the closest TSS) and (ii) TSS-distal (>1 kb from the closest TSS) enhancer regions. Based on the distance of active enhancers from the TSS, the genomic density distribution of active enhancers in inhibitor-treated and non-treated conditions was calculated. Kolmogorov–Smirnov (KS) test was performed to compare the distribution of active enhancers in the different conditions.

Motif enrichment analyses. Known motif enrichment and de novo motif discovery analyses for TFs-binding sites were performed using the findMotifsGenome.pl program in HOMER suit[73] using the default motif lengths and 200 nt long sequences centered on each enhancer region as input. The ten enriched known motifs with the lowest adjusted *P*-value from each condition and dataset were extracted. For Muerdter et al. dataset, all proximal and distal enhancers for each condition were used as a background. A heatmap was plotted using the $-\log_{10}$ (adjusted *P*-value). Lower *P*-value corresponds to higher enrichment in dark color. For motif enrichment in K562 promoters, all promoters analyzed by CapSTARR-seq were used as a background. The top de novo motif with the lowest *P*-value from each promoter category was demarcated.

ISRE binding across the promoter regions. The identified promoter regions were analyzed with respect to the enrichment of the interferon-stimulated response element (ISRE) and TFs individually. The binding site's information of TFs, IRF family (IRF1–9), and STAT1/2 of the human genome (hg19) were extracted from the latest version of the JASPAR database[74]. The individual TFs-binding regions were generated by bedtools merge (at least 1 base pair overlap), whereas the general ISRE was generated using the bedtools merge in all the IRF family of TFs and STAT1/STAT2. The TF-binding occurrences in the different categories of promoters were computed by counting intersects (bedtool intersect) of individual TFs-binding regions and ISRE across corresponding promoter regions.

Positional distribution of IFNa-specific TFs-binding sites. Overlapping binding sites of STAT1, STAT2, and IRF9 in IFNa-stimulated K562 cells were defined using the 'intersect' parameter from BEDTools v2.28.0[75], based on ChIP-seq data. For each induced gene, we identified the closest merged peak. The location of the closest merged peak was categorized into three groups: (i) the same promoter, located within ±1 kb from the TSS of the induced gene; (ii) intergenic, located at >1 kb from the TSS of any RefSeq gene and; (iii) another promoter, located ±1 kb from the TSS of any other RefSeq gene. Statistical significance of the positional distribution was assessed using the OLOGRAM tool[76] from the Pygtftk package[77].

Proximity of IFNa-induced genes in the human genome. To find the distribution of induced genes in the genome, the same number of IFNa-stimulated genes was extracted randomly from the human genome (hg19). The randomly selected genes were further clustered based on distance (≤100 kb) between the TSS of two induced genes. Further, the distance density distributions of observed and randomly selected genes were compared using Kolmogorov–Smirnov (KS) test.

Identification of IFNa-dependent clusters. The TSS coordinates of IFNa-induced genes or associated with induced Epromoters in K562 cells (as defined in Supplementary Data 2) were extracted with reference to the RefSeq annotations. The distance between each pair of TSSs was computed using a custom Python script. A cluster of IFNa-induced loci was defined if at least two TSS of induced genes or Epromoters were found to be closer than 100 kb. The clusters were then classified in function of the type of genes they contained: "induced gene only", "induced gene & Epromoter", "induced Epromoter only", or whether they contained a constitutive Epromoter (Supplementary Data 1). To assess whether the clustered genes were located within the same TAD, we used the coordinates of TAD borders identified in K562 cells from highly resolutive Hi-C data[78]. Overlapping domains were merged into one single TAD domain as described in[79]. The detailed clusters are summarized in Supplementary Data 5.

Identification of LPS-dependent clusters in primary macrophages. LPS-induced genes ($\log_2$(FC) > 1; *P*-val. <0.01) and binding sites of interferon-associated TFs (IRF1, IRF8, STAT1) were extracted from the Mancino et al. study[29]. The dataset comprises gene expression and TF-binding sites (mm9) in mouse primary macrophages before and after LPS stimulation for 4 h. The TSS coordinates of LPS-induced genes were extracted with reference to the RefSeq annotations. LPS-induced clusters were identified based on distance (≤100 Kb) as described for the IFNa-response clusters. The promoters binding with TF(s) were considered only if the TF-binding site(s) overlapped at least 1 bp with the promoter region (±1 Kb from the TSS). We then determined the number of promoters per cluster that bind a given TF. The detailed description of the LPS-induces clusters is described in Supplementary Data 6.

3D chromatin organization of the *IL15RA/IL2RA* locus. CTCF ChIP-seq and Hi-C data in THP-1 + PMA cells from the data of Phanstiel et al. study[35] were visualized using the New WashU Epigenome Browser (epigenomegateway.wustl.edu)[80]. TADs called from Hi-C experiments (HindIII) THP-1 cells were taken from[81]. DNA interactions centered on the *IL15RA* promoter with the associated CHiCAGO score were taken from published Promoter Capture Hi-C[36] and visualized using the New WashU Epigenome Browser.

Statistical analysis. GraphPad 7.0 was used for the statistical analysis of luciferase, ChIP-qPCR gene expression assays. For ChIP-qPCR and luciferase essays, two-sided Student's *t*-test was performed. Statistical analysis of gene expression assay during different time of stimulation was done using two-way ANOVA test comparing the WT samples to each mutant. *P*-values are mentioned on the figure when it is significant.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support this study are available from the corresponding author upon reasonable request. The RNA-Seq, CapSTARR-seq and ChIP-Seq data generated in this study have been submitted to the Gene Expression Omnibus (GEO) under the accession code GSE159462. Public datasets can be found under accession codes: ENCSR000FAU, ENCSR000FBC, ENCSR000EGL, ENCSR669GJD, GSE183296, GSE63525, GSE56123, GSE96800, GSE89663. All generated and publicly available datasets are listed in the Supplementary Data 4. Public databases used in this study are: Jaspar 2020 [http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=1169499883_qhnKn3jHG401qgfVk2LXUvfvZ2rU&db=hg38&c=chr19&g=hub_186875_JASPAR2020_TFBS_hg38]; Interferome database [http://www.interferome.org/interferome/home.jspx]. Source data are provided with this paper.

Code availability

All custom scripts used in this study are available at GitHub (https://github.com/Spicuglia-Lab/IFN-Data-Analysis) and archived at https://doi.org/10.5281/zenodo.5507612[83].

Received: 23 December 2020; Accepted: 14 October 2021;
Published online: 18 November 2021

References

1. Chatterjee, S. & Ahituv, N. Gene regulatory elements, major drivers of human disease. *Annu. Rev. Genomics Hum. Genet.* https://doi.org/10.1146/annurev-genom-091416-035537 (2017).
2. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.* **21**, 71–87 (2020).
3. Medina-Rivera, A., Santiago-Algarra, D., Puthier, D. & Spicuglia, S. Widespread enhancer activity from core promoters. *Trends Biochem. Sci.* https://doi.org/10.1016/j.tibs.2018.03.004 (2018).
4. Santiago-Algarra, D., Dao, L. T. M., Pradel, L., Espana, A. & Spicuglia, S. Recent advances in high-throughput approaches to dissect enhancer function. *F1000Res* **6**, 939 (2017).
5. Dao, L. T. M. et al. Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat. Genet.* **49**, 1073–1081 (2017).
6. Arnold, C. D. et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
7. Zabidi, M. A. et al. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2015).
8. Nguyen, T. A. et al. High-throughput functional comparison of promoter and enhancer activities. *Genome Res.* **26**, 1023–1033 (2016).
9. Corrales, M. et al. Clustering of Drosophila housekeeping promoters facilitates their expression. *Genome Res.* **27**, 1153–1161 (2017).

10. Engreitz, J. M. et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* https://doi.org/10.1038/nature20149 (2016).
11. Diao, Y. et al. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods* **14**, 629–635 (2017).
12. Rajagopal, N. et al. High-throughput mapping of regulatory DNA. *Nat. Biotechnol.* **34**, 167–174 (2016).
13. Jung, I. et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.* **51**, 1442–1449 (2019).
14. Dao, L. T. M. & Spicuglia, S. Transcriptional regulation by promoters with enhancer function. *Transcription*, https://doi.org/10.1080/21541264.2018.1486150 (2018).
15. Muerdter, F. et al. Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods* **15**, 141–149 (2018).
16. Smale, S. T. & Natoli, G. Transcriptional control of inflammatory responses. *Cold Spring Harb. Perspect. Biol.* **6**, a016261 (2014).
17. Platanias, L. C. Mechanisms of type-I- and type-II-interferon-mediated signalling. *Nat. Rev. Immunol.* **5**, 375–386 (2005).
18. Vanhille, L. et al. High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nat. Commun.* **6**, 6905 (2015).
19. Dao, L. T. M., Vanhille, L., Griffon, A., Fernandez, N. & Spicuglia, S. CapStarr-seq protocol. *Protoc. Exch.* https://doi.org/10.1038/protex.2015.096 (2015).
20. Duncan, C. J. et al. Human IFNAR2 deficiency: Lessons for antiviral immunity. *Sci. Transl. Med.* **7**, 307ra154 (2015).
21. Shaw, A. E. et al. Fundamental properties of the mammalian innate immune system revealed by multispecies comparison of type I interferon responses. *PLoS Biol.* **15**, e2004086 (2017).
22. Chen, K., Liu, J. & Cao, X. Regulation of type I interferon signaling in immunity and inflammation: A comprehensive review. *J. Autoimmun.* **83**, 1–11 (2017).
23. Perng, Y. C. & Lenschow, D. J. ISG15 in antiviral immunity and beyond. *Nat. Rev. Microbiol.* **16**, 423–439 (2018).
24. Rusinova, I. et al. Interferome v2.0: an updated database of annotated interferon-regulated genes. *Nucleic Acids Res.* **41**, D1040–D1046 (2013).
25. Kikuchi, M. et al. Newly identified interferon tau-responsive Hes family BHLH transcription factor 4 and cytidine/uridine monophosphate kinase 2 genes in peripheral blood granulocytes during early pregnancy in cows. *Domest. Anim. Endocrinol.* **68**, 64–72 (2019).
26. Jin, P. et al. Molecular signatures of maturing dendritic cells: implications for testing the quality of dendritic cell therapies. *J. Transl. Med.* **8**, 4 (2010).
27. Diamond, M. S. & Farzan, M. The broad-spectrum antiviral functions of IFIT and IFITM proteins. *Nat. Rev. Immunol.* **13**, 46–57 (2013).
28. Hornung, V., Hartmann, R., Ablasser, A. & Hopfner, K. P. OAS proteins and cGAS: unifying concepts in sensing and responding to cytosolic nucleic acids. *Nat. Rev. Immunol.* **14**, 521–528 (2014).
29. Mancino, A. et al. A dual cis-regulatory code links IRF8 to constitutive and inducible gene expression in macrophages. *Genes Dev.* **29**, 394–408 (2015).
30. Alleva, D. G., Kaser, S. B., Monroy, M. A., Fenton, M. J. & Beller, D. I. IL-15 functions as a potent autocrine regulator of macrophage proinflammatory cytokine production: evidence for differential receptor subunit utilization associated with stimulation or inhibition. *J. Immunol.* **159**, 2941–2951 (1997).
31. Mortier, E. et al. Macrophage- and dendritic-cell-derived interleukin-15 receptor alpha supports homeostasis of distinct CD8+ T cell subsets. *Immunity* **31**, 811–822 (2009).
32. Malek, T. R. The biology of interleukin-2. *Annu. Rev. Immunol.* **26**, 453–479 (2008).
33. Wammers, M. et al. Reprogramming of pro-inflammatory human macrophages to an anti-inflammatory phenotype by bile acids. *Sci. Rep.* **8**, 255 (2018).
34. Valitutti, S. et al. The expression of functional IL-2 receptor on activated macrophages depends on the stimulus applied. *Immunology* **67**, 44–50 (1989).
35. Phanstiel, D. H. et al. Static and dynamic DNA loops form AP-1-bound activation hubs during macrophage development. *Mol. Cell* **67**, 1037–1048 e1036 (2017).
36. Javierre, B. M. et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384 e1319 (2016).
37. Catarino, R. R., Neumayr, C. & Stark, A. Promoting transcription over long distances. *Nat. Genet.* **49**, 972–973 (2017).
38. Henriques, T. et al. Widespread transcriptional pausing and elongation control at enhancers. *Genes Dev.* **32**, 26–41 (2018).
39. Mikhaylichenko, O. et al. The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes Dev.* **32**, 42–57 (2018).
40. Rennie, S. et al. Transcription start site analysis reveals widespread divergent transcription in D. melanogaster and core promoter-encoded enhancer activities. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gky244 (2018).
41. Field, A. & Adelman, K. Evaluating enhancer function and transcription. *Annu. Rev. Biochem.* **89**, 213–234 (2020).
42. Core, L. J. et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).
43. Scruggs, B. S. et al. Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Mol. Cell* **58**, 1101–1112 (2015).
44. Schaffner, W. Enhancers, enhancers-from their discovery to today's universe of transcription enhancers. *Biol. Chem.* **396**, 311–327 (2015).
45. Feuerborn, A. & Cook, P. R. Why the activity of a gene depends on its neighbors. *Trends Genet.: TIG* **31**, 483–490 (2015).
46. Cook, P. R. & Marenduzzo, D. Transcription-driven genome organization: a model for chromosome structure and the regulation of gene expression tested through simulations. *Nucleic Acids Res.* **46**, 9895–9906 (2018).
47. Fanucchi, S., Shibayama, Y., Burd, S., Weinberg, M. S. & Mhlanga, M. M. Chromosomal contact permits transcription between coregulated genes. *Cell* **155**, 606–620 (2013).
48. Jin, F. et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).
49. Cuartero, S. et al. Control of inducible gene expression links cohesin to hematopoietic progenitor self-renewal and differentiation. *Nat. Immunol.* **19**, 932–941 (2018).
50. Schneider, W. M., Chevillotte, M. D. & Rice, C. M. Interferon-stimulated genes: a complex web of host defenses. *Annu. Rev. Immunol.* **32**, 513–545 (2014).
51. Mitchelmore, J., Grinberg, N. F., Wallace, C. & Spivakov, M. Functional effects of variation in transcription factor binding highlight long-range gene regulation by epromoters. *Nucleic Acids Res.* **48**, 2866–2879 (2020).
52. Li, W., Wong, W. H. & Jiang, R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res.* **47**, e60 (2019).
53. Wang, X. et al. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat. Commun.* **9**, 5380 (2018).
54. Sams, D. S. et al. Neuronal CTCF is necessary for basal and experience-dependent gene regulation, memory formation, and genomic structure of BDNF and Arc. *Cell Rep.* **17**, 2418–2430 (2016).
55. Netea, M. G. et al. A guiding map for inflammation. *Nat. Immunol.* **18**, 826–831 (2017).
56. Barrat, F. J., Crow, M. K. & Ivashkiv, L. B. Interferon target-gene expression and epigenomic signatures in health and disease. *Nat. Immunol.* **20**, 1574–1583 (2019).
57. Hua, J. T. et al. Risk SNP-mediated promoter-enhancer switching drives prostate cancer through lncRNA PCAT19. *Cell* **174**, 564–575 e518 (2018).
58. Gao, P. et al. Biology and clinical implications of the 19q13 aggressive prostate cancer susceptibility locus. *Cell* **174**, 576–589 e518 (2018).
59. Wu, X. & Sharp, P. A. Divergent transcription: a driving force for new gene origination? *Cell* **155**, 990–996 (2013).
60. Andersson, R., Sandelin, A. & Danko, C. G. A unified architecture of transcriptional regulatory elements. *Trends Genet.: TIG* **31**, 426–433 (2015).
61. Arenas-Mena, C. The origins of developmental gene regulation. *Evol. Dev.* **19**, 96–107 (2017).
62. Xie, C. et al. Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genet.* **8**, e1002942 (2012).
63. Carelli, F. N., Liechti, A., Halbert, J., Warnefors, M. & Kaessmann, H. Repurposing of promoters and enhancers during mammalian evolution. *Nat. Commun.* **9**, 4066 (2018).
64. Majic, P. & Payne, J. L. Enhancers facilitate the birth of de novo genes and gene integration into regulatory networks. *Mol. Biol. Evolution* **37**, 1165–1178 (2020).
65. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
66. Ramirez, F., Dundar, F., Diehl, S., Gruning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
67. Stempor, P. & Ahringer, J. SeqPlots-Interactive software for exploratory data analyses, pattern discovery and visualization in genomics. *Wellcome Open Res.* **1**, 14 (2016).
68. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
69. McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
70. Naito, Y., Hino, K., Bono, H. & Ui-Tei, K. CRISPRdirect: software for designing CRISPR/Cas guide RNA with reduced off-target sites. *Bioinformatics* **31**, 1120–1123 (2015).
71. Mali, P. et al. RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).

72. Chu, V. T. et al. Efficient CRISPR-mediated mutagenesis in primary immune cells using CrispRGold and a C57BL/6 Cas9 transgenic mouse line. *Proc. Natl Acad. Sci. USA* **113**, 12514–12519 (2016).
73. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
74. Fornes, O. et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).
75. Quinlan, A. R. BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinforma.* **47**, 11 12 11–11 12 34 (2014).
76. Ferre, Q. et al. OLOGRAM: Determining significance of total overlap length between genomic regions sets. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btz810 (2019).
77. Lopez, F. et al. Explore, edit and leverage genomic annotations using Python GTF toolkit. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btz116 (2019).
78. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
79. Paulsen, J. et al. Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. *Genome Biol.* **18**, 21 (2017).
80. Li, D., Hsu, S., Purushotham, D., Sears, R. L. & Wang, T. WashU Epigenome Browser update 2019. *Nucleic Acids Res.* **47**, W158–W165 (2019).
81. Lin, D. et al. Digestion-ligation-only Hi-C is an efficient and cost-effective method for chromosome conformation capture. *Nat. Genet.* **50**, 754–763 (2018).
82. Hong, P. et al. The DLO Hi-C tool for digestion-ligation-only Hi-C chromosome conformation capture data analysis. *Genes* **11**, https://doi.org/10.3390/genes11030289 (2020).
83. Singh, H., Sadouni, N., Charbonnier, G. & Spicuglia, S. Epromoters function as a hub to recruit key transcription factors required for the inflammatory response. *Zenodo*, https://doi.org/10.5281/zenodo.5507612 (2021)

Acknowledgements

We thank the Transcriptomics and Genomics Marseille-Luminy (TGML) platform for sequencing the CapSTARR-seq and ChIP-seq samples and the Marseille-Luminy cell biology platform for the management of cell culture. TGML is a member of the France Genomique consortium (ANR-10-INBS-0009). Work in the laboratory of S. Spicuglia was supported by recurrent funding from Institut National de la Santé et de la Recherche Médicale and Aix-Marseille University and by specific grants from Canceropôle PACA, A*MIDEX (ANR-11-IDEX-0001-02), INCA (PLBIO018-031 INCA_12619), Ligue contre le Cancer (*Equipe Labellisée 2018*), ANR (ANR-18-CE12-0019 and ANR-17-CE12-0035), European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie (grant agreement No 860002), Institut Marseille Maladies Rares (MarMaRa), and Bettencourt Schueller Foundation (*Prix coup d'élan pour la recherche française*). This work was also supported by ECOS/ANUIES/SEP/CONACYT grant M17S02/291235. Work in A. Medina-Rivera's lab was supported by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica – Universidad Nacional Autónoma de México (PAPIIT-UNAM) grant [IA201119]. David Santiago-Algarra was supported by a PhD fellowship from the Conacyt (Mexico).

Author contributions

S.S. conceived and supervised the project and secured funding. D.S., C.S. and S.S. conceptualized and designed the experiments. D.S., C.S., L.T.M.D., and S.H. performed the experiments. H.S. performed the majority of the bioinformatic analyses, with the help of D.S. and C.S. N.S. and G.C. contributed to the processing of N.G.S. data. J.C., A.M.R., and L.R. contributed to motif analyses. D.S., C.S., and S.S. wrote the manuscript. All authors contributed to reading, discussion, and commenting on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-26861-0.

Correspondence and requests for materials should be addressed to Salvatore Spicuglia.

Peer review information *Nature Communications* thanks Gioacchino Natoli, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

4.3 OLOGRAM : Modélisation de la distribution des croisements de données génomiques

Bioinformatics, 36(6), 2020, 1920–1922
doi: 10.1093/bioinformatics/btz810
Advance Access Publication Date: 5 November 2019
Applications Note

Genome analysis

OLOGRAM: determining significance of total overlap length between genomic regions sets

Q. Ferré[1,2,3,†], G. Charbonnier[1,3,†], N. Sadouni[1,3], F. Lopez[1,3], Y. Kermezli[1,3,4], S. Spicuglia[1,3], C. Capponi[2], B. Ghattas[5] and D. Puthier[1,3,*]

[1]Aix Marseille Univ, INSERM, UMR U1090, TAGC, Marseille, France, [2]Aix Marseille Univ, CNRS, UMR 7020, LIS, Qarma, Marseille, France, [3]Equipe Labellisée LIGUE contre le Cancer, [4]The Laboratory of Applied Molecular Biology and Immunology, Tlemcen University, Tlemcen, Algeria and [5]Aix Marseille Univ, CNRS, UMR 7373, IMM, Marseille, France

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Bonnie Berger

Received on May 9, 2019; revised on September 21, 2019; editorial decision on October 25, 2019; accepted on October 30, 2019

Abstract

Motivation: Various bioinformatics analyses provide sets of genomic coordinates of interest. Whether two such sets possess a functional relation is a frequent question. This is often determined by interpreting the statistical significance of their overlaps. However, only few existing methods consider the lengths of the overlap, and they do not provide a resolutive *P*-value.
Results: Here, we introduce *OLOGRAM*, which performs overlap statistics between sets of genomic regions described in BEDs or GTF. It uses Monte Carlo simulation, taking into account both the distributions of region and inter-region lengths, to fit a negative binomial model of the total overlap length. Exclusion of user-defined genomic areas during the shuffling is supported.
Availability and implementation: This tool is available through the command line interface of the *pygtftk* toolkit. It has been tested on Linux and OSX and is available on Bioconda and from https://github.com/dputhier/pygtftk under the GNU GPL license.
Contact: denis.puthier@univ-amu.fr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Current genomic analysis methods can localize a variety of sets of genomic regions, such as epigenomic features, resulting in a BED file giving their coordinates. To determine whether two such sets have a functional relationship, a typical approach is to look for significant co-localization by assessing the statistical significance of the amount of overlap between them (Haiminen *et al.*, 2008).

A comprehensive review of such methods is available through the *Coloc-stats* web interface (Simovski *et al.*, 2018), showing the biggest difference between them to be their null model. Many, such as *GREAT* (McLean *et al.*, 2010) or *CEAS* (Ji *et al.*, 2006) use a binomial test considering only the intersections of the peak centers with the query regions, while *BEDTOOLS fisher* (Quinlan and Hall, 2010) uses the number of intersecting 'bins' (whose size depends on the input regions) to compute a hypergeometric test.

Generating an empirical null distribution by random shuffling of the regions within the sets is another possibility. For example, *pybedtools* incorporates a wrapper for this (Dale *et al.*, 2011) which was also used to tackle the N-fold overlap problem (Aszódi, 2012). For a more realistic null model, conservation of inter-segment length during the shuffling was first proposed by the *Genomic HyperBrowser* (Sandve *et al.*, 2010). However, the *P*-value they provide is only empirical and limited in its resolution by shuffling depth, itself limited by computation time.

Here we propose a new method, implemented in a tool named *OLOGRAM (OverLap Of Genomic Regions Analysis using Monte Carlo)*, to conveniently assess the significance of overlaps by fitting a Negative Binomial model on overlap statistics of interest via a Monte Carlo method.

2 Materials and methods

2.1 Permutation and intersection computation

Let A and B be two sets of genomic regions with no overlaps within A nor B. For each subset $E_{A,k}$ (resp. $E_{B,k}$) of A (resp. B) only for chromosome k, let $L(E_{A,k})$ and $I(E_{A,k})$ be respectively the lists of regions' sizes and inter-regions distances (from end to start).

A shuffle is generated by performing independent random permutations of $L(E_{A,k})$ and $I(E_{A,k})$ for all chromosomes separately, and separately for A and B. This method differs from the classical

BEDTOOLS shuffle which sets regions at random positions. The Genome HyperBrowser showed the relevance of this idea.

Our approach can also exclude regions from the shuffle by shuffling across a shorter, concatenated 'sub-genome' generated by removing the excluded regions from both sets. This allows to compute enrichment relative to the genome minus excluded regions. For example, one can remove low mappability regions, or consider only accessible (i.e. DNAse I HyperSensitive) regions.

The tool then computes the regions' intersections between the ith shuffle of A and the ith of B, for all shuffles. This is done in RAM with a custom sweep-line (Shamos and Hoey, 1976) algorithm of $O(n)$ complexity to avoid disk I/O overhead. As intersections are only computed once per shuffle, the use of other algorithms such as Interval Trees with $O(n\log(n))$ complexity is not justified.

2.2 Discussion of statistical modeling

The null hypothesis (H_0) is that the regions of A are located independently of B. As such, we do not expect them to overlap more than expected by chance, if the regions were independently randomly placed on the genome.

Here, we propose a new statistical framework to model this problem. Under (H_0), for all regions A_i of A and B_j of B, consider the Bernoulli random variables $I_{i,j} = \mathbb{1}_{A_i \cap B_j \neq \varnothing}$.

They have very small probabilities $p_{i,j}$ (region sizes are typically small relative to chromosome size), that differ (each region has a different length, hence different intersection probability), and are dependent (the regions do not overlap).

Let N be the number of intersections and S the total number of overlapping nucleotides. Then $N = \sum_{i,j} I_{i,j}$ is a sum of dependant Bernoulli r.v. and can be modeled with a beta-binomial (Yu and Zelterman, 2008), itself modeled with a Negative Binomial. Unlike with *BEDTOOLS shuffle*, the dependency of the $I_{i,j}$ makes Poisson modeling unadapted.

Then consider $S = \sum_{i,j} \Lambda_{i,j}$ where $\Lambda_{i,j}$ is the length of the intersection between A_i and B_j. This sum has N nonzero terms, making it a Compound Negative Binomial. Furthermore, empirically $\Lambda_{i,j}$ will often follow a logarithmic distribution, so S can be approximated via a negative binomial (Omair *et al.*, 2018).

The assumptions taken here are confirmed in practice by a fitting test. Consequently, we reckon our model is plausible with N and S following negative binomial distributions of under (H_0) unknown parameters, approximated via this Monte Carlo approach. As such, we use them as test statistics: the P-value associated to their value in the true data is used to accept or reject the alternative hypothesis (H_1) that the regions of the query tend to overlap the reference.

3 Implementation

Our method is implemented as a plugin to *pygtftk* (Lopez *et al.*, 2019) and can be passed a GTF/BED stream or file (examples in documentation and Supplementary Data). Most of the code is written in Python 3, with performance-critical operations written in C++ and/or Cython (Behnel *et al.*, 2011). To preserve RAM, the total number of shuffles to be computed is divided into batches.

The tool will compute the overlap between the supplied BED region file and (i) any desired GTF feature, or (ii) features derived from GTF file attributes (e.g. 'gene_biotype'), or (iii) additional regions supplied as BEDs. It will output overlap statistics and the associated P-values.

The computing cost scales with the total number of lines in the reference and query files. A typical pairwise enrichment analysis of 10k regions against 10k takes 62 s on an 2, 5 GHz Intel Core i7 processor. 200k against 200k takes 11 min.

3.1 Results

Supplementary Table S1 presents the applicability conditions and functionalities of various tools and approaches including *GREAT*, *CEAS*, *Bedtools Fisher*, *Genomic HyperBrowser* and *LOLA* (Sheffield and Bock, 2016).

An example of *OLOGRAM* output is available in Supplementary Figure S1. We showcase interactions with *pygtftk* in Supplementary Figure S2, and the importance of considering both S and N in Supplementary Figure S3.

Using biological and artificial testing data, we found both S and N indeed follow a negative binomial distribution; this is shown in particular in Supplementary Figure S4 with the example of S on artificial data. A small total number of shuffles results in a noisy distribution, but whose two first moments (expectation, variance) remain similar than with a larger number of shuffles, making them sufficient to estimate the underlying distributions. We believe 200 shuffles (default parameter) to be an acceptable compromise between computing cost and precision of evaluation in most cases.

Fitting a distribution (as opposed to an empirical P-value) allows for better assessment of extreme overlaps presumably not encountered while shuffling. To confirm the goodness of fit, a fitting quality is given as $1 - V$ where V is Cramér's V score (Cramér, 1946) for the contingency table of observed versus expected histogram bins. It works best when the individual probability of intersection is not too small, meaning the query and reference regions are not too small and/or scarce compared to each other.

We compare our tool to other existing approaches in Supplementary Table S2, showing that OLOGRAM can provide meaningful insights by being resolutive at low P-values. Discussion of those results can be found in Supplementary Note S1. The full code to reproduce the analyses presented is available at: https://github.com/dputhier/ologram_supp_mat, showcasing Snakemake integration.

4 Conclusion

We have implemented a method which allows to consider the information found in the number of overlapping base pairs, with a shuffling paradigm that conserves inter-region length, used to fit a negative binomial model. New features are being developed, including support for multiple overlaps between $n \geq 2$ sets.

Funding

Q.F., G.C., N.S., S.S. and D.P. were supported by recurrent funding from INSERM and Aix Marseille Univ and specific grants from A*MIDEX (A-M-AAP-EI-17-63-170228-17.32-SPICUGLIA-HLS), Institut National du Cancer (PLBIO018-031 INCA_12619) and Ligue contre le Cancer (Equipe Labellisée). Y.K. was supported by the Franco-Algerian partenariat Hubert Curien (PHC) Tassili (15MDU935).

Conflict of Interest: none declared.

References

Aszódi,A. (2012) MULTOVL: fast multiple overlaps of genomic regions. *Bioinformatics*, **28**, 3318–3319.

Behnel,S. *et al.* (2011) Cython: the best of both worlds. *Comput. Sci. Eng.*, **13**, 31–39.

Cramér,H. (1946) *Mathematical Methods of Statistics*. Princeton University Press. ISBN: 0-691-08004-6.

Dale,R.K. *et al.* (2011) Pybedtools: a flexible python library for manipulating genomic datasets and annotations. *Bioinformatics*, **27**, 3423.

Haiminen,N. *et al.* (2008) Determining significance of pairwise co-occurrences of events in bursty sequences. *BMC Bioinformatics*, **9**, 336.

Ji,X. *et al.* (2006) CEAS: cis-regulatory element annotation system. *Nucleic Acids Res.*, **34**, W551–W554.

Lopez,F. *et al.* (2019) Explore, edit and leverage genomic annotations using python GTF toolkit. *Bioinformatics*, **35**, 3487.

McLean,C.Y. *et al.* (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.

Omair,M.A. *et al.* (2018) A bivariate model based on compound negative binomial distribution. *Rev. Colomb. Estad.*, **41**, 87–108.

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

Sandve,G.K. *et al*. (2010) The genomic HyperBrowser: inferential genomics at the sequence level. *Genome Biol*., **11**, R121.

Shamos,M.I. and Hoey,D. (1976). Geometric intersection problems. In: *17th Annual Symposium on Foundations of Computer Science (SFCS 1976)*, pp. 208–215.

Sheffield,N.C. and Bock,C. (2016) LOLA: enrichment analysis for genomic region sets and regulatory elements in R and bioconductor. *Bioinformatics* **32**, 587–589.

Simovski,B. *et al*. (2018) Coloc-stats: a unified web interface to perform colocalization analysis of genomic features. *Nucleic Acids Res*., **46**, W186–W193.

Yu,C. and Zelterman,D. (2008) Sums of exchangeable Bernoulli random variables for family and litter frequency data. *Comput. Stat. Data Anal*., **52**, 1636–1649.

Bibliographie

[Abderrahmani et al., 2001] Abderrahmani, A., Steinmann, M., Plaisance, V., Niederhauser, G., Haefliger, J.-A., Mooser, V., Bonny, C., Nicod, P., and Waeber, G. (2001). The transcriptional repressor rest determines the cell-specific expression of the human mapk8ip1 gene encoding ib1 (jip-1). *Molecular and Cellular Biology*, 21(21) :7256–7267.

[Adoue et al., 2019] Adoue, V., Binet, B., Malbec, A., Fourquet, J., Romagnoli, P., van Meerwijk, J. P., Amigorena, S., and Joffre, O. P. (2019). The histone methyltransferase setdb1 controls t helper cell lineage integrity by repressing endogenous retroviruses. *Immunity*, 50(3) :629–644.

[Alomairi et al., 2020] Alomairi, J., Molitor, A. M., Sadouni, N., Hussain, S., Torres, M., Saadi, W., Dao, L. T., Charbonnier, G., Santiago-Algarra, D., Andrau, J. C., et al. (2020). Integration of high-throughput reporter assays identify a critical enhancer of the ikzf1 gene. *PloS one*, 15(5) :e0233191.

[Amabile et al., 2016] Amabile, A., Migliara, A., Capasso, P., Biffi, M., Cittaro, D., Naldini, L., and Lombardo, A. (2016). Inheritable silencing of endogenous genes by hit-and-run targeted epigenetic editing. *Cell*, 167(1) :219–232.

[Andersson and Sandelin, 2020] Andersson, R. and Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. *Nature Reviews Genetics*, 21(2) :71–87.

[Arnold et al., 2013] Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by starr-seq. *Science*, 339(6123) :1074–1077.

[Aszódi, 2012] Aszódi, A. (2012). Multovl : fast multiple overlaps of genomic regions. *Bioinformatics*, 28(24) :3318–3319.

[Avsec et al., 2021a] Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D. R. (2021a). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10) :1196–1203.

[Avsec et al., 2021b] Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., et al. (2021b). Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53(3) :354–366.

[Bai et al., 2020] Bai, X., Shi, S., Ai, B., Jiang, Y., Liu, Y., Han, X., Xu, M., Pan, Q., Wang, F., Wang, Q., et al. (2020). Endb : a manually curated database of experimentally supported enhancers for human and mouse. *Nucleic Acids Research*, 48(D1) :D51–D57.

[Bailey et al., 2009] Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). Meme suite : tools for motif discovery and searching. *Nucleic acids research*, 37(suppl_2) :W202–W208.

[Bajic et al., 2004] Bajic, V. B., Tan, S. L., Suzuki, Y., and Sugano, S. (2004). Promoter prediction analysis on the whole human genome. *Nature biotechnology*, 22(11) :1467–1473.

[Bakoulis et al., 2022] Bakoulis, S., Krautz, R., Alcaraz, N., Salvatore, M., and Andersson, R. (2022). Endogenous retroviruses co-opted as divergently transcribed regulatory elements shape the regulatory landscape of embryonic stem cells. *Nucleic acids research*, 50(4) :2111–2127.

[Balestrieri et al., 2018] Balestrieri, C., Alfarano, G., Milan, M., Tosi, V., Prosperini, E., Nicoli, P., Palamidessi, A., Scita, G., Diaferia, G. R., and Natoli, G. (2018). Co-optation of tandem dna repeats for the maintenance of mesenchymal identity. *Cell*, 173(5) :1150–1164.

[Banerji et al., 1983] Banerji, J., Olson, L., and Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell*, 33(3) :729–740.

[Banerji et al., 1981] Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote sv40 dna sequences. *Cell*, 27(2) :299–308.

[Bao et al., 2015] Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile Dna*, 6(1) :1–6.

[Barakat et al., 2018] Barakat, T. S., Halbritter, F., Zhang, M., Rendeiro, A. F., Perenthaler, E., Bock, C., and Chambers, I. (2018). Functional dissection of the enhancer repertoire in human embryonic stem cells. *Cell stem cell*, 23(2) :276–288.

[Barra and Fachinetti, 2018] Barra, V. and Fachinetti, D. (2018). The dark side of centromeres : types, causes and consequences of structural abnormalities implicating centromeric dna. *Nature Communications*, 9(1) :1–17.

[Barski et al., 2007] Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4) :823–837.

[Baubec and Schübeler, 2014] Baubec, T. and Schübeler, D. (2014). Genomic patterns and context specific interpretation of dna methylation. *Current opinion in genetics & development*, 25 :85–92.

[Benbarche et al., 2022] Benbarche, S., Lopez, C. K., Salataj, E., Aid, Z., Thirant, C., Laiguillon, M.-C., Lecourt, S., Belloucif, Y., Vaganay, C., Antonini, M., et al. (2022). Screening of eto2-glis2–induced super enhancers identifies targetable cooperative dependencies in acute megakaryoblastic leukemia. *Science advances*, 8(6) :eabg9455.

[Benjamini and Speed, 2012] Benjamini, Y. and Speed, T. P. (2012). Summarizing and correcting the gc content bias in high-throughput sequencing. *Nucleic acids research*, 40(10) :e72–e72.

[Biggar and Crabtree, 2001] Biggar, S. R. and Crabtree, G. R. (2001). Cell signaling can direct either binary or graded transcriptional responses. *The EMBO journal*, 20(12) :3167–3176.

[Bodapati et al., 2020] Bodapati, S., Daley, T. P., Lin, X., Zou, J., and Qi, L. S. (2020). A benchmark of algorithms for the analysis of pooled crispr screens. *Genome biology*, 21(1) :1–13.

[Boland et al., 2014] Boland, M. J., Nazor, K. L., and Loring, J. F. (2014). Epigenetic regulation of pluripotency and differentiation. *Circulation research*, 115(2) :311–324.

[Bonasio et al., 2010] Bonasio, R., Tu, S., and Reinberg, D. (2010). Molecular signals of epigenetic states. *science*, 330(6004) :612–616.

[Bourque et al., 2018] Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., et al. (2018). Ten things you should know about transposable elements. *Genome biology*, 19(1) :1–12.

[Boursot and Bonhomme, 1986] Boursot, P. and Bonhomme, F. (1986). Génétique et évolution du génome mitochondrial des métazoaires. *Génétique sélection evolution*, 18(1) :73–98.

[Boyle et al., 2008] Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2) :311–322.

[Bradner et al., 2017] Bradner, J. E., Hnisz, D., and Young, R. A. (2017). Transcriptional addiction in cancer. *Cell*, 168(4) :629–643.

[Brand et al., 1985] Brand, A. H., Breeden, L., Abraham, J., Sternglanz, R., and Nasmyth, K. (1985). Characterization of a ?silencer ? in yeast : a dna sequence with properties opposite to those of a transcriptional enhancer. *Cell*, 41(1) :41–48.

[Braun et al., 2017] Braun, S. M., Kirkland, J. G., Chory, E. J., Husmann, D., Calarco, J. P., and Crabtree, G. R. (2017). Rapid and reversible epigenome editing by endogenous chromatin regulators. *Nature communications*, 8(1) :1–8.

[Bray et al., 2016] Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5) :525–527.

[Britten and Kohne, 1968] Britten, R. J. and Kohne, D. E. (1968). Repeated sequences in dna : Hundreds of thousands of copies of dna sequences have been incorporated into the genomes of higher organisms. *Science*, 161(3841) :529–540.

[Buenrostro et al., 2013] Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, 10(12) :1213–1218.

[Buenrostro et al., 2015] Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561) :486–490.

[Cai et al., 2021] Cai, Y., Zhang, Y., Loh, Y. P., Tng, J. Q., Lim, M. C., Cao, Z., Raju, A., Lieberman Aiden, E., Li, S., Manikandan, L., et al. (2021). H3k27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. *Nature communications*, 12(1) :1–22.

[Carullo and Day, 2019] Carullo, N. V. and Day, J. J. (2019). Genomic enhancers in brain health and disease. *Genes*, 10(1) :43.

[Castro-Mondragon et al., 2022] Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., et al. (2022). Jaspar 2022 : the 9th release of the open-access database of transcription factor binding profiles. *Nucleic acids research*, 50(D1) :D165–D173.

[Cayrou et al., 2015] Cayrou, C., Ballester, B., Peiffer, I., Fenouil, R., Coulombe, P., Andrau, J.-C., van Helden, J., and Méchali, M. (2015). The chromatin environment shapes dna replication origin organization and defines origin classes. *Genome research*, 25(12) :1873–1885.

[Cesana et al., 2011] Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., and Bozzoni, I. (2011). A long noncoding rna controls muscle differentiation by functioning as a competing endogenous rna. *Cell*, 147(2) :358–369.

[Chatterjee and Ahituv, 2017] Chatterjee, S. and Ahituv, N. (2017). Gene regulatory elements, major drivers of human disease. *Annual review of genomics and human genetics*, 18 :45–63.

[Chaudhri et al., 2020] Chaudhri, V. K., Dienger-Stambaugh, K., Wu, Z., Shrestha, M., and Singh, H. (2020). Charting the cis-regulome of activated b cells by coupling structural and functional genomics. *Nature Immunology*, 21(2) :210–220.

[Chavez et al., 2015] Chavez, A., Scheiman, J., Vora, S., Pruitt, B. W., Tuttle, M., PR Iyer, E., Lin, S., Kiani, S., Guzman, C. D., Wiegand, D. J., et al. (2015). Highly efficient cas9-mediated transcriptional programming. *Nature methods*, 12(4) :326–328.

[Chen et al., 2014] Chen, L., Kostadima, M., Martens, J. H., Canu, G., Garcia, S. P., Turro, E., Downes, K., Macaulay, I. C., Bielczyk-Maczynska, E., Coe, S., et al. (2014). Transcriptional diversity during lineage commitment of human blood progenitors. *Science*, 345(6204) :1251033.

[Chen, 2004] Chen, N. (2004). Using repeat masker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics*, 5(1) :4–10.

[Chong et al., 1995] Chong, J. A., Tapia-Ramirez, J., Kim, S., Toledo-Aral, J. J., Zheng, Y., Boutros, M. C., Altshuller, Y. M., Frohman, M. A., Kraner, S. D., and Mandel, G. (1995). Rest : a mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell*, 80(6) :949–957.

[Chopra et al., 2012] Chopra, V. S., Kong, N., and Levine, M. (2012). Transcriptional repression via antilooping in the drosophila embryo. *Proceedings of the National Academy of Sciences*, 109(24) :9460–9464.

[Chopra and Levine, 2009] Chopra, V. S. and Levine, M. (2009). Combinatorial patterning mechanisms in the drosophila embryo. *Briefings in Functional Genomics and Proteomics*, 8(4) :243–249.

[Chuong et al., 2016] Chuong, E. B., Elde, N. C., and Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science*, 351(6277) :1083–1087.

[Chuong et al., 2017] Chuong, E. B., Elde, N. C., and Feschotte, C. (2017). Regulatory activities of transposable elements : from conflicts to benefits. *Nature Reviews Genetics*, 18(2) :71–86.

[Cieslak et al., 2020] Cieslak, A., Charbonnier, G., Tesio, M., Mathieu, E.-L., Belhocine, M., Touzart, A., Smith, C., Hypolite, G., Andrieu, G. P., Martens, J. H., et al. (2020). Blueprint of human thymopoiesis reveals molecular mechanisms of stage-specific tcr enhancer activation. *Journal of Experimental Medicine*, 217(9).

[Cordaux and Batzer, 2009] Cordaux, R. and Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature reviews genetics*, 10(10) :691–703.

[Cortes-Ciriano et al., 2017] Cortes-Ciriano, I., Lee, S., Park, W.-Y., Kim, T.-M., and Park, P. J. (2017). A molecular portrait of microsatellite instability across multiple cancers. *Nature communications*, 8(1) :1–12.

[Cournac et al., 2016] Cournac, A., Koszul, R., and Mozziconacci, J. (2016). The 3d folding of metazoan genomes correlates with the association of similar repetitive elements. *Nucleic acids research*, 44(1) :245–255.

[Cuellar et al., 2017] Cuellar, T. L., Herzner, A.-M., Zhang, X., Goyal, Y., Watanabe, C., Friedman, B. A., Janakiraman, V., Durinck, S., Stinson, J., Arnott, D., et al. (2017). Silencing of retrotransposons by setdb1 inhibits the interferon response in acute myeloid leukemia. *Journal of Cell Biology*, 216(11) :3535–3549.

[Cusanovich et al., 2015] Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., and Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237) :910–914.

[Dailey, 2015] Dailey, L. (2015). High throughput technologies for the functional discovery of mammalian enhancers : new approaches for understanding transcriptional regulatory network dynamics. *Genomics*, 106(3) :151–158.

[Dalla-Favera et al., 1982] Dalla-Favera, R., Bregni, M., Erikson, J., Patterson, D., Gallo, R. C., and Croce, C. M. (1982). Human c-myc onc gene is located on the region of chromosome 8 that is translocated in burkitt lymphoma cells. *Proceedings of the National Academy of Sciences*, 79(24) :7824–7827.

[Dao et al., 2017] Dao, L. T., Galindo-Albarrán, A. O., Castro-Mondragon, J. A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., Charbonnier, G., Griffon, A., Vanhille, L., Stephen, T., et al. (2017). Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nature genetics*, 49(7) :1073–1081.

[Dao and Spicuglia, 2018] Dao, L. T. and Spicuglia, S. (2018). Transcriptional regulation by promoters with enhancer function. *Transcription*, 9(5) :307–314.

[Davey and Blaxter, 2010] Davey, J. W. and Blaxter, M. L. (2010). Radseq : next-generation population genetics. *Briefings in functional genomics*, 9(5-6) :416–423.

[de Almeida et al., 2021] de Almeida, B. P., Reiter, F., Pagani, M., and Stark, A. (2021). Deepstarr predicts enhancer activity from dna sequence and enables the de novo design of enhancers. *bioRxiv*.

[de Villiers and Schaffner, 1981] de Villiers, J. and Schaffner, W. (1981). A small segment of polyoma virus dna enhances the expression of a cloned β-globin gene over a distance of 1400 base pairs. *Nucleic acids research*, 9(23) :6251–6264.

[Dean, 2011] Dean, A. (2011). In the loop : long range chromatin interactions and gene regulation. *Briefings in functional genomics*, 10(1) :3–10.

[Della Rosa and Spivakov, 2020] Della Rosa, M. and Spivakov, M. (2020). Silencers in the spotlight. *Nature Genetics*, 52(3) :244–245.

[Deng and Roberts, 2005] Deng, W. and Roberts, S. G. (2005). A core promoter element downstream of the tata box that is recognized by tfiib. *Genes & development*, 19(20) :2418–2423.

[Depledge et al., 2019] Depledge, D. P., Srinivas, K. P., Sadaoka, T., Bready, D., Mori, Y., Placantonakis, D. G., Mohr, I., and Wilson, A. C. (2019). Direct rna sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nature communications*, 10(1) :1–13.

[Derrien et al., 2012] Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., et al. (2012). The gencode v7 catalog of human long noncoding rnas : analysis of their gene structure, evolution, and expression. *Genome research*, 22(9) :1775–1789.

[Di Tommaso et al., 2017] Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4) :316–319.

[Diao et al., 2017] Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., Lin, K. C., Huang, H., Liu, T., Marina, R. J., et al. (2017). A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nature methods*, 14(6) :629–635.

[Dietrich et al., 2012] Dietrich, N., Lerdrup, M., Landt, E., Agrawal-Singh, S., Bak, M., Tommerup, N., Rappsilber, J., Södersten, E., and Hansen, K. (2012). Rest–mediated recruitment of polycomb repressor complexes in mammalian cells. *PLoS genetics*, 8(3) :e1002494.

[Ding et al., 2008] Ding, N., Zhou, H., Esteve, P.-O., Chin, H. G., Kim, S., Xu, X., Joseph, S. M., Friez, M. J., Schwartz, C. E., Pradhan, S., et al. (2008). Mediator links epigenetic silencing of neuronal gene expression with x-linked mental retardation. *Molecular cell*, 31(3) :347–359.

[DiStefano, 2018] DiStefano, J. K. (2018). The emerging role of long noncoding rnas in human disease. *Disease Gene Identification*, pages 91–110.

[Dobin et al., 2013] Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star : ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1) :15–21.

[Doench, 2018] Doench, J. G. (2018). Am i ready for crispr ? a user's guide to genetic screens. *Nature Reviews Genetics*, 19(2) :67–80.

[Donda et al., 1996] Donda, A., Schulz, M., Bürki, K., De Libero, G., and Uematsu, Y. (1996). Identification and characterization of a human cd4 silencer. *European journal of immunology*, 26(2) :493–500.

[Doni Jayavelu et al., 2020] Doni Jayavelu, N., Jajodia, A., Mishra, A., and Hawkins, R. D. (2020). Candidate silencer elements for the human and mouse genomes. *Nature communications*, 11(1) :1–15.

[Douse et al., 2020] Douse, C. H., Tchasovnikarova, I. A., Timms, R. T., Protasio, A. V., Seczynska, M., Prigozhin, D. M., Albecka, A., Wagstaff, J., Williamson, J. C., Freund, S., et al. (2020). Tasor is a pseudo-parp that directs hush complex assembly and epigenetic transposon control. *Nature communications*, 11(1) :1–16.

[Dreos et al., 2015] Dreos, R., Ambrosini, G., Périer, R. C., and Bucher, P. (2015). The eukaryotic promoter database : expansion of epdnew and new promoter analysis tools. *Nucleic acids research*, 43(D1) :D92–D96.

[Drongitis et al., 2019] Drongitis, D., Aniello, F., Fucci, L., and Donizetti, A. (2019). Roles of transposable elements in the different layers of gene expression regulation. *International Journal of Molecular Sciences*, 20(22) :5755.

[Eagen, 2018] Eagen, K. P. (2018). Principles of chromosome architecture revealed by hi-c. *Trends in biochemical sciences*, 43(6) :469–478.

[Ecco et al., 2017] Ecco, G., Imbeault, M., and Trono, D. (2017). Krab zinc finger proteins. *Development*, 144(15) :2719–2729.

[Engreitz et al., 2016] Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., McDonel, P. E., Guttman, M., and Lander, E. S. (2016). Local regulation of gene expression by lncrna promoters, transcription and splicing. *Nature*, 539(7629) :452–455.

[Eraslan et al., 2019] Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. (2019). Deep learning : new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7) :389–403.

[Ernst and Kellis, 2012] Ernst, J. and Kellis, M. (2012). Chromhmm : automating chromatin-state discovery and characterization. *Nature methods*, 9(3) :215–216.

[Ernst et al., 2016] Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T. S., and Kellis, M. (2016). Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nature biotechnology*, 34(11) :1180–1190.

[Essletzbichler et al., 2014] Essletzbichler, P., Konopka, T., Santoro, F., Chen, D., Gapp, B. V., Kralovics, R., Brummelkamp, T. R., Nijman, S. M., and Bürckstümmer, T. (2014). Megabase-scale deletion using crispr/cas9 to generate a fully haploid human cell line. *Genome research*, 24(12) :2059–2065.

[Ferré et al., 2021] Ferré, Q., Capponi, C., and Puthier, D. (2021). Ologram-modl : mining enriched n-wise combinations of genomic features with monte carlo and dictionary learning. *NAR genomics and bioinformatics*, 3(4) :lqab114.

[Fornes et al., 2020] Fornes, O., Castro-Mondragon, J. A., Khan, A., Van der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., et al. (2020). Jaspar 2020 : update of the open-access database of transcription factor binding profiles. *Nucleic acids research*, 48(D1) :D87–D92.

[Fueyo et al., 2022] Fueyo, R., Judd, J., Feschotte, C., and Wysocka, J. (2022). Roles of transposable elements in the regulation of mammalian transcription. *Nature Reviews Molecular Cell Biology*, pages 1–17.

[Fulco et al., 2016] Fulco, C. P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S. R., Perez, E. M., Kane, M., Cleary, B., Lander, E. S., and Engreitz, J. M. (2016). Systematic mapping of functional enhancer–promoter connections with crispr interference. *Science*, 354(6313) :769–773.

[Gasperini et al., 2020] Gasperini, M., Tome, J. M., and Shendure, J. (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics*, 21(5) :292–310.

[Gaszner and Felsenfeld, 2006] Gaszner, M. and Felsenfeld, G. (2006). Insulators : exploiting transcriptional and epigenetic mechanisms. *Nature Reviews Genetics*, 7(9) :703–713.

[Gemayel et al., 2010] Gemayel, R., Vinces, M. D., Legendre, M., and Verstrepen, K. J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual review of genetics*, 44 :445–477.

[Georgopoulos, 2017] Georgopoulos, K. (2017). The making of a lymphocyte : the choice among disparate cell fates and the ikaros enigma. *Genes & Development*, 31(5) :439–450.

[Gilbert et al., 2013] Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., Stern-Ginossar, N., Brandman, O., Whitehead, E. H., Doudna, J. A., et al. (2013). Crispr-mediated modular rna-guided regulation of transcription in eukaryotes. *Cell*, 154(2) :442–451.

[Gillies et al., 1983] Gillies, S. D., Morrison, S. L., Oi, V. T., and Tonegawa, S. (1983). A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell*, 33(3) :717–728.

[Giorgetti et al., 2010] Giorgetti, L., Siggers, T., Tiana, G., Caprara, G., Notarbartolo, S., Corona, T., Pasparakis, M., Milani, P., Bulyk, M. L., and Natoli, G. (2010). Noncooperative interactions between transcription factors and clustered dna binding sites enable graded transcriptional responses to environmental inputs. *Molecular cell*, 37(3) :418–428.

[Giresi et al., 2007] Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., and Lieb, J. D. (2007). Faire (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome research*, 17(6) :877–885.

[Gisselbrecht et al., 2020] Gisselbrecht, S. S., Palagi, A., Kurland, J. V., Rogers, J. M., Ozadam, H., Zhan, Y., Dekker, J., and Bulyk, M. L. (2020). Transcriptional silencers in drosophila serve a dual role as transcriptional enhancers in alternate cellular contexts. *Molecular cell*, 77(2) :324–337.

[Gong and Maquat, 2011] Gong, C. and Maquat, L. E. (2011). lncrnas transactivate stau1-mediated mrna decay by duplexing with 3 ? utrs via alu elements. *Nature*, 470(7333) :284–288.

[Goodbourn et al., 1985] Goodbourn, S., Zinn, K., and Maniatis, T. (1985). Human β-interferon gene expression is regulated by an inducible enhancer element. *Cell*, 41(2) :509–520.

[Gorzynski et al., 2022] Gorzynski, J. E., Goenka, S. D., Shafin, K., Jensen, T. D., Fisk, D. G., Grove, M. E., Spiteri, E., Pesout, T., Monlong, J., Baid, G., et al. (2022). Ultrarapid nanopore genome sequencing in a critical care setting. *The New England journal of medicine*.

[Grapotte et al., 2021] Grapotte, M., Saraswat, M., Bessière, C., Menichelli, C., Ramilowski, J. A., Severin, J., Hayashizaki, Y., Itoh, M., Tagami, M., Murata, M., et al. (2021). Discovery of widespread transcription initiation at microsatellites predictable by sequence-based deep neural network. *Nature communications*, 12(1) :1–18.

[Griffiths et al., 2000] Griffiths, A. J., Miller, J. H., Suzuki, D. T., Lewontin, R. C., and Gelbart, W. M. (2000). Structure of dna. In *An Introduction to Genetic Analysis. 7th edition*. WH Freeman.

[Grüning et al., 2018] Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., and Köster, J. (2018). Bioconda : sustainable and comprehensive software distribution for the life sciences. *Nature methods*, 15(7) :475–476.

[Gualdrini et al., 2022] Gualdrini, F., Polletti, S., Simonatto, M., Prosperini, E., Pileri, F., and Natoli, G. (2022). H3k9 trimethylation in active chromatin restricts the usage of functional ctcf sites in sine b2 repeats. *Genes & Development*.

[Gupta et al., 2017] Gupta, R. M., Hadaya, J., Trehan, A., Zekavat, S. M., Roselli, C., Klarin, D., Emdin, C. A., Hilvering, C. R., Bianchi, V., Mueller, C., et al. (2017). A genetic variant associated with five vascular diseases is a distal regulator of endothelin-1 gene expression. *Cell*, 170(3) :522–533.

[Haberle and Stark, 2018] Haberle, V. and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nature reviews Molecular cell biology*, 19(10) :621–637.

[Halfon, 2020] Halfon, M. S. (2020). Silencers, enhancers, and the multifunctional regulatory genome. *Trends in Genetics*, 36(3) :149–151.

[Hansen and Hodges, 2022] Hansen, T. J. and Hodges, E. (2022). Identifying transcription factor-bound gene activators and silencers in the chromatin accessible human genome using atac-starr-seq. *bioRxiv*.

[Hao et al., 2015] Hao, B., Naik, A. K., Watanabe, A., Tanaka, H., Chen, L., Richards, H. W., Kondo, M., Taniuchi, I., Kohwi, Y., Kohwi-Shigematsu, T., et al. (2015). An anti-silencer–and satb1-dependent chromatin hub regulates rag1 and rag2 gene expression during thymocyte development. *Journal of Experimental Medicine*, 212(5) :809–824.

[Harris et al., 2005] Harris, M. B., Mostecki, J., and Rothman, P. B. (2005). Repression of an interleukin-4-responsive promoter requires cooperative bcl-6 function. *Journal of Biological Chemistry*, 280(13) :13114–13121.

[Hause et al., 2016] Hause, R. J., Pritchard, C. C., Shendure, J., and Salipante, S. J. (2016). Classification and characterization of microsatellite instability across 18 cancer types. *Nature medicine*, 22(11) :1342–1350.

[He et al., 2005] He, X., He, X., Dave, V. P., Zhang, Y., Hua, X., Nicolas, E., Xu, W., Roe, B. A., and Kappes, D. J. (2005). The zinc finger transcription factor th-pok regulates cd4 versus cd8 t-cell lineage commitment. *Nature*, 433(7028) :826–833.

[Heinz et al., 2010] 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. *Molecular cell*, 38(4) :576–589.

[Heizmann et al., 2018] Heizmann, B., Kastner, P., and Chan, S. (2018). The ikaros family in lymphocyte development. *Current opinion in immunology*, 51 :14–23.

[Henson et al., 2014] Henson, D. M., Chou, C., Sakurai, N., and Egawa, T. (2014). A silencer-proximal intronic region is required for sustained cd4 expression in postselection thymocytes. *The Journal of Immunology*, 192(10) :4620–4627.

[Heyer et al., 2019] Heyer, E. E., Deveson, I. W., Wooi, D., Selinger, C. I., Lyons, R. J., Hayes, V. M., O ?Toole, S. A., Ballinger, M. L., Gill, D., Thomas, D. M., et al. (2019). Diagnosis of fusion genes using targeted rna sequencing. *Nature communications*, 10(1) :1–12.

[Hilton et al., 2015] Hilton, I. B., D'ippolito, A. M., Vockley, C. M., Thakore, P. I., Crawford, G. E., Reddy, T. E., and Gersbach, C. A. (2015). Epigenome editing by a crispr-cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature biotechnology*, 33(5) :510–517.

[Horii et al., 2013] Horii, T., Morita, S., Kimura, M., Kobayashi, R., Tamura, D., Takahashi, R.-u., Kimura, H., Suetake, I., Ohata, H., Okamoto, K., et al. (2013). Genome engineering of mammalian haploid embryonic stem cells using the cas9/rna system. *PeerJ*, 1 :e230.

[Hornung et al., 2014] Hornung, V., Hartmann, R., Ablasser, A., and Hopfner, K.-P. (2014). Oas proteins and cgas : unifying concepts in sensing and responding to cytosolic nucleic acids. *Nature Reviews Immunology*, 14(8) :521–528.

[Huang et al., 2019] Huang, D., Petrykowska, H. M., Miller, B. F., Elnitski, L., and Ovcharenko, I. (2019). Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. *Genome research*, 29(4) :657–667.

[Huang et al., 1999] Huang, Y., Myers, S. J., and Dingledine, R. (1999). Transcriptional repression by rest : recruitment of sin3a and histone deacetylase to neuronal genes. *Nature neuroscience*, 2(10) :867–872.

[Huang et al., 2021] Huang, Z., Liang, N., Goñi, S., Damdimopoulos, A., Wang, C., Ballaire, R., Jager, J., Niskanen, H., Han, H., Jakobsson, T., et al. (2021). The corepressors gps2 and smrt control enhancer and silencer remodeling via erna transcription during inflammatory activation of macrophages. *Molecular cell*, 81(5) :953–968.

[Humphrey et al., 2001] Humphrey, G. W., Wang, Y., Russanova, V. R., Hirai, T., Qin, J., Nakatani, Y., and Howard, B. H. (2001). Stable histone deacetylase complexes distinguished by the presence of sant domain proteins corest/kiaa0071 and mta-l1. *Journal of Biological Chemistry*, 276(9) :6817–6824.

[Hussain et al.,] Hussain, S., Sadouni, N., van Essen, D., Lopez, P., Dao, L., Charbonnier, G., Torres, M., Lecellier, C., Sexton, T., Saccani, S., and Spicuglia, S. (.). Znf263 tandem repeats are important contributors of silencer elements in t cells. *In submission*.

[Ichiyanagi et al., 2021] Ichiyanagi, T., Katoh, H., Mori, Y., Hirafuku, K., Boyboy, B. A., Kawase, M., and Ichiyanagi, K. (2021). B2 sine copies serve as a transposable boundary of dna methylation and histone modifications in the mouse. *Molecular biology and evolution*, 38(6) :2380–2395.

[Inoue and Ahituv, 2015] Inoue, F. and Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. *Genomics*, 106(3) :159–164.

[Inukai et al., 2017] Inukai, S., Kock, K. H., and Bulyk, M. L. (2017). Transcription factor–dna binding : beyond binding site motifs. *Current opinion in genetics & development*, 43 :110–119.

[Ishino et al., 1987] Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., and Nakata, A. (1987). Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in escherichia coli, and identification of the gene product. *Journal of bacteriology*, 169(12) :5429–5433.

[Jansen et al., 2002] Jansen, R., Embden, J. D. v., Gaastra, W., and Schouls, L. M. (2002). Identification of genes that are associated with dna repeats in prokaryotes. *Molecular microbiology*, 43(6) :1565–1575.

[Jinek et al., 2012] Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-rna–guided dna endonuclease in adaptive bacterial immunity. *science*, 337(6096) :816–821.

[Jumper et al., 2021] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873) :583–589.

[Jurka, 2000] Jurka, J. (2000). Repbase update : a database and an electronic journal of repetitive elements. *Trends in genetics*, 16(9) :418–420.

[Jurka et al., 2005] Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*, 110(1-4) :462–467.

[Juven-Gershon and Kadonaga, 2010] Juven-Gershon, T. and Kadonaga, J. T. (2010). Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Developmental biology*, 339(2) :225–229.

[Kaaij et al., 2019] Kaaij, L. J., Mohn, F., van der Weide, R. H., de Wit, E., and Bühler, M. (2019). The chahp complex counteracts chromatin looping at ctcf sites that emerged from sine expansions in mouse. *Cell*, 178(6) :1437–1451.

[Kadonaga, 2012] Kadonaga, J. T. (2012). Perspectives on the rna polymerase ii core promoter. *Wiley Interdisciplinary Reviews : Developmental Biology*, 1(1) :40–51.

[Kaplan et al., 2009] Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., LeProust, E. M., Hughes, T. R., Lieb, J. D., Widom, J., et al. (2009). The dna-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236) :362–366.

[Karki et al., 2018] Karki, R., Place, D., Samir, P., Mavuluri, J., Sharma, B. R., Balakrishnan, A., Malireddi, R. S., Geiger, R., Zhu, Q., Neale, G., et al. (2018). Irf8 regulates transcription of naips for nlrc4 inflammasome activation. *Cell*, 173(4) :920–933.

[Kassiotis and Stoye, 2016] Kassiotis, G. and Stoye, J. P. (2016). Immune responses to endogenous retroelements : taking the bad with the good. *Nature Reviews Immunology*, 16(4) :207–219.

[Kaya-Okur et al., 2019] Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G., Ahmad, K., and Henikoff, S. (2019). Cut&tag for efficient epigenomic profiling of small samples and single cells. *Nature communications*, 10(1) :1–10.

[Kim et al., 2019] Kim, J. H., Rege, M., Valeri, J., Dunagin, M. C., Metzger, A., Titus, K. R., Gilgenast, T. G., Gong, W., Beagan, J. A., Raj, A., et al. (2019). Ladl : light-activated dynamic looping for endogenous gene expression control. *Nature methods*, 16(7) :633–639.

[Kim et al., 2018] Kim, Y. H., Marhon, S. A., Zhang, Y., Steger, D. J., Won, K.-J., and Lazar, M. A. (2018). Rev-erbα dynamically modulates chromatin looping to control circadian gene transcription. *Science*, 359(6381) :1274–1277.

[Klein et al., 2020] Klein, J. C., Agarwal, V., Inoue, F., Keith, A., Martin, B., Kircher, M., Ahituv, N., and Shendure, J. (2020). A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nature methods*, 17(11) :1083–1091.

[Klein et al., 2018] Klein, J. C., Chen, W., Gasperini, M., and Shendure, J. (2018). Identifying novel enhancer elements with crispr-based screens. *ACS chemical biology*, 13(2) :326–332.

[Konermann et al., 2015] Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., Hsu, P. D., Habib, N., Gootenberg, J. S., Nishimasu, H., et al. (2015). Genome-scale transcriptional activation by an engineered crispr-cas9 complex. *Nature*, 517(7536) :583–588.

[Kopp and Mendell, 2018] Kopp, F. and Mendell, J. T. (2018). Functional classification and experimental dissection of long noncoding rnas. *Cell*, 172(3) :393–407.

[Köster and Rahmann, 2012] Köster, J. and Rahmann, S. (2012). Snakemake ?a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19) :2520–2522.

[Kouros-Mehr et al., 2006] Kouros-Mehr, H., Slorach, E. M., Sternlicht, M. D., and Werb, Z. (2006). Gata-3 maintains the differentiation of the luminal cell fate in the mammary gland. *Cell*, 127(5) :1041–1055.

[Krijger and De Laat, 2016] Krijger, P. H. L. and De Laat, W. (2016). Regulation of disease-associated gene expression in the 3d genome. *Nature reviews Molecular cell biology*, 17(12) :771–782.

[Kuo and Schlissel, 2009] Kuo, T. C. and Schlissel, M. S. (2009). Mechanisms controlling expression of the rag locus during lymphocyte development. *Current opinion in immunology*, 21(2) :173–178.

[Kuwahara et al., 2003] Kuwahara, K., Saito, Y., Takano, M., Arai, Y., Yasuno, S., Nakagawa, Y., Takahashi, N., Adachi, Y., Takemura, G., Horie, M., et al. (2003). Nrsf regulates the fetal cardiac gene program and maintains normal cardiac structure and function. *The EMBO journal*, 22(23) :6310–6321.

[Labun et al., 2019] Labun, K., Montague, T. G., Krause, M., Torres Cleuren, Y. N., Tjeldnes, H., and Valen, E. (2019). Chopchop v3 : expanding the crispr web toolbox beyond genome editing. *Nucleic acids research*, 47(W1) :W171–W174.

[Lagrange et al., 1998] Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D., and Ebright, R. H. (1998). New core promoter element in rna polymerase ii-dependent transcription : sequence-specific dna binding by transcription factor iib. *Genes & development*, 12(1) :34–44.

[Laimins et al., 1986] Laimins, L., Holmgren-Koenig, M., and Khoury, G. (1986). Transcriptional" silencer" element in rat repetitive sequences associated with the rat insulin 1 gene locus. *Proceedings of the National Academy of Sciences*, 83(10) :3151–3155.

[Lander et al., 2001] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. .

[Lee et al., 2020] Lee, D., Shi, M., Moran, J., Wall, M., Zhang, J., Liu, J., Fitzgerald, D., Kyono, Y., Ma, L., White, K. P., et al. (2020). Starrpeaker : uniform processing and accurate identification of starr-seq active regions. *Genome biology*, 21(1) :1–24.

[Leiz et al., 2021] Leiz, J., Rutkiewicz, M., Birchmeier, C., Heinemann, U., and Schmidt-Ott, K. M. (2021). Technologies for profiling the impact of genomic variants on transcription factor binding. *Medizinische Genetik*, 33(2) :147–155.

[Leonard and Méchali, 2013] Leonard, A. C. and Méchali, M. (2013). Dna replication origins. *Cold Spring Harbor perspectives in biology*, 5(10) :a010116.

[Leporcq et al., 2020] Leporcq, C., Spill, Y., Balaramane, D., Toussaint, C., Weber, M., and Bardet, A. F. (2020). Tfmotifview : a webserver for the visualization of transcription factor motifs in genomic regions. *Nucleic acids research*, 48(W1) :W208–W217.

[Lettice et al., 2012] Lettice, L. A., Williamson, I., Wiltshire, J. H., Peluso, S., Devenney, P. S., Hill, A. E., Essafi, A., Hagman, J., Mort, R., Grimes, G., et al. (2012). Opposing functions of the ets factor family define shh spatial expression in limb buds and underlie polydactyly. *Developmental cell*, 22(2) :459–467.

[Lewis et al., 2005] Lewis, B. A., Sims III, R. J., Lane, W. S., and Reinberg, D. (2005). Functional characterization of core promoter elements : Dpe-specific transcription requires the protein kinase ck2 and the pc4 coactivator. *Molecular cell*, 18(4) :471–481.

[Li et al., 2006] Li, H., Xu, D., Toh, B.-H., and Liu, J.-P. (2006). Tgf-β and cancer : Is smad3 a repressor of htert gene ? *Cell research*, 16(2) :169–173.

[Li et al., 2014] Li, W., Xu, H., Xiao, T., Cong, L., Love, M. I., Zhang, F., Irizarry, R. A., Liu, J. S., Brown, M., and Liu, X. S. (2014). Mageck enables robust identification of essential genes from genome-scale crispr/cas9 knockout screens. *Genome biology*, 15(12) :1–12.

[Liu et al., 2009] Liu, A. M., New, D. C., Lo, R. K., and Wong, Y. H. (2009). Reporter gene assays. In *Cell-Based Assays for High-Throughput Screening*, pages 109–123. Springer.

[Lopes et al., 2021] Lopes, R., Sprouffske, K., Sheng, C., Uijttewaal, E. C., Wesdorp, A. E., Dahinden, J., Wengert, S., Diaz-Miyar, J., Yildiz, U., Bleu, M., et al. (2021). Systematic dissection of transcriptional regulatory networks by genome-scale and single-cell crispr screens. *Science advances*, 7(27) :eabf5733.

[López-Flores and Garrido-Ramos, 2012] López-Flores, I. and Garrido-Ramos, M. (2012). The repetitive dna content of eukaryotic genomes. *Repetitive DNA*, 7 :1–28.

[Lovén et al., 2013] Lovén, J., Hoke, H. A., Lin, C. Y., Lau, A., Orlando, D. A., Vakoc, C. R., Bradner, J. E., Lee, T. I., and Young, R. A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, 153(2) :320–334.

[Lu et al., 2014] Lu, T., Aron, L., Zullo, J., Pan, Y., Kim, H., Chen, Y., Yang, T.-H., Kim, H.-M., Drake, D., Liu, X. S., et al. (2014). Rest and stress resistance in ageing and alzheimer ?s disease. *Nature*, 507(7493) :448–454.

[Lupinacci et al., 2018] Lupinacci, R. M., Goloudina, A., Buhard, O., Bachet, J.-B., Maréchal, R., Demetter, P., Cros, J., Bardier-Dupas, A., Collura, A., Cervera, P., et al. (2018). Prevalence of microsatellite instability in intraductal papillary mucinous neoplasms of the pancreas. *Gastroenterology*, 154(4) :1061–1065.

[Lupo et al., 2013] Lupo, A., Cesaro, E., Montano, G., Zurlo, D., Izzo, P., and Costanzo, P. (2013). Krab-zinc finger proteins : a repressor family displaying multiple biological functions. *Current genomics*, 14(4) :268–278.

[Macfarlan et al., 2012] Macfarlan, T. S., Gifford, W. D., Driscoll, S., Lettieri, K., Rowe, H. M., Bonanomi, D., Firth, A., Singer, O., Trono, D., and Pfaff, S. L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*, 487(7405) :57–63.

[Madsen et al., 2020] Madsen, J. G., Madsen, M. S., Rauch, A., Traynor, S., Van Hauwaert, E. L., Haakonsson, A. K., Javierre, B. M., Hyldahl, M., Fraser, P., and Mandrup, S. (2020). Highly interconnected enhancer communities control lineage-determining genes in human mesenchymal stem cells. *Nature Genetics*, 52(11) :1227–1238.

[Maeshima et al., 2014] Maeshima, K., Imai, R., Tamura, S., and Nozaki, T. (2014). Chromatin as dynamic 10-nm fibers. *Chromosoma*, 123(3) :225–237.

[Manion, 2002] Manion, M. (2002). Mcgraw-hill encyclopedia of science and technology. *Reference & user services quarterly*, 42(2) :178.

[Mansour et al., 2014] Mansour, M. R., Abraham, B. J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A. D., Etchin, J., Lawton, L., Sallan, S. E., Silverman, L. B., et al. (2014). An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science*, 346(6215) :1373–1377.

[Maston et al., 2006] Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7 :29–59.

[Mathieu et al., 2014] Mathieu, E.-L., Belhocine, M., Dao, L., Puthier, D., and Spicuglia, S. (2014). Functions of lncrna in development and diseases. *Medecine Sciences : M/S*, 30(8-9) :790–796.

[Matys et al., 2006] Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). Transfac® and its module transcompel® : transcriptional gene regulation in eukaryotes. *Nucleic acids research*, 34(suppl_1) :D108–D110.

[Medina-Rivera et al., 2018] Medina-Rivera, A., Santiago-Algarra, D., Puthier, D., and Spicuglia, S. (2018). Widespread enhancer activity from core promoters. *Trends in biochemical sciences*, 43(6) :452–468.

[Melnikov et al., 2012] Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C. G., Kinney, J. B., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature biotechnology*, 30(3) :271–277.

[Meyer and Liu, 2014] Meyer, C. A. and Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics*, 15(11) :709–721.

[Miga et al., 2015] Miga, K. H., Eisenhart, C., and Kent, W. J. (2015). Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments. *Nucleic acids research*, 43(20) :e133–e133.

[Mojica et al., 2005] Mojica, F. J., Díez-Villaseñor, C., García-Martínez, J., Soria, E., et al. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of molecular evolution*, 60(2) :174–182.

[Mojica et al., 1993] Mojica, F. J., Juez, G., and Rodriguez-Valera, F. (1993). Transcription at different salinities of haloferax mediterranei sequences adjacent to partially modified psti sites. *Molecular microbiology*, 9(3) :613–621.

[Montague et al., 2014] Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M., and Valen, E. (2014). Chopchop : a crispr/cas9 and talen web tool for genome editing. *Nucleic acids research*, 42(W1) :W401–W407.

[Morgan et al., 2017] Morgan, S. L., Mariano, N. C., Bermudez, A., Arruda, N. L., Wu, F., Luo, Y., Shankar, G., Jia, L., Chen, H., Hu, J.-F., et al. (2017). Manipulation of nuclear architecture through crispr-mediated chromosomal looping. *Nature communications*, 8(1) :1–9.

[Morin et al., 2008] Morin, R. D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T. J., McDonald, H., Varhol, R., Jones, S. J., and Marra, M. A. (2008). Profiling the hela s3 transcriptome using randomly primed cdna and massively parallel short-read sequencing. *Biotechniques*, 45(1) :81–94.

[Mouri et al., 2022] Mouri, K., Dewey, H. B., Castro, R., Berenzy, D., Kales, S., and Tewhey, R. (2022). Whole genome functional characterization of re1 silencers using a modified massively parallel reporter assay. *bioRxiv*.

[Muerdter et al., 2018] Muerdter, F., Boryń, Ł. M., Woodfin, A. R., Neumayr, C., Rath, M., Zabidi, M. A., Pagani, M., Haberle, V., Kazmar, T., Catarino, R. R., et al. (2018). Resolving systematic errors in widely used enhancer activity assays in human cells. *Nature methods*, 15(2) :141–149.

[Newburger and Bulyk, 2009] Newburger, D. E. and Bulyk, M. L. (2009). Uniprobe : an online database of protein binding microarray data on protein–dna interactions. *Nucleic acids research*, 37(suppl_1) :D77–D82.

[Ng et al., 2012] Ng, S.-Y., Johnson, R., and Stanton, L. W. (2012). Human long non-coding rnas promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *The EMBO journal*, 31(3) :522–533.

[Ngan et al., 2020] Ngan, C. Y., Wong, C. H., Tjong, H., Wang, W., Goldfeder, R. L., Choi, C., He, H., Gong, L., Lin, J., Urban, B., et al. (2020). Chromatin interaction analyses elucidate the roles of prc2-bound silencers in mouse development. *Nature genetics*, 52(3) :264–272.

[Nguyen et al., 2016] Nguyen, T. A., Jones, R. D., Snavely, A. R., Pfenning, A. R., Kirchner, R., Hemberg, M., and Gray, J. M. (2016). High-throughput functional comparison of promoter and enhancer activities. *Genome research*, 26(8) :1023–1033.

[Northcott et al., 2014] Northcott, P. A., Lee, C., Zichner, T., Stütz, A. M., Erkek, S., Kawauchi, D., Shih, D. J., Hovestadt, V., Zapatka, M., Sturm, D., et al. (2014). Enhancer hijacking activates gfi1 family oncogenes in medulloblastoma. *Nature*, 511(7510) :428–434.

[Novo et al., 2018] Novo, C. L., Javierre, B.-M., Cairns, J., Segonds-Pichon, A., Wingett, S. W., Freire-Pritchett, P., Furlan-Magaril, M., Schoenfelder, S., Fraser, P., and Rugg-Gunn, P. J. (2018). Long-range enhancer interactions are prevalent in mouse embryonic stem cells and are reorganized upon pluripotent state transition. *Cell reports*, 22(10) :2615–2627.

[Nurk et al., 2021] Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., et al. (2021). The complete sequence of a human genome. *bioRxiv*.

[Nurk et al., 2022] Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. *Science*, 376(6588) :44–53.

[Ogbourne and Antalis, 1998] Ogbourne, S. and Antalis, T. M. (1998). Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochemical Journal*, 331(1) :1–14.

[O'Geen et al., 2007] O'Geen, H., Squazzo, S. L., Iyengar, S., Blahnik, K., Rinn, J. L., Chang, H. Y., Green, R., and Farnham, P. J. (2007). Genome-wide analysis of kap1 binding suggests autoregulation of krab-znfs. *PLoS genetics*, 3(6) :e89.

[Okada et al., 2010] Okada, N., Sasaki, T., Shimogori, T., and Nishihara, H. (2010). Emergence of mammals by emergency : exaptation. *Genes to Cells*, 15(8) :801–812.

[Ooi and Wood, 2007] Ooi, L. and Wood, I. C. (2007). Chromatin crosstalk in development and disease : lessons from rest. *Nature Reviews Genetics*, 8(7) :544–554.

[Ostapcuk et al., 2018] Ostapcuk, V., Mohn, F., Carl, S. H., Basters, A., Hess, D., Iesmantavicius, V., Lampersberger, L., Flemr, M., Pandey, A., Thomä, N. H., et al. (2018). Activity-dependent neuroprotective protein recruits hp1 and chd4 to control lineage-specifying genes. *Nature*, 557(7707) :739–743.

[Ou et al., 2017] Ou, H. D., Phan, S., Deerinck, T. J., Thor, A., Ellisman, M. H., and O ?shea, C. C. (2017). Chromemt : Visualizing 3d chromatin structure and compaction in interphase and mitotic cells. *Science*, 357(6349) :eaag0025.

[Oughtred et al., 2019] Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., Kolas, N., O ?Donnell, L., Leung, G., McAdam, R., et al. (2019). The biogrid interaction database : 2019 update. *Nucleic acids research*, 47(D1) :D529–D541.

[Pancaldi et al., 2016] Pancaldi, V., Carrillo-de Santa-Pau, E., Javierre, B. M., Juan, D., Fraser, P., Spivakov, M., Valencia, A., and Rico, D. (2016). Integrating epigenomic data and 3d genomic structure with a new measure of chromatin assortativity. *Genome biology*, 17(1) :1–19.

[Pang and Snyder, 2020] Pang, B. and Snyder, M. P. (2020). Systematic identification of silencers in human cells. *Nature genetics*, 52(3) :254–263.

[Park, 2009] Park, P. J. (2009). Chip–seq : advantages and challenges of a maturing technology. *Nature reviews genetics*, 10(10) :669–680.

[Patwardhan et al., 2012] Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., Lee, C., Andrie, J. M., Lee, S.-I., Cooper, G. M., et al. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nature biotechnology*, 30(3) :265–270.

[Patwardhan et al., 2009] Patwardhan, R. P., Lee, C., Litvin, O., Young, D. L., Pe'er, D., and Shendure, J. (2009). High-resolution analysis of dna regulatory elements by synthetic saturation mutagenesis. *Nature biotechnology*, 27(12) :1173–1175.

[Payer and Burns, 2019] Payer, L. M. and Burns, K. H. (2019). Transposable elements in human genetic disease. *Nature Reviews Genetics*, 20(12) :760–772.

[Payer et al., 2019] Payer, L. M., Steranka, J. P., Ardeljan, D., Walker, J., Fitzgerald, K. C., Calabresi, P. A., Cooper, T. A., and Burns, K. H. (2019). Alu insertion variants alter mrna splicing. *Nucleic acids research*, 47(1) :421–431.

[Peng et al., 2020] Peng, T., Zhai, Y., Atlasi, Y., Ter Huurne, M., Marks, H., Stunnenberg, H. G., and Megchelenbrink, W. (2020). Starr-seq identifies active, chromatin-masked, and dormant enhancers in pluripotent mouse embryonic stem cells. *Genome biology*, 21(1) :1–27.

[Petrykowska et al., 2008] Petrykowska, H. M., Vockley, C. M., and Elnitski, L. (2008). Detection and characterization of silencers and enhancer-blockers in the greater cftr locus. *Genome research*, 18(8) :1238–1246.

[Pombo and Dillon, 2015] Pombo, A. and Dillon, N. (2015). Three-dimensional genome architecture : players and mechanisms. *Nature reviews Molecular cell biology*, 16(4) :245–257.

[Privalsky, 2004] Privalsky, M. L. (2004). The role of corepressors in transcriptional regulation by nuclear hormone receptors. *Annu. Rev. Physiol.*, 66 :315–360.

[Qi et al., 2015] Qi, H., Liu, M., Emery, D. W., and Stamatoyannopoulos, G. (2015). Functional validation of a constitutive autonomous silencer element. *PloS one*, 10(4) :e0124588.

[Quail et al., 2012] Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., and Gu, Y. (2012). A tale of three next generation sequencing platforms : comparison of ion torrent, pacific biosciences and illumina miseq sequencers. *BMC genomics*, 13(1) :1–13.

[Queen and Baltimore, 1983] Queen, C. and Baltimore, D. (1983). Immunoglobulin gene transcription is activated by downstream sequence elements. *Cell*, 33(3) :741–748.

[Quinlan, 2014] Quinlan, A. R. (2014). Bedtools : the swiss-army tool for genome feature analysis. *Current protocols in bioinformatics*, 47(1) :11–12.

[Rabbani et al., 2016] Rabbani, B., Nakaoka, H., Akhondzadeh, S., Tekin, M., and Mahdieh, N. (2016). Next generation sequencing : implications in personalized medicine and pharmacogenomics. *Molecular biosystems*, 12(6) :1818–1830.

[Reece et al., 2012] Reece, J. B., Taylor, M. R., Simon, E. J., and Dickey, J. L. (2012). *Campbell biology : concepts & connections*. Benjamin Cummings San Francisco, CA.

[Rees and Liu, 2018] Rees, H. A. and Liu, D. R. (2018). Base editing : precision chemistry on the genome and transcriptome of living cells. *Nature reviews genetics*, 19(12) :770–788.

[Reyes and Huber, 2018] Reyes, A. and Huber, W. (2018). Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic acids research*, 46(2) :582–592.

[Rhee and Pugh, 2011] Rhee, H. S. and Pugh, B. F. (2011). Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, 147(6) :1408–1419.

[Rinn et al., 2007] Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J., Segal, E., et al. (2007). Functional demarcation of active and silent chromatin domains in human hox loci by noncoding rnas. *cell*, 129(7) :1311–1323.

[Roopra et al., 2004] Roopra, A., Qazi, R., Schoenike, B., Daley, T. J., and Morrison, J. F. (2004). Localized domains of g9a-mediated histone methylation are required for silencing of neuronal genes. *Molecular cell*, 14(6) :727–738.

[Ross-Innes et al., 2012] Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., Brown, G. D., Gojis, O., Ellis, I. O., Green, A. R., et al. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 481(7381) :389–393.

[Rossi et al., 2015] Rossi, A., Kontarakis, Z., Gerri, C., Nolte, H., Hölper, S., Krüger, M., and Stainier, D. Y. (2015). Genetic compensation induced by deleterious mutations but not gene knockdowns. *Nature*, 524(7564) :230–233.

[Rothbart and Strahl, 2014] Rothbart, S. B. and Strahl, B. D. (2014). Interpreting the language of histone and dna modifications. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1839(8) :627–643.

[Rovira et al., 2021] Rovira, M., Atla, G., Maestro, M. A., Grau, V., García-Hurtado, J., Maqueda, M., Mosquera, J. L., Yamada, Y., Kerr-Conte, J., Pattou, F., et al. (2021). Rest is a major negative regulator of endocrine differentiation during pancreas organogenesis. *Genes & development*, 35(17-18) :1229–1242.

[Sacoto et al., 2020] Sacoto, M. J. G., Tchasovnikarova, I. A., Torti, E., Forster, C., Andrew, E. H., Anselm, I., Baranano, K. W., Briere, L. C., Cohen, J. S., Craigen, W. J., et al. (2020). De novo variants in the atpase module of morc2 cause a neurodevelopmental disorder with growth retardation and variable craniofacial dysmorphism. *The American Journal of Human Genetics*, 107(2) :352–363.

[Sandelin et al., 2004] Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B. (2004). Jaspar : an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(suppl_1) :D91–D94.

[Sandve et al., 2010] Sandve, G. K., Gundersen, S., Rydbeck, H., Glad, I. K., Holden, L., Holden, M., Liestøl, K., Clancy, T., Ferkingstad, E., Johansen, M., et al. (2010). The genomic hyperbrowser : inferential genomics at the sequence level. *Genome biology*, 11(12) :1–12.

[Sanna et al., 2008] Sanna, C. R., Li, W.-H., and Zhang, L. (2008). Overlapping genes in the human and mouse genomes. *BMC genomics*, 9(1) :1–11.

[Santiago-Algarra et al., 2017] Santiago-Algarra, D., Dao, L. T., Pradel, L., España, A., and Spicuglia, S. (2017). Recent advances in high-throughput approaches to dissect enhancer function. *F1000Research*, 6.

[Santiago-Algarra et al., 2021] Santiago-Algarra, D., Souaid, C., Singh, H., Dao, L., Sadouni, N., Hussain, S., Medina-Rivera, A., Ramirez-Navarro, L., Castro-Mondragon, J. A., Charbonnier, G., et al. (2021). Epromoters function as a hub to recruit key transcription factors required for the inflammatory response. *Nature communications*, 12(1) :1–18.

[Sato, 2007] Sato, N. (2007). Origin and evolution of plastids : genomic view on the unification and diversity of plastids. In *The structure and function of plastids*, pages 75–102. Springer.

[Schmidt et al., 2012] Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Gonçalves, Â., Kutter, C., Brown, G. D., Marshall, A., Flicek, P., and Odom, D. T. (2012). Waves of retrotransposon expansion remodel genome organization and ctcf binding in multiple mammalian lineages. *Cell*, 148(1-2) :335–348.

[Schmidt et al., 2015] Schmidt, S. F., Larsen, B. D., Loft, A., Nielsen, R., Madsen, J. G. S., and Mandrup, S. (2015). Acute tnf-induced repression of cell identity genes is mediated by nfκb-directed redistribution of cofactors from super-enhancers. *Genome research*, 25(9) :1281–1294.

[Schoenherr and Anderson, 1995] Schoenherr, C. J. and Anderson, D. J. (1995). The neuron-restrictive silencer factor (nrsf) : a coordinate repressor of multiple neuron-specific genes. *Science*, 267(5202) :1360–1363.

[Scholl et al., 1996] Scholl, T., Stevens, M. B., Mahanta, S., and Strominger, J. L. (1996). A zinc finger protein that represses transcription of the human mhc class ii gene, dpa. *The Journal of Immunology*, 156(4) :1448–1457.

[Seczynska et al., 2022] Seczynska, M., Bloor, S., Cuesta, S. M., and Lehner, P. J. (2022). Genome surveillance by hush-mediated silencing of intronless mobile elements. *Nature*, 601(7893) :440–445.

[Segert et al., 2021] Segert, J. A., Gisselbrecht, S. S., and Bulyk, M. L. (2021). Transcriptional silencers : Driving gene expression with the brakes on. *Trends in Genetics*, 37(6) :514–527.

[Sertil et al., 2003] Sertil, O., Kapoor, R., Cohen, B. D., Abramova, N., and Lowry, C. V. (2003). Synergistic repression of anaerobic genes by mot3 and rox1 in saccharomyces cerevisiae. *Nucleic acids research*, 31(20) :5831–5837.

[Shalem et al., 2014] Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G., et al. (2014). Genome-scale crispr-cas9 knockout screening in human cells. *Science*, 343(6166) :84–87.

[Shin et al., 2009] Shin, H., Liu, T., Manrai, A. K., and Liu, X. S. (2009). Ceas : cis-regulatory element annotation system. *Bioinformatics*, 25(19) :2605–2606.

[Siersbæk et al., 2014] Siersbæk, R., Rabiee, A., Nielsen, R., Sidoli, S., Traynor, S., Loft, A., Poulsen, L. L. C., Rogowska-Wrzesinska, A., Jensen, O. N., and Mandrup, S. (2014). Transcription factor cooperativity in early adipogenic hotspots and super-enhancers. *Cell reports*, 7(5) :1443–1455.

[Simovski et al., 2018] Simovski, B., Kanduri, C., Gundersen, S., Titov, D., Domanska, D., Bock, C., Bossini-Castillo, L., Chikina, M., Favorov, A., Layer, R. M., et al. (2018). Coloc-stats : a unified web interface to perform colocalization analysis of genomic features. *Nucleic acids research*, 46(W1) :W186–W193.

[Skene and Henikoff, 2017] Skene, P. J. and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of dna binding sites. *elife*, 6 :e21856.

[Smith et al., 2006] Smith, A. D., Sumazin, P., Xuan, Z., and Zhang, M. Q. (2006). Dna motifs in human and mouse proximal promoters predict tissue-specific expression. *Proceedings of the National Academy of Sciences*, 103(16) :6275–6280.

[Smith and Meissner, 2013] Smith, Z. D. and Meissner, A. (2013). Dna methylation : roles in mammalian development. *Nature Reviews Genetics*, 14(3) :204–220.

[Sobocińska et al., 2021] Sobocińska, J., Molenda, S., Machnik, M., and Oleksiewicz, U. (2021). Krab-zfp transcriptional regulators acting as oncogenes and tumor suppressors : An overview. *International journal of molecular sciences*, 22(4) :2212.

[Sparber et al., 2019] Sparber, P., Filatova, A., Khantemirova, M., and Skoblov, M. (2019). The role of long non-coding rnas in the pathogenesis of hereditary diseases. *BMC Medical Genomics*, 12(2) :63–78.

[Stamos et al., 2021] Stamos, D. B., Clubb, L. M., Mitra, A., Chopp, L. B., Nie, J., Ding, Y., Das, A., Venkataganesh, H., Lee, J., El-Khoury, D., et al. (2021). The histone demethylase lsd1 regulates multiple repressive gene programs during t cell development. *Journal of Experimental Medicine*, 218(12) :e20202012.

[Stedman et al., 2008] Stedman, W., Kang, H., Lin, S., Kissil, J. L., Bartolomei, M. S., and Lieberman, P. M. (2008). Cohesins localize with ctcf at the kshv latency control region and at cellular c-myc and h19/igf2 insulators. *The EMBO journal*, 27(4) :654–666.

[Subramanian et al., 2020] Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, 14 :1177932219899051.

[Sundaram and Wysocka, 2020] Sundaram, V. and Wysocka, J. (2020). Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. *Philosophical Transactions of the Royal Society B*, 375(1795) :20190347.

[Suzuki et al., 2015] Suzuki, A., Wakaguri, H., Yamashita, R., Kawano, S., Tsuchihara, K., Sugano, S., Suzuki, Y., and Nakai, K. (2015). Dbtss as an integrative platform for transcriptome, epigenome and genome sequence variation data. *Nucleic acids research*, 43(D1) :D87–D91.

[Takimoto et al., 2010] Takimoto, T., Wakabayashi, Y., Sekiya, T., Inoue, N., Morita, R., Ichiyama, K., Takahashi, R., Asakawa, M., Muto, G., Mori, T., et al. (2010). Smad2 and smad3 are redundantly essential for the tgf-β–mediated regulation of regulatory t plasticity and th1 development. *The Journal of immunology*, 185(2) :842–855.

[Tanenbaum et al., 2014] Tanenbaum, M. E., Gilbert, L. A., Qi, L. S., Weissman, J. S., and Vale, R. D. (2014). A protein-tagging system for signal amplification in gene expression and fluorescence imaging. *Cell*, 159(3) :635–646.

[Tang et al., 2016] Tang, X., Kim, J., Zhou, L., Wengert, E., Zhang, L., Wu, Z., Carromeu, C., Muotri, A. R., Marchetto, M. C., Gage, F. H., et al. (2016). Kcc2 rescues functional deficits in human neurons derived from patients with rett syndrome. *Proceedings of the National Academy of Sciences*, 113(3) :751–756.

[Taniuchi et al., 2004] Taniuchi, I., Ellmeier, W., and Littman, D. R. (2004). The cd4 ? cd8 lineage choice : New insights into epigenetic regulation during t cell development. *Advances in immunology*, 83 :55–89.

[Taub et al., 1982] Taub, R., Kirsch, I., Morton, C., Lenoir, G., Swan, D., Tronick, S., Aaronson, S., and Leder, P. (1982). Translocation of the c-myc gene into the immunoglobulin heavy chain locus in human burkitt lymphoma and murine plasmacytoma cells. *Proceedings of the National Academy of Sciences*, 79(24) :7837–7841.

[Tchasovnikarova et al., 2017] Tchasovnikarova, I. A., Timms, R. T., Douse, C. H., Roberts, R. C., Dougan, G., Kingston, R. E., Modis, Y., and Lehner, P. J. (2017). Hyperactivation of hush complex function by charcot–marie–tooth disease mutation in morc2. *Nature genetics*, 49(7) :1035–1044.

[Tchasovnikarova et al., 2015] Tchasovnikarova, I. A., Timms, R. T., Matheson, N. J., Wals, K., Antrobus, R., Göttgens, B., Dougan, G., Dawson, M. A., and Lehner, P. J. (2015). Epigenetic silencing by the hush complex mediates position-effect variegation in human cells. *Science*, 348(6242) :1481–1485.

[Thiollier et al., 2012] Thiollier, C., Lopez, C. K., Gerby, B., Ignacimouttou, C., Poglio, S., Duffourd, Y., Guégan, J., Rivera-Munoz, P., Bluteau, O., Mabialah, V., et al. (2012). Characterization of novel genomic alterations and therapeutic approaches using acute megakaryoblastic leukemia xenograft models. *Journal of Experimental Medicine*, 209(11) :2017–2031.

[Thomas-Chollier et al., 2008] Thomas-Chollier, M., Sand, O., Turatsinze, J.-V., Janky, R., Defrance, M., Vervisch, E., Brohee, S., and van Helden, J. (2008). Rsat : regulatory sequence analysis tools. *Nucleic acids research*, 36(suppl_2) :W119–W127.

[Touat et al., 2020] Touat, M., Li, Y. Y., Boynton, A. N., Spurr, L. F., Iorgulescu, J. B., Bohrson, C. L., Cortes-Ciriano, I., Birzu, C., Geduldig, J. E., Pelton, K., et al. (2020). Mechanisms and therapeutic implications of hypermutation in gliomas. *Nature*, 580(7804) :517–523.

[Trizzino et al., 2017] Trizzino, M., Park, Y., Holsbach-Beltrame, M., Aracena, K., Mika, K., Caliskan, M., Perry, G. H., Lynch, V. J., and Brown, C. D. (2017). Transposable elements are the primary source of novelty in primate gene regulation. *Genome research*, 27(10) :1623–1633.

[Tunbak et al., 2020] Tunbak, H., Enriquez-Gasca, R., Tie, C. H., Gould, P. A., Mlcochova, P., Gupta, R. K., Fernandes, L., Holt, J., van der Veen, A. G., Giampazolias, E., et al. (2020). The hush complex is a gatekeeper of type i interferon through epigenetic regulation of line-1s. *Nature communications*, 11(1) :1–15.

[Tunyasuvunakool et al., 2021] Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873) :590–596.

[van Ouwerkerk et al., 2020] van Ouwerkerk, A. F., Hall, A. W., Kadow, Z. A., Lazarevic, S., Reyat, J. S., Tucker, N. R., Nadadur, R. D., Bosada, F. M., Bianchi, V., Ellinor, P. T., et al. (2020). Epigenetic and transcriptional networks underlying atrial fibrillation. *Circulation research*, 127(1) :34–50.

[Vanhille et al., 2015] Vanhille, L., Griffon, A., Maqbool, M. A., Zacarias-Cabeza, J., Dao, L., Fernandez, N., Ballester, B., Andrau, J. C., and Spicuglia, S. (2015). High-throughput and quantitative assessment of enhancer activity in mammals by capstarr-seq. *Nature communications*, 6(1) :1–10.

[Vaquerizas et al., 2009] Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors : function, expression and evolution. *Nature Reviews Genetics*, 10(4) :252–263.

[Villiers et al., 1982] Villiers, J. d., Olson, L., Tyndall, C., and Schaffner, W. (1982). Transcriptional ?enhancers ? from sv40 and polyoma virus show a cell type preference. *Nucleic acids research*, 10(24) :7965–7976.

[Visel et al., 2007] Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L. A. (2007). Vista enhancer browser ?a database of tissue-specific human enhancers. *Nucleic acids research*, 35(suppl_1) :D88–D92.

[Vu et al., 2018] Vu, T. N., Deng, W., Trac, Q. T., Calza, S., Hwang, W., and Pawitan, Y. (2018). A fast detection of fusion genes from paired-end rna-seq data. *BMC genomics*, 19(1) :1–13.

[Wang et al., 2015] Wang, J., Vicente-García, C., Seruggia, D., Moltó, E., Fernandez-Miñán, A., Neto, A., Lee, E., Gómez-Skarmeta, J. L., Montoliu, L., Lunyak, V. V., et al. (2015). Mir retrotransposon sequences provide insulators to the human genome. *Proceedings of the National Academy of Sciences*, 112(32) :E4428–E4437.

[Wang et al., 2008] Wang, L., Charroux, B., Kerridge, S., and Tsai, C.-C. (2008). Atrophin recruits hdac1/2 and g9a to modify histone h3k9 and to determine cell fates. *EMBO reports*, 9(6) :555–562.

[Wang et al., 2011] Wang, Q., Liu, J.-q., Chen, Z., Zheng, K.-w., Chen, C.-y., Hao, Y.-h., and Tan, Z. (2011). G-quadruplex formation at the 3 ? end of telomere dna inhibits its extension by telomerase, polymerase and unwinding by helicase. *Nucleic acids research*, 39(14) :6229–6237.

[Wang et al., 2014] Wang, T., Wei, J. J., Sabatini, D. M., and Lander, E. S. (2014). Genetic screens in human cells using the crispr-cas9 system. *Science*, 343(6166) :80–84.

[Wang et al., 2018] Wang, X., He, L., Goggin, S. M., Saadat, A., Wang, L., Sinnott-Armstrong, N., Claussnitzer, M., and Kellis, M. (2018). High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nature communications*, 9(1) :1–15.

[Weingarten-Gabbay et al., 2019] Weingarten-Gabbay, S., Nir, R., Lubliner, S., Sharon, E., Kalma, Y., Weinberger, A., and Segal, E. (2019). Systematic interrogation of human promoters. *Genome research*, 29(2) :171–183.

[Westbrook et al., 2005] Westbrook, T. F., Martin, E. S., Schlabach, M. R., Leng, Y., Liang, A. C., Feng, B., Zhao, J. J., Roberts, T. M., Mandel, G., Hannon, G. J., et al. (2005). A genetic screen for candidate tumor suppressors identifies rest. *Cell*, 121(6) :837–848.

[White, 2015] White, M. A. (2015). Understanding how cis-regulatory function is encoded in dna sequence using massively parallel reporter assays and designed sequences. *Genomics*, 106(3) :165–170.

[Whyte et al., 2013] Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2) :307–319.

[Wicker et al., 2007] Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12) :973–982.

[Wilkinson et al., 2016] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1) :1–9.

[Wingender et al., 1996] Wingender, E., Dietze, P., Karas, H., and Knüppel, R. (1996). Transfac : a database on transcription factors and their dna binding sites. *Nucleic acids research*, 24(1) :238–241.

[Xuan Lin et al., 2019] Xuan Lin, Q. X., Sian, S., An, O., Thieffry, D., Jha, S., and Benoukraf, T. (2019). Methmotif : an integrative cell specific database of transcription factor binding motifs coupled with dna methylation profiles. *Nucleic acids research*, 47(D1) :D145–D154.

[Yang et al., 2004] Yang, J., Mani, S. A., Donaher, J. L., Ramaswamy, S., Itzykson, R. A., Come, C., Savagner, P., Gitelman, I., Richardson, A., and Weinberg, R. A. (2004). Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *cell*, 117(7) :927–939.

[Yannoutsos et al., 2004] Yannoutsos, N., Barreto, V., Misulovin, Z., Gazumyan, A., Yu, W., Rajewsky, N., Peixoto, B. R., Eisenreich, T., and Nussenzweig, M. C. (2004). A cis element in the recombination activating gene locus regulates gene expression by counteracting a distant silencer. *Nature immunology*, 5(4) :443–450.

[Yoon et al., 2012] Yoon, J.-H., Abdelmohsen, K., Srikantan, S., Yang, X., Martindale, J. L., De, S., Huarte, M., Zhan, M., Becker, K. G., and Gorospe, M. (2012). Lincrna-p21 suppresses target mrna translation. *Molecular cell*, 47(4) :648–655.

[Yu et al., 2020] Yu, Z., Feng, J., Wang, W., Deng, Z., Zhang, Y., Xiao, L., Wang, Z., Liu, C., Liu, Q., Chen, S., et al. (2020). The egfr-znf263 signaling axis silences six3 in glioblastoma epigenetically. *Oncogene*, 39(15) :3163–3178.

[Zabidi et al., 2015] Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2015). Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, 518(7540) :556–559.

[Zeng et al., 2021] Zeng, W., Chen, S., Cui, X., Chen, X., Gao, Z., and Jiang, R. (2021). Silencerdb : a comprehensive database of silencers. *Nucleic acids research*, 49(D1) :D221–D228.

[Zhang et al., 2016] Zhang, X., Choi, P. S., Francis, J. M., Imielinski, M., Watanabe, H., Cherniack, A. D., and Meyerson, M. (2016). Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nature genetics*, 48(2) :176–182.

[Zhao et al., 2006] Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Singh Sandhu, K., Singh, U., et al. (2006). Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions. *Nature genetics*, 38(11) :1341–1347.

[Zweidler-McKay et al., 1996] Zweidler-McKay, P. A., Grimes, H. L., Flubacher, M. M., and Tsichlis, P. N. (1996). Gfi-1 encodes a nuclear zinc finger protein that binds dna and functions as a transcriptional repressor. *Molecular and cellular biology*, 16(8) :4024–4034.