Décryptage de dynamiques épigénomiques au cours de la thymopoïèse et de la spermatogénèse en appliquant une méthodologie de recherche reproductible à des données de séquençage à haut débit

Guillaume Charbonnier

AIX-MARSEILLE UNIVERSITÉ
ECOLE DOCTORALE 62
TAGC/Inserm U1090

Thèse présentée pour obtenir le grade universitaire de docteur

Discipline : Sciences de la vie et de la santé
Spécialité : Génomique et Bioinformatique

Guillaume CHARBONNIER

Décryptage de dynamiques épigénomiques au cours de la thymopoïèse et de la spermatogénèse en appliquant une méthodologie de recherche reproductible à des données de séquençage à haut débit

Soutenue le 4 Octobre 2019 devant le jury composé de :

Dr Carl HERRMANN (GL)	DKFZ (Heidelberg)	Rapporteur
Dr Marco Antonio MENDOZA (CR2)	ISSB (Paris)	Rapporteur
Dr Sophie ROUSSEAUX (DR2)	IAB (Grenoble)	Examinateur
Dr Catherine NGUYEN (DR2)	TAGC (Marseille)	Examinateur
Dr Denis PUTHIER (MCU)	TAGC (Marseille)	Co-directeur de thèse
Dr Salvatore SPICUGLIA (DR2)	TAGC (Marseille)	Directeur de thèse

Numéro national de thèse/suffixe local : 2019AIXM0285/074ED62

Résumé

Cette dernière décennie, le développement de nombreuses méthodes expérimentales basées sur les technologies de séquençage à haut débit ont élargi les possibilités d'exploration du fonctionnement du génome des organismes. Le niveau d'expression de l'ensemble des gènes, l'accessibilité et les modifications locales de la chromatine, ainsi que les sites de fixations de protéines sur l'ADN sont des informations accessibles par des approches largement démocratisées. Les données produites par ces approches sont caractérisées par leur taille importante, leur complexité et leur réemployabilité pour des applications scientifiques au delà de celle pour laquelle elles ont été initialement générées. Ces caractéristiques ont favorisé la création de projets internationaux fructueux pour la génération et le partage de ces données à la communauté scientifique.

Les traitements informatiques qu'il est possible de leur appliquer sont d'une grande diversité en termes de principes, outils, algorithmes, paramètres et tests statististiques. Des choix différents peuvent parfois mener à des résultats divergents débattus par des scientifiques. Il est alors primordial de disposer d'une méthodologie pour le traitement de ces données qui permette d'assurer une reproductibilité complète des résultats tout en accordant une flexibilité de développement permettant l'application de différents traitements en parallèle.

Une implémentation d'une telle méthodologie est présentée dans cette thèse, complémentée de trois outils utiles pour le traitement de certains types de données issues du séquençage. Cette implémentation a été réalisée grâce à, mais aussi pour des projets d'études de mécanismes épigénétiques, principalement centrés autour de la thymopoïèse humaine et de la spermatogénèse murine. Pour autant, les principes à suivre dans la méthodologie sont pensés pour être les plus généraux possibles et peuvent être appliqués pour produire tous types d'analyses basées sur un enchaînement d'outils en ligne de commande.

Les analyses de données ChIP-Seq, ATAC-Seq, WGBS et RNA-Seq du premier projet ont permis premièrement de replacer de manière cohérente et ordonnée les paysages épigénétiques des sous-populations thymiques par rapport à ceux de toutes les populations issues de lignées hématopoïétiques. Deuxièmement, ces analyses ont montré qu'une majorité des éléments régulateurs distaux sont déméthylées de manière constitutive au cours de la différentiation des lymphocytes T, et ce indépendamment de leur état d'activation. L'un d'eux, l'amplificateur du gène du TCRA, Eα, est dans un état ouvert et déméthylé dès les étapes les plus précoces de la thymopoïèse alors qu'il ne s'active que tardivement. Les données transcriptomiques ont révélé que le niveau d'expression des facteurs de transcription HOXA5-9 diminuent au moment où Eα s'active. Finalement, ces analyses ont permis de guider la découverte par d'autres approches biologiques du rôle d'HOXA9 dans le blocage du mécanisme de réarrangements du TCRA.

Concernant l'étude de la spermatogénèse murine, l'analyse de données ChIP-Seq à différents stades de différentation a permis de caractériser l'acétylation et la butyrylation des lysines 5 et 8 de l'histone 4 comme étant quatre modifications post-traductionnelles fortement corrélées et marquant les gènes hautement transcrits. L'analyse de données MNase-Seq a permis de lever le voile sur la dynamique d'éviction, de remplacement par des protamines, et de repositionnement des nucléosomes au cours de la spermiogénèse. D'autres analyses MNase-Seq ont permis de caractériser l'impact de l'incorporation d'un variant d'histone, H2AL2, et de l'expression d'un facteur de transcription spécifique de la spermiogénèse, Nut, sur la structure de la chromatine.

Abstract

The development over the last decade of numerous experimental methods based on high throughput sequencing technologies has broadened the possibilities for exploring the functioning of the genome of organisms. The level of expression of all genes, accessibility and local modifications of chromatin, as well as protein binding sites on DNA are information accessible by widely democratized approaches. The data produced by these approaches are characterized by their large size, complexity and reusability for scientific applications beyond the one they were initially generated. These characteristics have favored the creation of successful international projects for the generation and sharing of these data to the scientific community.

The computer treatments that can be applied to them are of a great diversity in terms of principles, tools, algorithms, parameters and statistic tests. Different choices can sometimes lead to divergent results debated by scientists. It is then essential to have a methodology for the processing of these data which allows to ensure a complete reproducibility of the results while granting flexibility of development allowing the application of different treatments in parallel.

An implementation of such a methodology is presented in this thesis, complemented by three useful tools for the treatment of certain types of data from sequencing. This implementation was realized thanks to, but also for projects of studies of epigenetic mechanisms, mainly centered around the human thymopoiesis and the murine spermatogenesis. However, the principles to be followed in the methodology are designed to be as general as possible and can be applied to produce all types of analysis based on a sequence of command line tools.

ChIP-Seq, ATAC-Seq, WGBS and RNA-Seq data analyzes from the first project allowed the epigenetic landscapes of the thymic subpopulations to be consistently and orderly compared with those of all populations from hematopoietic lineages. Secondly, these analyzes have shown that a majority of the distal regulatory elements are constitutively demethylated during the differentiation of T lymphocytes, regardless of their activation state. One of them, the TCRA gene enhancer, Eα, is in an open and demethylated state as early as the earliest stages of thymopoiesis while it only activates late. Transcriptomic data revealed that the expression level of transcription factors HOXA5-9 decreases as Eα activates. Finally, these analyzes guided the discovery by other biological approaches of the role of HOXA9 in blocking the TCRA rearrangement mechanism.

Regarding the study of murine spermatogenesis, the analysis of ChIP-Seq data at different stages of differentiation made it possible to characterize the acetylation and butyrylation of lysines 5 and 8 of histone 4 as four strongly correlated post-translational modifications marking the highly transcribed genes. MNase-Seq data analysis has shed light on the dynamics of eviction, replacement and repositioning of nucleosomes during spermiogenesis. Other MNase-Seq analyzes have characterized the impact of the incorporation of a histone variant, H2AL2, and the expression of a spermiogenesis-specific transcription factor, Nut, on the structure of chromatin.

Keywords: bioinformatics, reproducible research, epigenomics, thymopoiesis, spermatogenesis, .*-Seq, snakemake, python, R, conda, linux

Mots clés: bioinformatique, recherche reproductible, épigénomique, thymopoïèse, spermatogénèse, .*-seq, snakemake, python, R, conda, linux

Remerciements

Mes premiers remerciements vont aux Docteurs Carl Herrmann, Marco Antonio Mendoza et Sophie Rousseaux pour avoir accepté de porter un regard critique sur mon travail de thèse.

Je tiens à remercier Denis Puthier, pour avoir été l'enseignant puis le mentor qui m'a le plus inspiré durant mes études, et pour avoir été le premier à m'accorder sa confiance. Merci pour tes conseils et ta relecture avisée. Je remercie également Salvatore Spicuglia, Vahid Asnafi et Saadi Khochbin qui m'ont poussé à terminer cette thèse. Un grand merci à vous tous pour m'avoir permis de participer à l'exploration de thématiques si passionnantes, vos enseignements et vos disponibilités. J'espère que nous pourrons continuer à collaborer ensemble pour les années à venir.

Je remercie les directeurs successifs du TAGC, Catherine Nguyen pour m'avoir accueilli au sein du laboratoire et m'avoir permis de m'inscrire en thèse, et Pascal Rihet, pour avoir enduré la fantaisie de mes aspirations professionnelles. Merci également à Myriam Ramadour pour avoir géré le casse-tête administratif de ces pérégrinations.

Je tiens à témoigner toute ma reconnaissance aux personnes avec qui j'ai pu collaborer et interagir sur les différents projets pour leur sympathie et leur expertise. Je pense notamment aux expérimentalistes, Sophie Barral, Thierry Buchou, Agata Cieslak et Charlotte Smith, ainsi qu'aux ingénieurs de la plateforme de séquençage TGML, Hortense Vachon, Lisa Bargier et Béatrice Loriod, sans oublier les bioinformaticiens Sophie Rousseaux, Quentin Ferré, Florent Chuffart et Fabrice Lopez.

Je tiens à remercier tous les collègues du TAGC pour les bons moments que nous avons passé ensemble, avec une pensée pour la Justice League de l'épigénétique, Iris, Nori, Alex, José-David, Saadat, David, Himanshu, Charbel et Lydie ainsi que les chevaliers Jedi Florian et Jeanne, l'Imperialdramon Firas, et le prince du Qatar Michel.

Je garde une place particulière pour les amis qui m'ont soutenu dans les moments délicats, Nicolas, Céline et Lucie. Merci d'avoir été là. C'est également le moment de remercier un collègue devenu ami qui aurait pu être cité dans les paragraphes précédents, Eleusis Haradwaith. Sans toi ces premières années de thèse auraient été particulièrement ennuyeuses. Enfin, je tiens à remercier ma famille et ma compagne qui ont eu la patience de me supporter jusqu'au bout de cette aventure.

Table des matières

Table des figures

Liste des tableaux

Table des articles

Avant-propos

En 1985, des biologistes ont évoqué l'idée d'un projet titanesque : séquencer le génome humain dans son intégralité. On estimait alors qu'il faudrait trente ans pour lire l'ensemble des informations contenues dans les 24 chromosomes, soit 3,2 milliards de nucléotides (COOK-DEEGAN 1991). Malgré les techniques de séquençage de l'ADN encore balbutiantes, un consortium international s'est lancé dans l'aventure du Projet Génome Humain en 1990. En 2001, il livrait la séquence quasi-complète du génome humain pour un coût total d'environ 3 milliards de dollars. Aujourd'hui, grâce au séquençage à haut débit, il est possible de séquencer le génome d'une personne en quelques heures, pour quelques centaines d'euros et l'objectif des 100$ est évoqué comme réaliste pour les années à venir (HERPER 2019).

En parallèle de la réduction du coût de séquençage se sont développées de nombreuses techniques dérivées repoussant à chaque fois un peu plus les limites possibles de l'exploration du fonctionnement du Vivant (ANNEXE A). Un socle commun de caractéristiques peut se dégager de toutes ces approches. Elles sont :

— massives, *i.e.* chaque échantillon de données brutes contient entre quelques dizaines et centaines de giga-octets, et autant en données processées,

— complexes, *i.e.* il n'y a pas une unique façon d'analyser le contenu de chaque échantillon et

— réemployables au delà de l'application scientifique pour lesquelles elles ont été initiallement générées.

De part ces caractéristiques, ces techniques ont profondément transformé l'approche de la Biologie au point de faire émerger un domaine d'expertise à part entière : la Bioinformatique. Les expériences ne servent plus seulement à produire des résultats scientifiques publiables, mais constituent autant de ressources pouvant être croisées, consolidées ou réfutées par de futures analyses. Sur ce dernier point, la Bioinformatique n'échappe pas à la crise de la reproductibilité qui secoue régulièrement les milieux scientifiques. Néanmoins, elle représente par essence un cadre dans lequel l'objectif de reproductibilité parfaite semble atteignable, à défaut d'être atteint.

C'est avec cet objectif méthodologique que j'ai entrepris le travail de recherche retranscrit dans ce manuscrit. Ainsi, une importance particulière est accordée à la méthodologie qui est développée en détail après une mise en contexte et une présentation des résultats biologiques obtenus.

1 Contexte général

1.1 Organisation du génome eucaryote

Le génome représente l'ensemble du matériel génétique d'une espèce 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 et BONHOMME 1986), ainsi que du génome plastidial des végétaux (SATO 2007). Le génome nucléaire est 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 cette section.

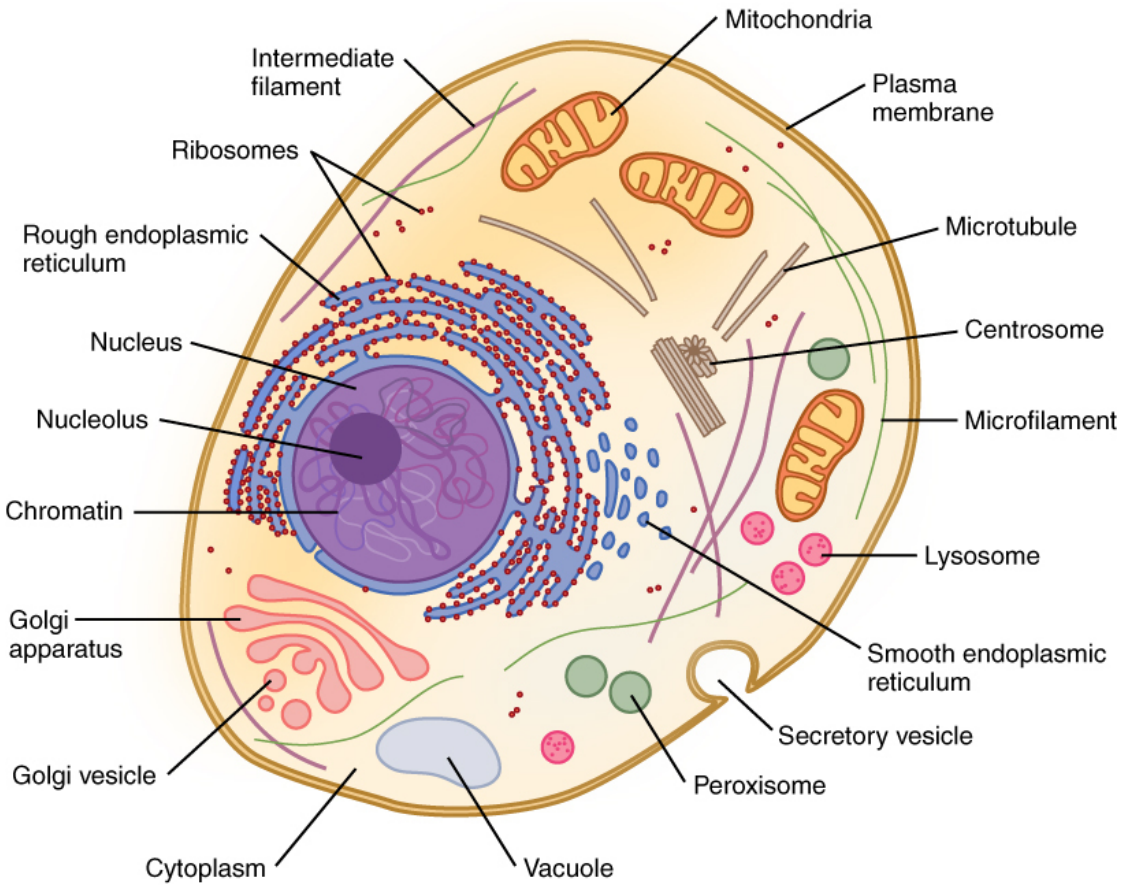
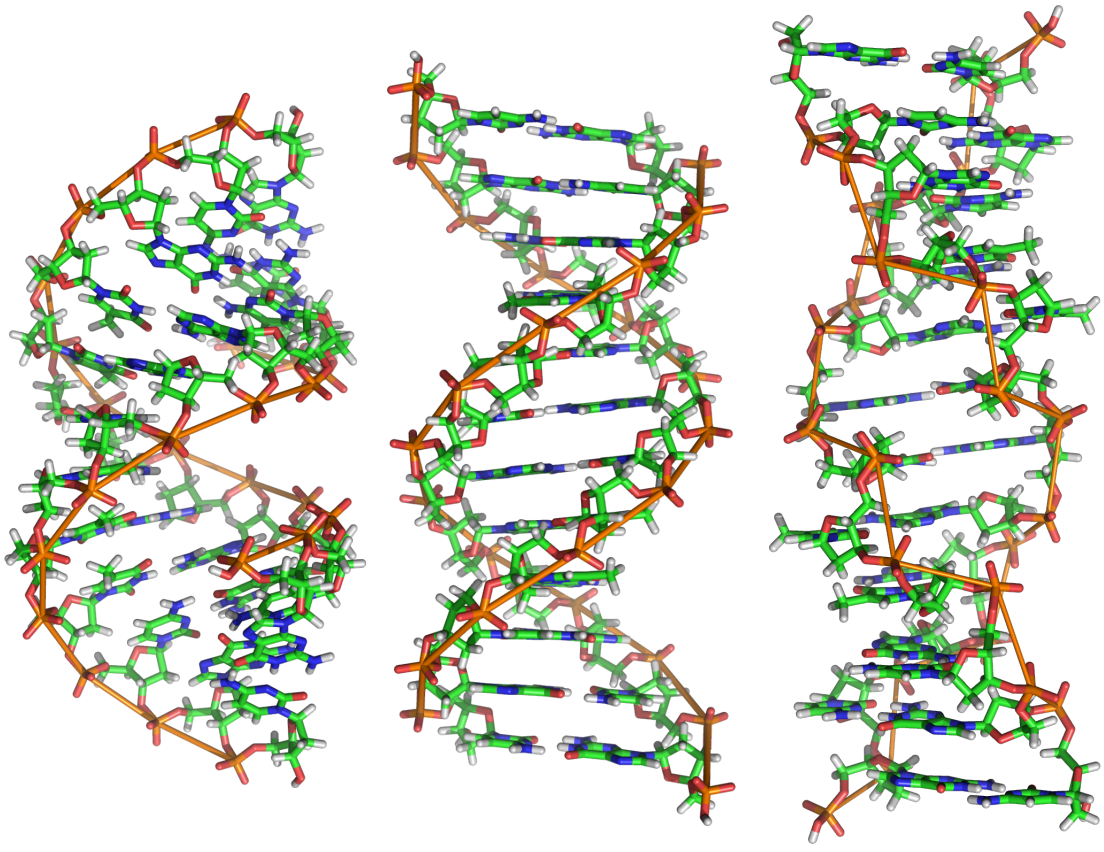


FIGURE 1.1 – Organisation d'une cellule eucaryote classique. © Richard Wheeler (Zephyris) / English Wikipedia / CC BY-SA 3.0

1.1.1 Structure de la chromatine

1.1.1.1 Structures de l'ADN

L'ADN a été pour la pemière fois décrit 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 (β-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), et ponctuellement en structures particulières : triplex, G-quadruplex, épingle à cheveux, ramification et jonction de Holliday (FIGURE 1.3).



(a) ADN A (b) ADN B (c) ADN Z

FIGURE 1.2 – Géométries en double hélice communes de l'ADN *in vivo*. © Richard Wheeler (Zephyris) / English Wikipedia / CC BY-SA 3.0

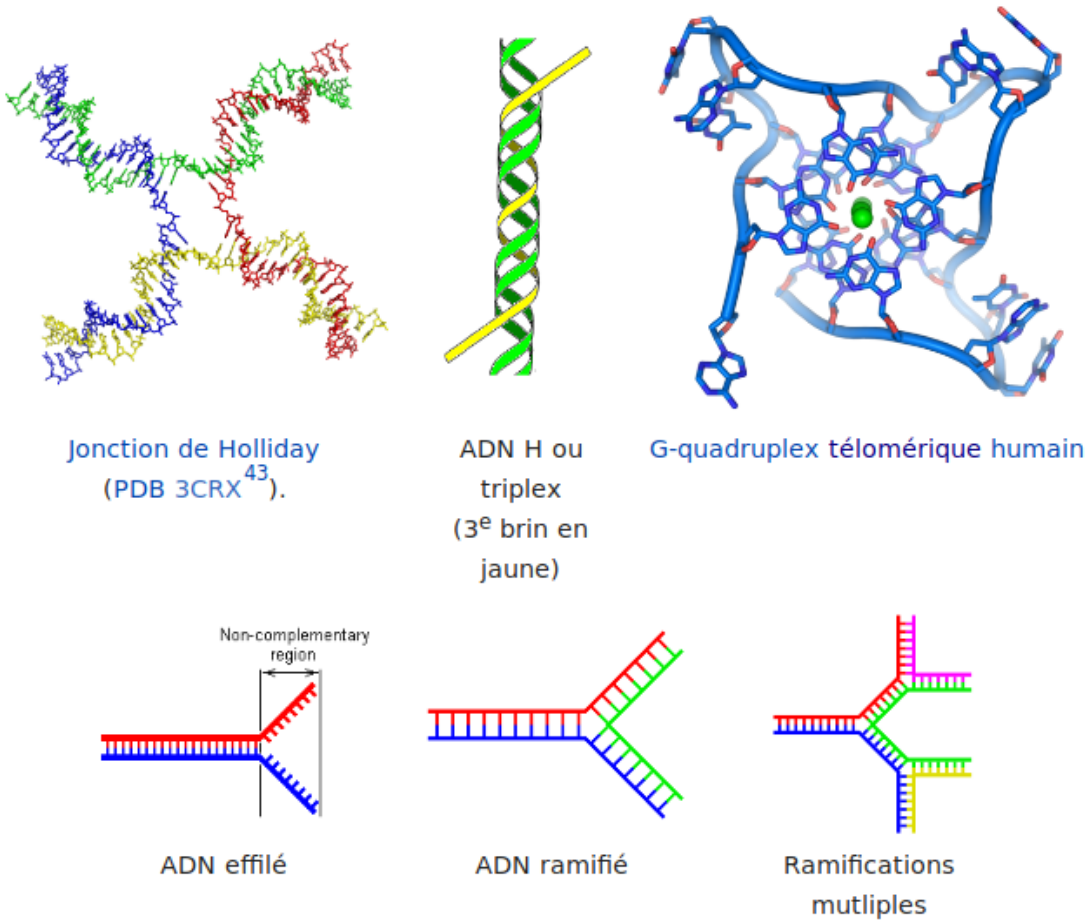


FIGURE 1.3 – Géométries ponctuelles de l'ADN *in vivo*.

1.1.1.2 Structure du nucléosome

Les histones constituent la famille principale de protéines associées à l'ADN. On distingue cinq classes d'histones. Les quatres premières H2A, H2B, H3 et H4 présentent la caractéristique de pouvoir s'associer en quadruple hétérodimères pour former un octamère autour duquel l'ADN s'enroule sur une longueur de 147 paires de bases (FIGURE 1.4). Cette association ADN-histones représente l'unité de base de la chromatine, le nucléosome.

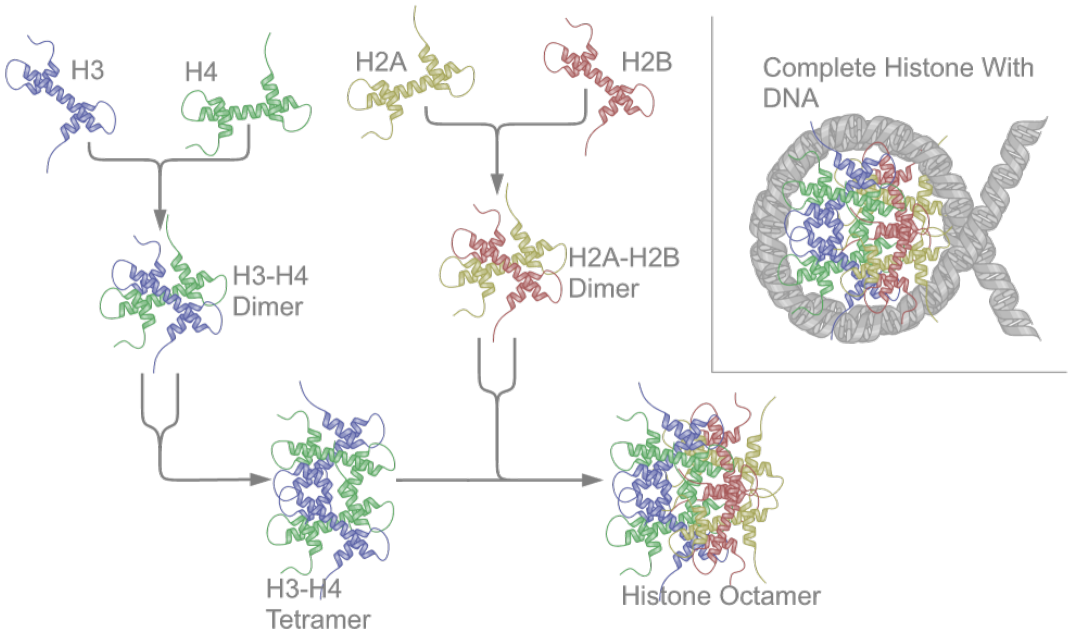


FIGURE 1.4 – Diagramme d'assemblage d'un nucléosome. © Richard Wheeler (Zephyris) / English Wikipedia / CC BY-SA 3.0

1.1.1.3 Organisation spatiale de la chromatine

Les nucléosomes successifs s'organisent en une structure de type "collier de perles" de 11 nm de diamètre. La dernière classe d'histone, H1, peut venir se fixer sur chaque nucléosome 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. 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, IMAI, TAMURA et al. 2014).

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

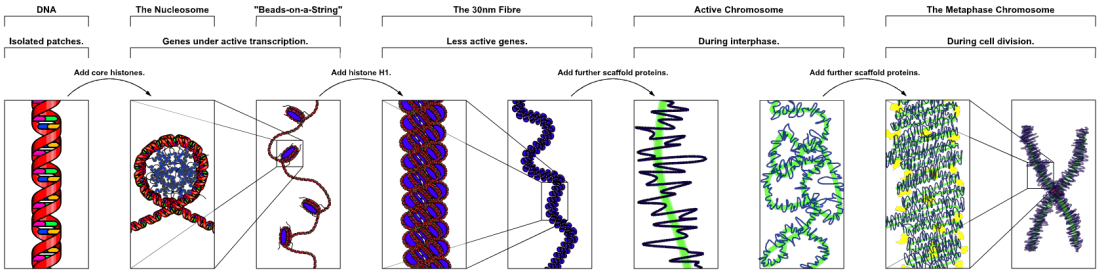


FIGURE 1.5 – Niveaux hiérarchiques des structures de compaction de l'ADN. © Richard Wheeler (Zephyris) / English Wikipedia / CC BY-SA 3.0

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 et 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 700 nm de diamètre, 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, PHAN, DEERINCK et al. 2017).

Le niveau de compaction le plus extrême de la chromatine est observé au cours de la spermiogénèse des métazoaires. La chromatine du spermatide est alors remodelée en une structure quasi-crystalline où la plupart des nucléosomes sont remplacés par des protamines. Une étude des mécanismes sous-jacents à ces remaniements chromatiniens est proposée en SECTION 2.2.

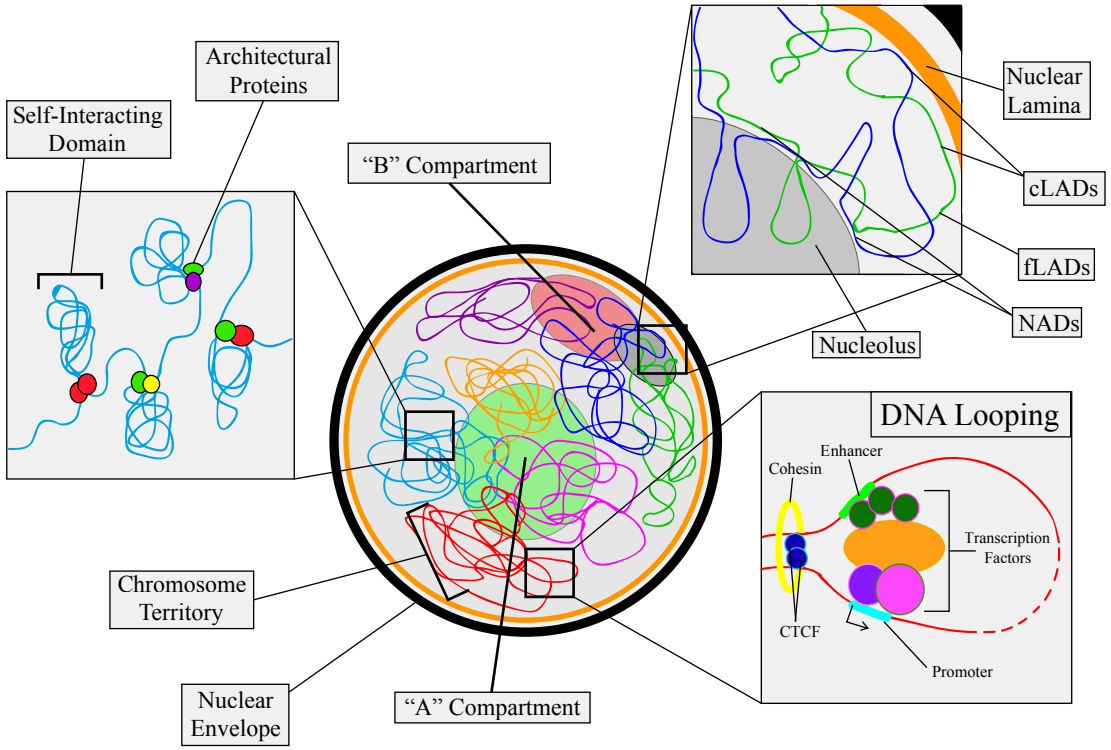


FIGURE 1.6 – Niveaux d'organisation spatiale de la chromatine. © Evin Wieser / CC BY-SA 4.0

1.1.2 Classification du génome en éléments fonctionnels

L'objectif de cette section est de présenter succintement les différents éléments fonctionnels du génome.

1.1.2.1 Gènes

Un gène est une séquence de nucléotides codant pour une molécule possédant une fonction. Cette molécule correspond classiquement à une protéine préalablement traduite par un 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 méchanismes (transcription, épissage, maturation, traduction) permettant d'aboutir à une protéine fonctionnelle. Enfin, les gènes permettant de produire des immunoglobulines (IG) ou des récepteurs de cellules T (TR) présentent un mécanisme de recombinaison de séquence au niveau de l'ADN suffisamment caractéristique pour les classifier à part (SECTION 1.2.1.3). Ces différentes familles sont classifiées en biotypes.

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 1.8a). 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 ou *template* et synthétisent donc des brins complémentaires anti-parallèles dans le sens 5'-3' considéré alors comme le sens en-avant ou *forward* (FIGURE 1.8b). 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

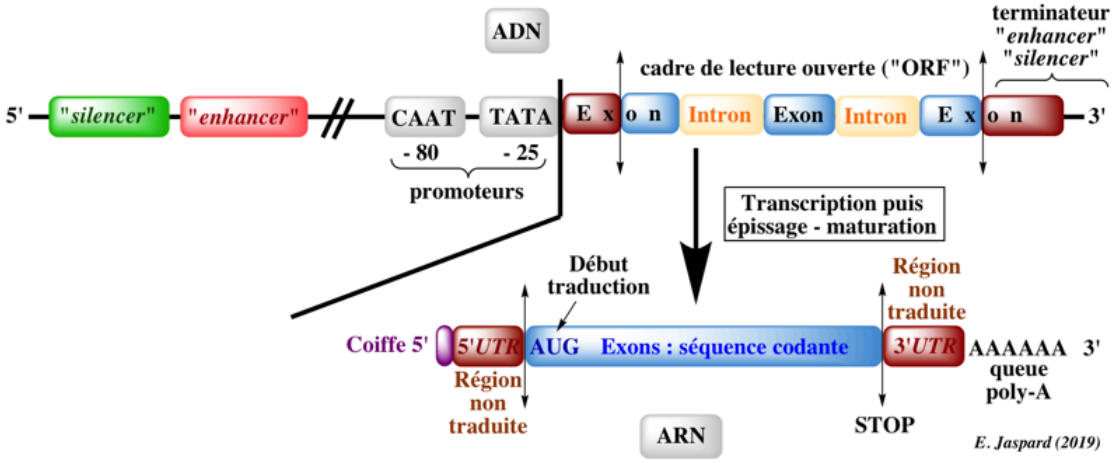


FIGURE 1.7 – Modèle de gène eucaryote classique. Extrait de E. Jaspard 2019.

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.

Les gènes sont répartis de façon relativement uniforme sur les deux brins d'ADN de chaque chromosome (KELLIS, WOLD, SNYDER et al. 2014), et environ 10% d'entres eux sont chevauchants avec un autre gène (SANNA, LI et ZHANG 2008). Afin de pouvoir définir sans ambiguité 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 (SECTION 1.1.2.3).

Un gène peut être associé à plusieurs transcrits pouvant différer par leur site d'initiation de la transcription (*Transcription Start Site* : TSS), les exons retenus après épissage et leur site de terminaison de la transcription (*Transcription End Site* : TES). Différentes isoformes d'un transcrit sont en général associées à différents types cellulaires d'un organisme. L'existence de TSS et TES 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 et HUBER 2018).

1.1.2.2 Éléments cis-régulateurs

Les éléments cis-régulateurs sont des régions voisines aux gènes ayant une fonction dans la régulation de leur transcription. Les éléments cis-régulateurs tiennent leur préfixe "cis" du fait qu'ils sont situés sur la même molécule d'ADN que le gène régulé, par opposition aux éléments trans-régulateurs correspondant aux gènes codant pour des protéines venant se fixer sur les éléments cis-régulateurs. Ces protéines sont appelées des facteurs de transcription. On distingue quatre classes d'éléments cis-régulateurs : promoteurs, amplificateurs, inactivateurs et isolateurs.

Promoteurs Un promoteur est une région d'ADN en amont d'un gène à proximité du TSS sur laquelle se fixe initialement l'ARN polymérase avant de démarrer la synthèse d'ARN. Les promoteurs eucaryotes sont classiquement associés à plusieurs éléments (BUTLER 2002) :

— le *TFIIB Recognition Element* (BRE) en 5' du site d'initiation de la transcription,

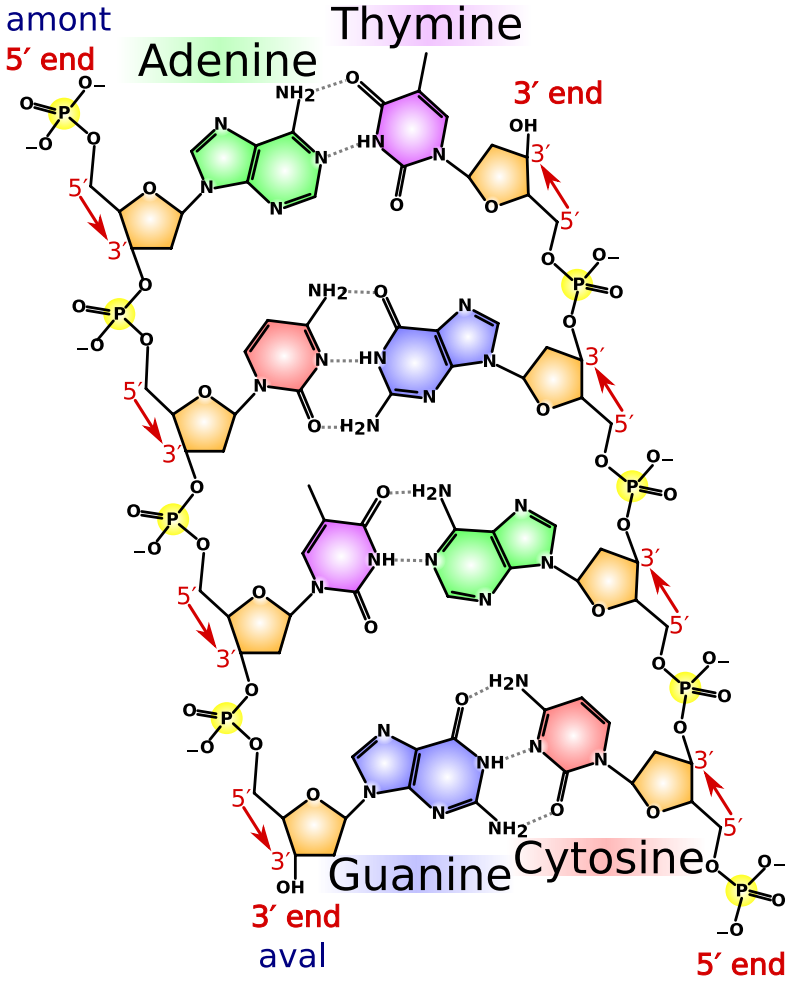
— la boîte TATA, en 5' du site d'initiation de la transcription et en 3' de BRE, entre les positions -23 et -33,

— la boîte CAAT, en 5' de la boîte GC, entre les positions -80 et -120,

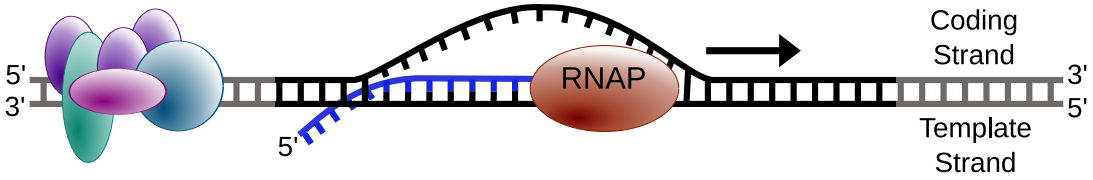
— la séquence *Inr*, ou *initiator*, correspondant au site d'initiation de la transcription,

— le *Motif Ten Element* (MTE), en 3' du site d'initiation de la transcription et

— le *Downstream Promoter Element* (DPE) en 3' du MTE.



(a) Conventions de directionnalité dans une séquence nucléotidique



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

FIGURE 1.8 – Terminologie de navigation dans une séquence nucléotidique. Extrait du domaine public.

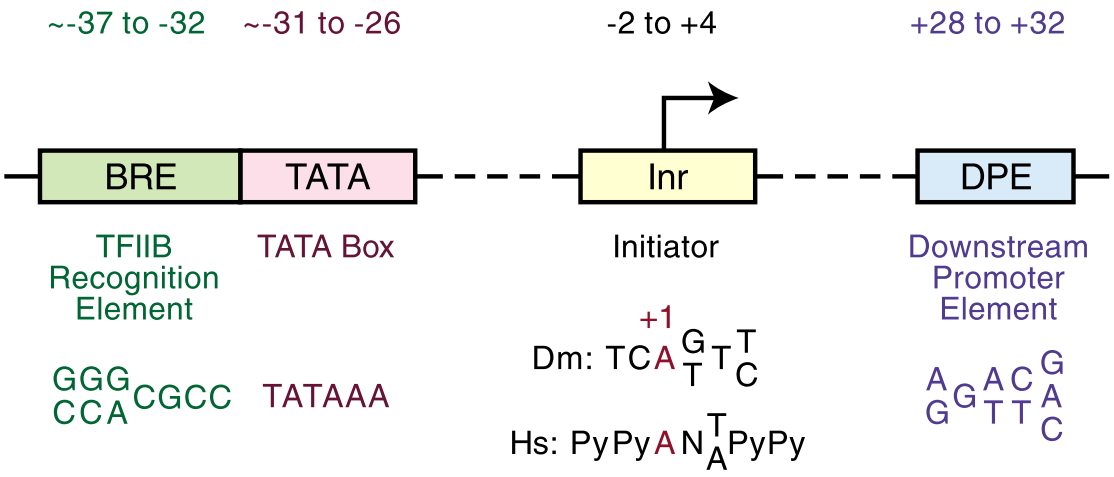


FIGURE 1.9 – Éléments classiquement associés dans un promoteur eucaryote. Dm : *Drosophila melanogaster* ; Hs : *Homo sapiens* ; Py : pyrimidine. Extrait de BUTLER 2002.

Ces éléments ne sont cependant pas observés chez tous les eucaryotes et sont seulement détectés pour une minorité des promoteurs d'un organisme. La séquence de ces éléments peut également être dégénérée relativement aux motifs consensus, *i.e.* certains nucléotides des motifs sont remplacés par d'autres (FIGURE 1.9). Des séquences promotrices plus ou moins complètes et dégénérées vont influencer sur l'affinité du promoteur pour les différentes protéines nécessaires à l'initiation de la transcription. En conséquence, certains promoteurs qualifiés de forts produiront, en l'absence d'autres éléments régulateurs, un niveau élevé de transcrits alors que des promoteurs qualifiés de faibles produiront un niveau moindre de transcrits. L'existence de ces différentes formes de promoteurs peut s'expliquer par l'utilité pour un organisme de disposer de promoteurs faibles ou forts en fonction des gènes associés. Un promoteur fort peut ainsi coder pour un gène dont le produit est utile en permanence dans la cellule, alors qu'un promoteur faible peut ne produire de transcrits qu'en présence d'un environnement en autres éléments régulateurs spécifiques d'un type cellulaire ou d'un contexte nécessitant la transcription de ce gène.

Un promoteur est finalement plus facilement définissable par des approches expérimentales pouvant détecter les sites de fixations de l'ARN polymérase (SUN, WU, WICKRAMASINGHE et al. 2011). La plupart sont aussi caractérisés par une absence de nucléosomes corrélée à un taux élevés de dinucléotides CpG (FENOUIL, CAUCHY, KOCH et al. 2012). Si les différents éléments constitutifs des promoteurs peuvent s'étendre sur une région entre 100 et 1000 paires de bases (bp) de long et chevauchant le TSS, il est commun en bioinformatique de définir plus simplement un promoteur comme étant l'ensemble de la région de 1, 2 voir 3 kilo-paires de bases (kb) en amont d'un TSS (SHIN, LIU, MANRAI et al. 2009). Dans ce cas, la partie du promoteur en aval du TSS est associée à la région 5'-UTR.

Amplificateurs Un amplificateur, ou *enhancer*, est une courte région d'ADN, généralement entre 50 et 1500 bp, pouvant être fixée par des facteurs de transcription qualifiés d'activateurs pour augmenter le niveau de transcription d'un gène à proximité plus ou moins immédiate (2 à 2000 kb en amont ou aval). Le modèle fonctionnel le plus probable expliquant le mode d'action des amplificateurs repose sur la création d'une boucle de l'ADN permettant l'apparition d'une proximité spatiale entre le ou les amplificateurs fixés par les facteurs de transcription et le promoteur du gène à activer (FIGURE 1.6 ; KOLOVOS, KNOCH, GROSVELD et al. 2012).

Inactivateurs Un inactivateur, ou *silencer*, présente des caractéristiques analogues à l'amplificateur, à la différence près qu'il recrute des facteurs de transcription répresseurs de la transcription. Les inactivateurs, comme les amplificateurs, sont des éléments régulateurs présentant une tendance générale à être actifs spécifiquement dans certains types cellulaires. Des régions amplificatrices sont également connues pour commuter en inactivateur en fonction du type et de l'environnement cellulaire (BESSIS, CHAMPTIAUX, CHATELIN et al. 1997). Certains promoteurs peuvent aussi avoir une fonction d'amplificateurs pour d'autres gènes voisins (ANNEXE B ; DAO, GALINDO-ALBARRÁN, CASTRO-MONDRAGON et al. 2017). Le même principe est suspecté pour la tenue d'une fonction d'inactivateur.

Isolateurs Un isolateur est un élément cis-régulateur qui joue le rôle barrière séparatrice entre sa région en amont et en aval. Un amplificateur ou inactivateur actif n'ont pas d'influence sur des gènes desquels ils sont séparés par un isolateur. Certains isolateurs séparent également des régions condensées et décondensées de la chromatine (FIGURE 1.10 ; GASZNER et 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 en SECTION 1.1.1.3.

1.1.2.3 Éléments à fonction non génique

Environ 80% du génome est couvert par des gènes d'après la définition fournie en SECTION 1.1.2.1. Sans s'imiscer dans le tumultueux débat de l'aspect fonctionnel, ou pas d'un grand nombre de ces gènes (KELLIS, WOLD, SNYDER et al. 2014), on peut citer brièvement des éléments du génome qui peuvent posséder conceptuellement une fonction indépendante d'une quelconque transcription menant à un produit fonctionnel.

Origines de réplication Les origines de réplications 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 et MÉCHALI 2013). Leur taille est comprise entre 100 et 1000 bp et la majorité de ces régions contiennent

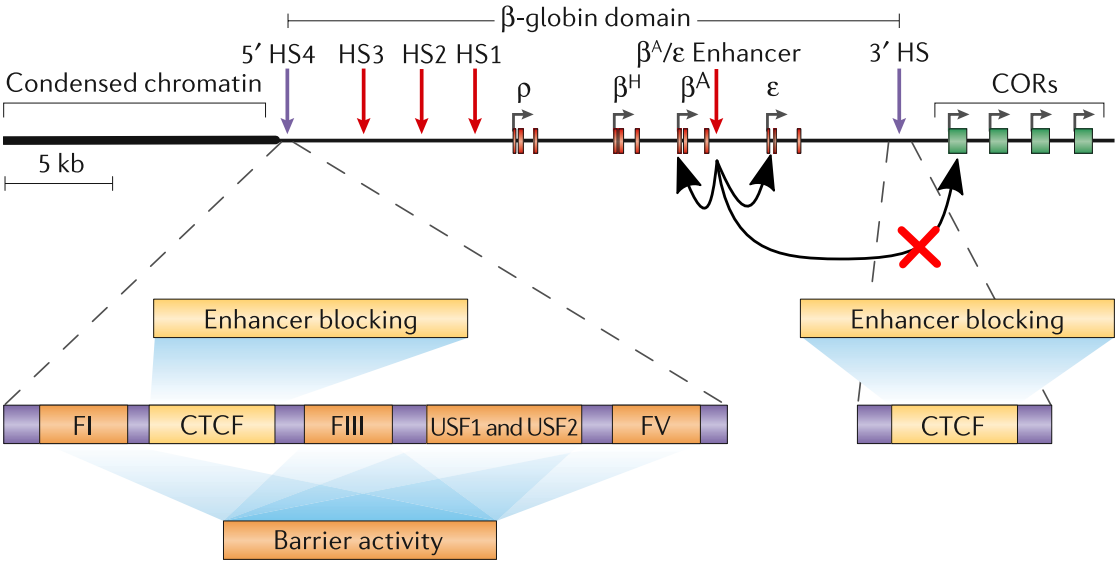


FIGURE 1.10 – Exemple de locus délimité par les deux types d'isolateurs : locus de la β-globine du poulet. Chez le poulet (*Gallus gallus*), les isolateurs 5'HS4 et 3'HS, indiqués par des flèches violettes définissent les limites d'une domaine chromatinien qui contient le cluster de gènes de la β-globine. Les amplificateurs 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 et FELSENFELD 2006

chez les eucaryotes des séquences répétées riches en guanine capable de former des G-quadruplex (CAYROU, BALLESTER, PEIFFER et al. 2015).

Centromères 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 mégabases) 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 et FACHINETTI 2018).

Télomères Les extrémités de chaque chromosome, ou télomères, sont constituées de séquences répétées riches en guanine propices à la formation de G-quadruplex (WANG, LIU, CHEN et al. 2011). En l'absence de mécanismes compensatoires, chaque chromosome perd approximativement 50 nucléotides par cycle de réplication à cause d'imperfections dans le mécanisme de réplication par l'ADN polymérase (GOMEZ, ARMANDO, FARINA et al. 2012). Les télomères assurent le rôle de tampon vis-à-vis des gènes périphériques et leur structure particulière permet l'action de la télomérase, une enzyme capable d'inverser le processus.

Transposons Pour finir par le type d'éléments aux fonctions les plus controversées, les transposons des séquences ADN capables de se déplacer et se dupliquer de manière autonome dans un génome, par un mécanisme appelé transposition. Malgré l'existence d'exemples ponctuels du rôle de séquences dérivées d'éléments transposables dans certaines fonctions géniques (JONES et GELLERT 2004), les éléments semblent pour la plupart inutiles à la cellule vivante, et parfois même défavorables. 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, BURNS, GEHRING et al. 2018).

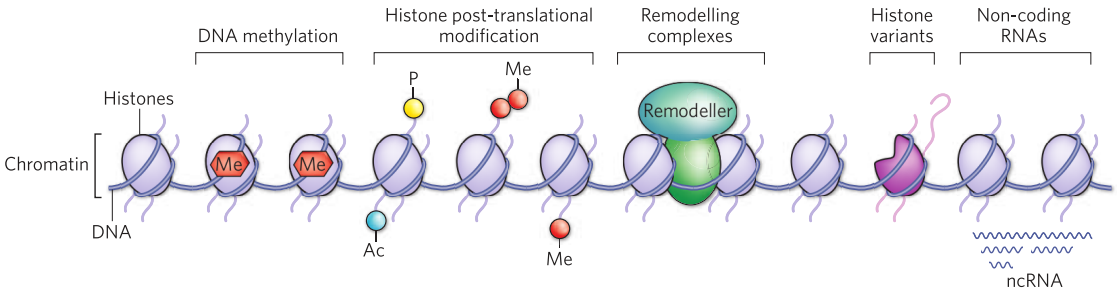


FIGURE 1.11 – Mécanismes épigénétiques impliqués dans la modification de la chromatine. Extrait de DULAC 2010

.

1.1.3 Vue d'ensemble de l'épigénétique

Alors que la plupart des cellules d'un organisme possèdent un génome identique, comment expliquer l'existence de 411 types cellulaires chez l'homme ? (VICKARYOUS et HALL 2006) A l'échelle d'un organisme, comment expliquer les différences phénotypiques que l'on observe par exemple chez les abeilles entre une reine et une ouvrière possédant pourtant le même génome ? (HERB, WOLSCHIN, HANSEN et al. 2012) De même, comment expliquer chez certains reptiles que la détermination du sexe soit définie par la température d'incubation de l'œuf ? (YATSU, MIYAGAWA, KOHNO et al. 2016) Dès 1942, Conrad Hal Waddington postule l'existence d'un ensemble complexe de processus permettant d'expliquer comment le génotype peut être modulé pour produire des phénotypes différents. Il introduit ainsi la notion d'épigénotype qui donnera naissance à une science dédiée à son étude : l'épigénétique (WADDINGTON 2012).

La plupart des recherches actuelles dans ce domaine convergent vers l'idée qu'une part importante des phénomènes épigénétiques puisse être expliquée par des modifications de la chromatine au travers de modifications de l'ADN et des histones (FIGURE 1.11). Ces modifications permettent principalement de moduler l'accessibilité et la reconnaissance des gènes pour les complexes impliqués dans la transcription et apportent par conséquent un niveau de régulation de l'expression des gènes. Une autre part des phénomènes épigénétiques est tenue par plusieurs familles d'ARN non-codant pour des protéines (PESCHANSKY et WAHLESTEDT 2014 ; FRÍAS-LASSERRE et VILLAGRA 2017). Ces phénomènes n'étant pas étudiés dans cette thèse, ils ne sont pas présentés ci-dessous.

1.1.3.1 Modifications de l'ADN

Méthylation des cytosines La méthylation des cytosines (5mC), et plus précisément celles suivies d'une guanine, a concentré 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 et

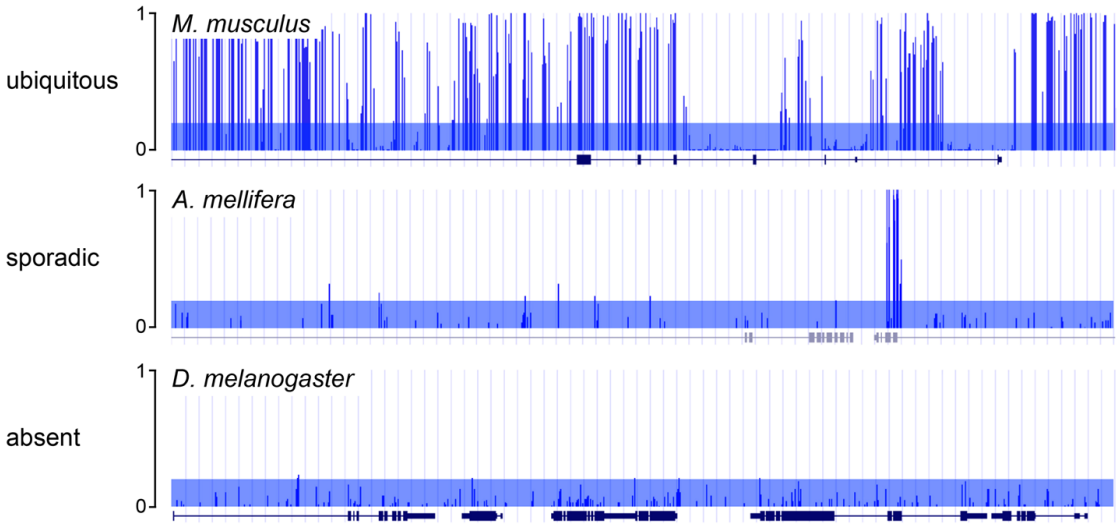


FIGURE 1.12 – Catégories principales de méthylomes chez les eucaryotes. Les méthylations ubiquitaires, sporadiques et absentes sont illustrées par des vues de navigateur de génome sur des résultats de *whole-genome bisulfite sequencing* (SECTION 1.3.2.5). Les trois vues chez la souris, l'abeille et la drosophile correspondent à des régions aléatoires de 40 kb. Les barres bleues correspondant à un ratio de méthylation de 0.2 indiquent le bruit artefactuel d'une expérience de bisulfite-seq. Extrait de BREILING et LYKO 2015.

MEISSNER 2013 ; BAUBEC et SCHÜBELER 2014) ou induire la fixation de protéines 5mC-spécifiques pour recruter des complexes de co-répresseurs (KLOSE et BIRD 2006). Le rapport de CpG méthylés et non méthylés varie grandement entre espèces. La méthylation est omniprésente chez l'humain et la souris, sporadique chez l'abeille, et absente chez la drosophile ce qui suggère des mécanismes de régulation différents entre espèces (FIGURE 1.12 ; BREILING et LYKO 2015). Une famille d'enzymes, les DNA méthyltransférases (DNMT) sont responsables du processus de méthylation des cytosines.

Déméthylation des cytosines Plus récemment, la découverte de l'activité dioxygénase de la famille des protéines *Ten eleven translocation* (TET) a attiré l'attention sur les formes oxydées successives de la méthylcytosine : l'hydroxyméthylcytosine (hmC), la formylcytosines (fC) et la carboxylcytosine (caC) (ARAVIND, BANDUKWALA, AGARWAL et al. 2009 ; ITO, DALESSIO, TARANOVA et al. 2010). L'hydroxyméthylcytosine se retrouve en tant que marqueur dans les amplificateurs et les corps de gènes actifs (LU, LIU, JIANG et al. 2014 ; HON, SONG, DU et al. 2014). On trouve 30 000 méthylcytosines, 1 000 hydroxyméthylytosines, 20 formylcytosines et 3 carboxylcytosines par million de cytosines dans des cellules embryonnaires de souris. Sachant que les formyl- et carboxylcytosines sont reconnues et excisées par un mécanisme de réparation de l'ADN médié par des Thymine-ADN-Glycosylases (TDG), les formes hydroxylées des cytosines pour-

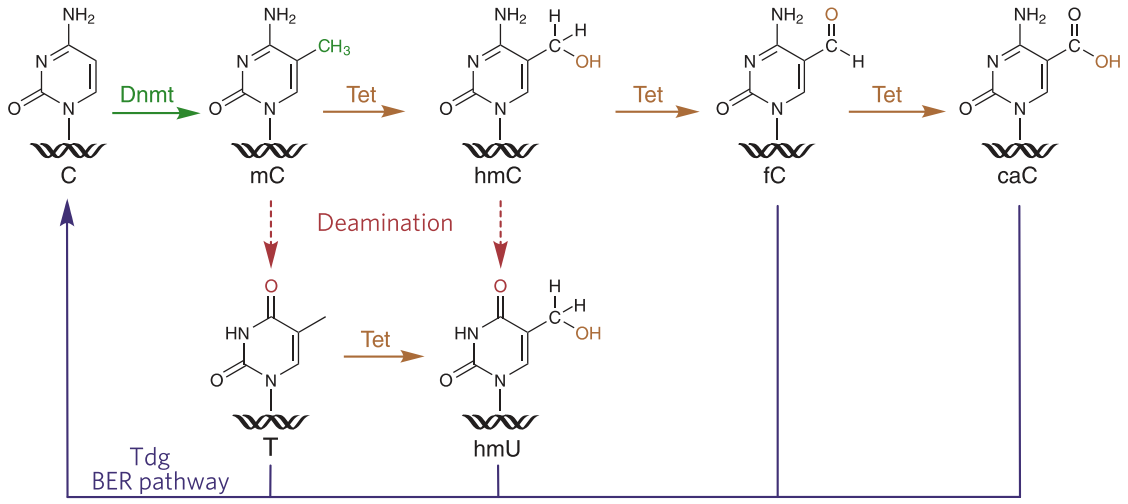


FIGURE 1.13 – Chemins de méthylation/déméthylation des cytosines. Les déaminations sont des phénomènes spontanés moins fréquents que les réactions catalysées par Tet. Revu et adapté de PFAFFENEDER, SPADA, WAGNER et al. 2014.

raient jouer le rôle d'intermédiaires de déméthylation. Les méthylcytosines et les hydroxyméthylcytosine produisent en se déaminant spontanément respectivement des thymines et des hydroxyuraciles produisant ainsi dans une moindre mesure d'autres intermédiaires de déméthylation (FIGURE 1.13 ; ITO, SHEN, DAI et al. 2011).

Autres modifications Les thymines peuvent se retrouver hydroxylées en hydroxyuraciles par les mêmes enzymes TET. Les hydroxyuraciles ainsi formées semblent jouer un rôle dans le recrutement des complexes de réparation de l'ADN entraînant la restauration conjointe des cytosines et des thymines (PFAFFENEDER, SPADA, WAGNER et al. 2014). Pour finir avec les modifications de l'ADN, on peut noter que chez plusieurs espèces procaryotes, ainsi que chez la drosophile et le ver *Caenorhabditis elegans*, la N6-méthylation des adénines semble remplacer le mécanisme de méthylation des cytosines (BREILING et LYKO 2015).

1.1.3.2 Variants d'histones

Les modifications d'histones peuvent se classifier en deux types. Le premier type de modifications d'histones est le remplacement des histones canoniques présentées en SECTION 1.1.1.2 par des variants différant de quelques acides aminés à de longs fragments (FIGURE 1.15). L'assemblage des histones canoniques sur l'ADN est un processus intimement lié à l'étape de réplication de l'ADN (phase S) au cours du cycle cellulaire (MARZLUFF et DURONIO 2002). Après la mitose, le niveau d'expression des gènes canoniques diminue et un remplacement des histones peut s'opérer sur des régions spécifiques du génome ce qui induit une modification des propriétés de la chromatine impactant des mécanismes tels

que la transcription, la recombinaison et la réparation de l'ADN. L'assemblage et l'incorporation des histones sont guidés par des protéines de la famille des chaperonnes d'histones. Leur réarrangement par glissement le long de l'ADN, leur éviction ou leur remplacement sont guidés par des remodeleurs ATP-dépendants (FIGURE 1.14 ; TYLER 2002 ; CLAPIER et CAIRNS 2009 ; DAS, TYLER et CHURCHILL 2010 ; BURGESS et ZHANG 2013).

Variants communs de H3 Chez les eucaryotes, les centromères sont définis fonctionnellement par la présence d'histones H3 centromériques (cenH3) plutôt que par leur séquence nucléotidique (MIZUGUCHI, XIAO, WISNIEWSKI et al. 2007). De la même façon, les gènes actifs et l'hétérochromatine télomérique sont marqués par le variant H3.3 (FILIPESCU, SZENKER et ALMOUZNI 2016). Ce dernier ne semble pas avoir de fonction différente de l'histone canonique mais possède la capacité de remplacer les nucléosomes où qu'ils se trouvent par un mécanisme d'intégration indépendant de l'étape de réplication. Le remplacement de H3 par H3.3 n'aurait d'autre utilité fonctionnelle que de prévenir l'apparition de trous dans le paysage nucléosomal.

Variants communs de H2A De nombreux variants de l'histone H2A ont été identifiés chez les eucaryotes. Le variant H2A.X semble contribuer au mécanisme de réparation des liaisons de l'ADN double brin (LOWNDES et TOH 2005). Le variant H2A.Z semble tenir un rôle dans des mécanismes de régulation lié à la transcription, ainsi que dans la réparation de l'ADN (ZLATANOVA et THAKAR 2008 ; ALTAF, AUGER, COVIC et al. 2009 ; MARQUES, LAFLAMME, GERVAIS et al. 2010 ; ADKINS, NIU, SUNG et al. 2013). MacroH2A et H2A.B sont deux variants associés au mécanisme d'inactivation d'un des chromsomes X ayant lieu à chaque réplication chez la femelle mammifère. macroH2A se retrouve enrichi dans l'hétérochromatine du chromosome X inactivé (COSTANZI et PEHRSON 1998 ; CHADWICK 2001). À l'inverse, H2A.B y est spécifiquement déplété et semble impliqué dans le mécanisme d'épissage alternatif des gènes (SOBOLEVA, PARKER, NEKRASOV et al. 2017).

Variants spécifiques de la spermatogénèse Les cellules germinales males expriment de nombreux variants d'histones spécifiques : H2AL1, H2AL2, H2AL3, H2BL1, H2BL2, TH2B, H3T (FIGURE 1.16 ; EL KENNANI, ADRAIT, SHAYTAN et al. 2017). Les études de ces variants tendent à montrer que leur incorporation dans les nucléosomes génère de l'instabilité menant à leur éviction permettant un remplacement correct par les protamines (GAUTIER, ABBOTT, MOLLA et al. 2004 ; GOVIN, ESCOFFIER, ROUSSEAUX et al. 2007 ; LI, DONALD et GOLUB 2005 ; TACHIWANA, KAGAWA, OSAKABE et al. 2010).

1.1.3.3 Modifications post-traductionnelles des histones

Le remplacement par des variants n'est pas la seule source de variabilité des histones. En effet, ces dernières sont également sujettes à de nombreuses mo-

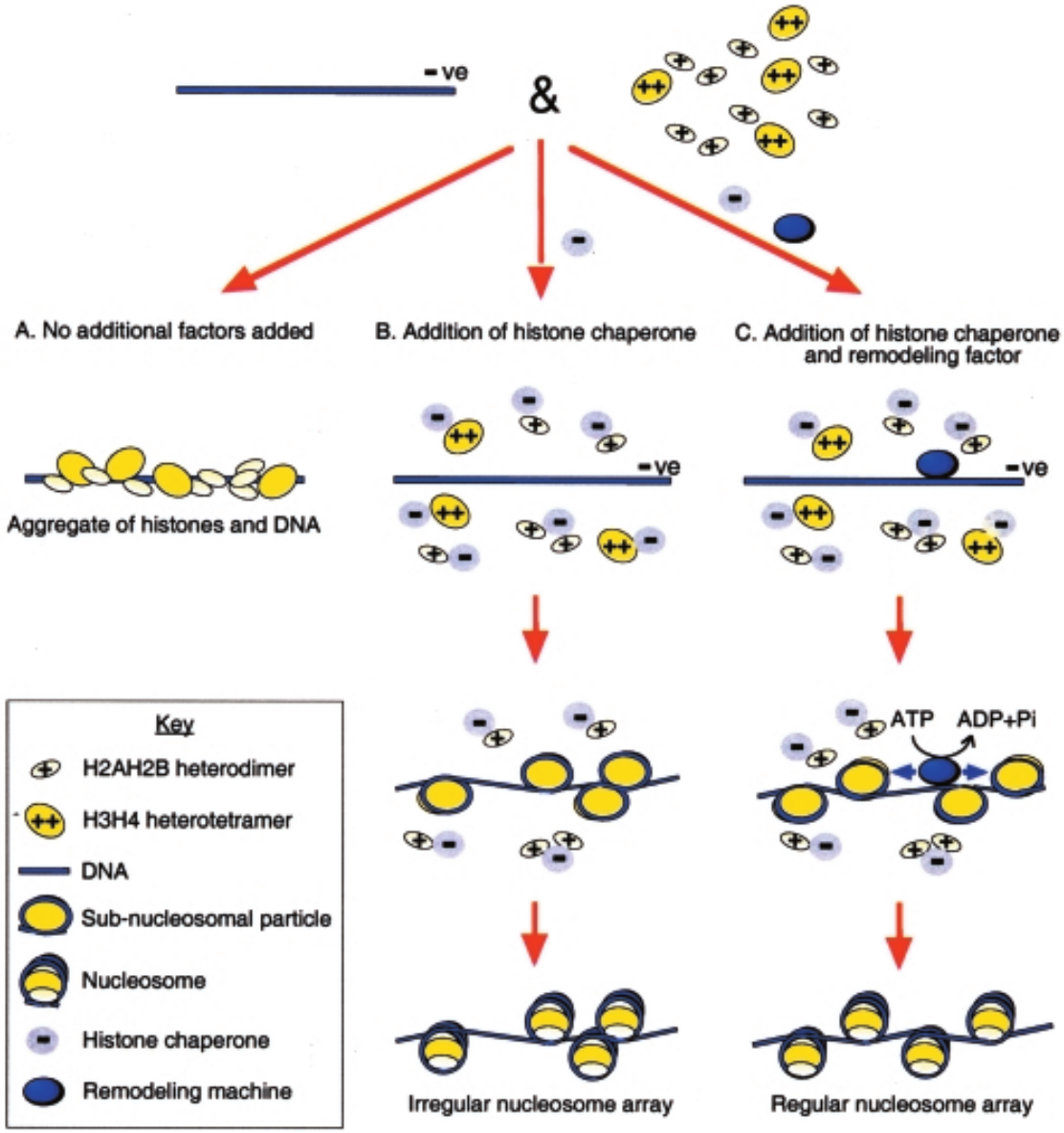


FIGURE 1.14 – Modèle d'assemblage de la chromatine grâce aux chaperonnes d'histones et aux remodeleurs ATP-dépendants. (A) Les histones chargées positivement et l'ADN chargé négativement possèdent une forte affinités et forment spontanéments des agrégats lorsque mélangés *in vitro*. (B) Les chaperonnes d'histones chargées négativement empêche l'agrégation directe des histones sur l'ADN et permet l'assemblage du nucléosome avant incorporation. (C) Les remodeleurs ATP-dépendants repositionnent les nucléosomes pour former une structure régulière. Extrait de TYLER 2002.

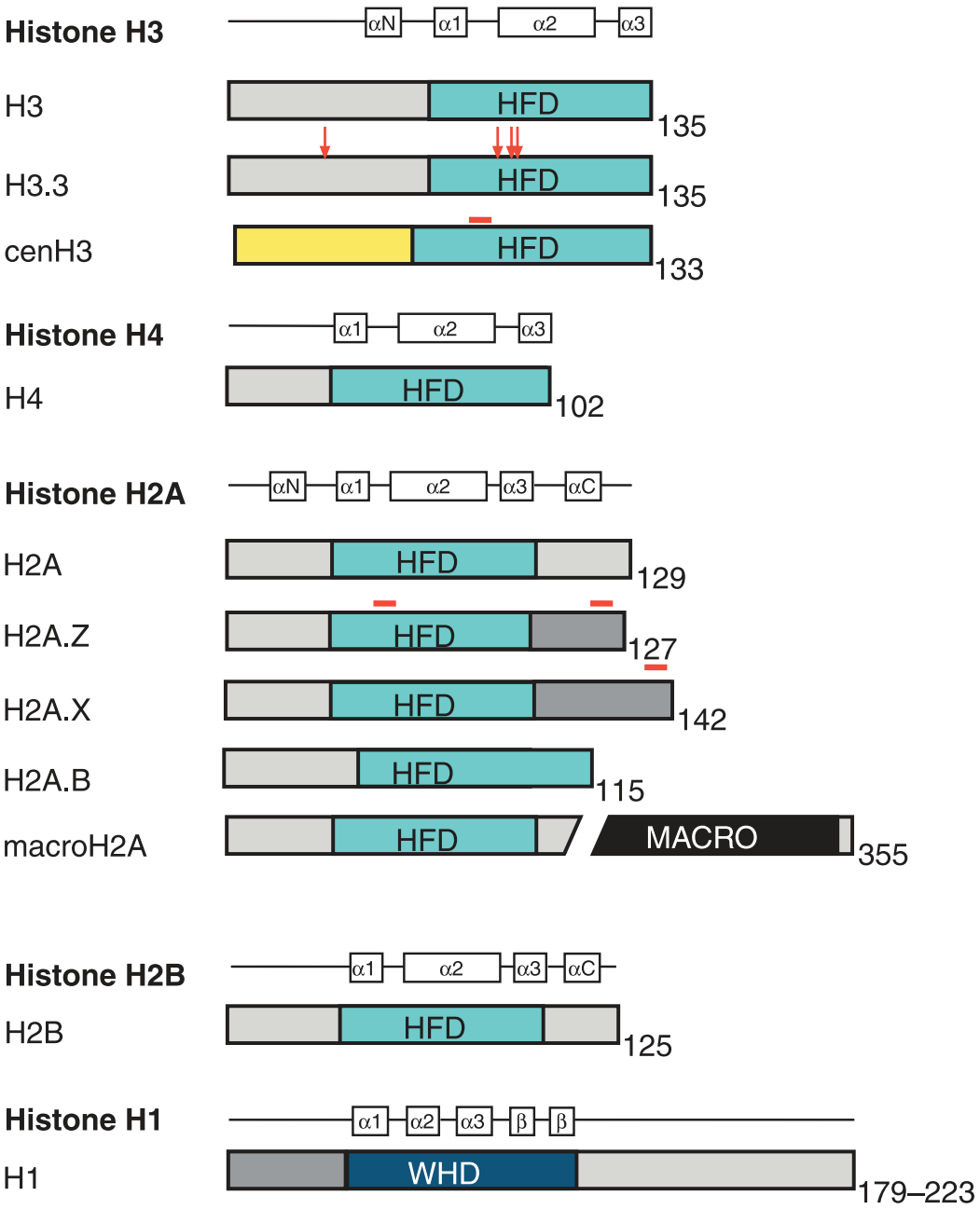


FIGURE 1.15 – Comparaison schématique de la structure des histones et des principaux variants. L'*Histone-Fold Domain* (HFD) correspond à la région permettant l'hétérodimérisation des histones. Le *Winged-Helix Domain* (WHD) correspond à la région de fixation à l'ADN. Les régions contenant des variations de séquence par rapport à l'histone canonique sont indiquées en rouge. MACRO est une région occupant les deux tiers de la région C-terminale du variant macroH2A et ne correspondant à aucune autre séquence histonique. La région N-terminale du variant cenH3 indiqué en jaune possède une variabilité telle qu'il n'est pas pas possible de l'aligner entre différents taxons. Extrait de HENIKOFF et SMITH 2015.

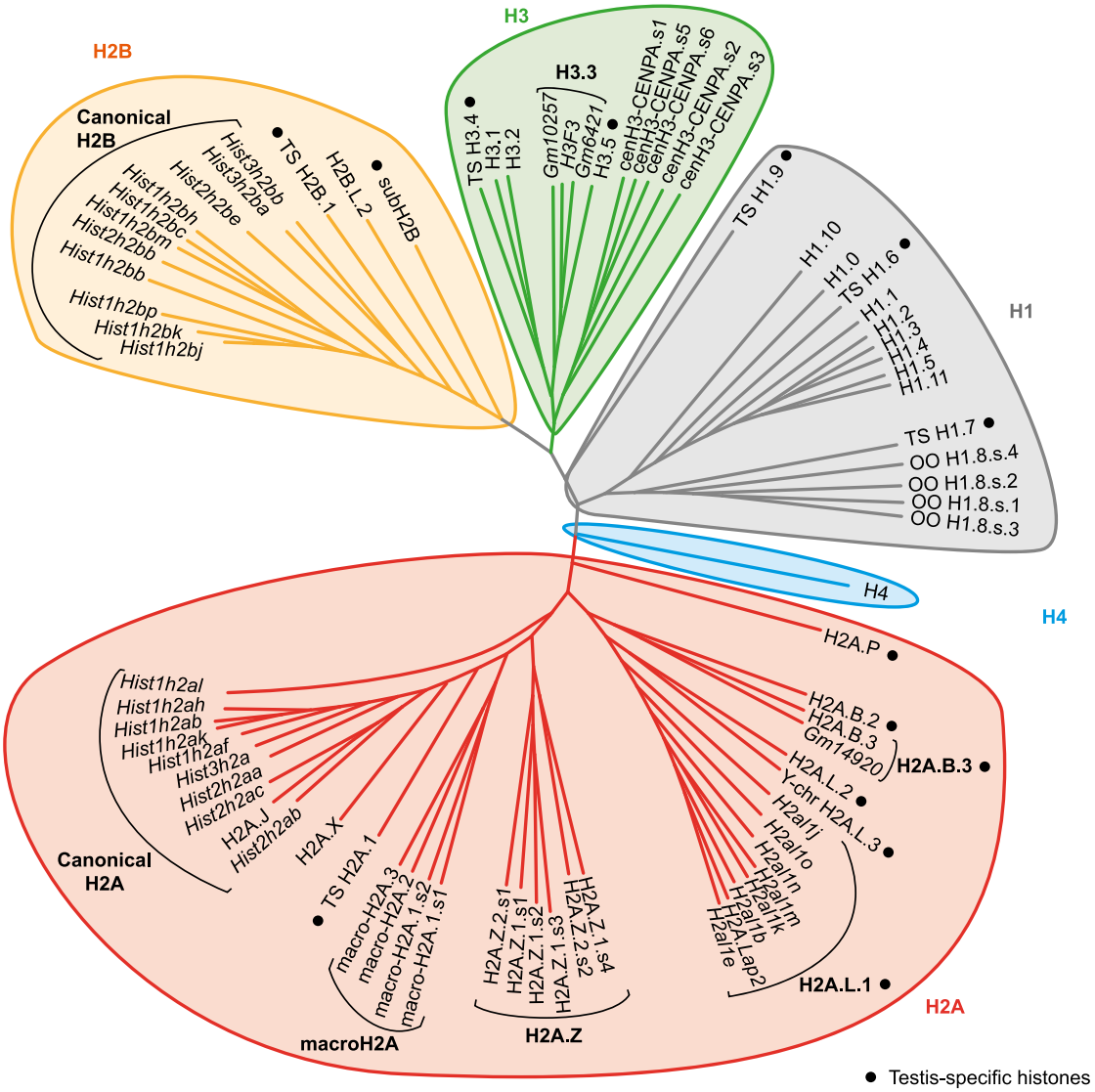


FIGURE 1.16 – Arbre phylogénétique des variants d'histones chez la souris. Les histones marquées d'un point noir sont spécifiques du testicule. Extrait de EL KENNANI, ADRAIT, SHAYTAN et al. 2017.

difications chimiques. Une vingtaine de ces modifications ont été identifiées (HUANG, SABARI, GARCIA et al. 2014 ; ZHAO et GARCIA 2015), affectant principalement les lysines, et dans une moindre mesure les sérines, arginines, théonines, tyrosines, glutamates et histidines (FIGURE 1.17). Elles touchent principalement les queues N- et C-terminales qui dépassent de leur structure, mais aussi le domaine globulaire constituant le cœur du nucléosome (LAWRENCE, DAUJAT et SCHNEIDER 2016).

Ces modifications sont déposées sur les histones par des familles d'enzymes rassemblées sous la dénomination fonctionnelle de *writers*. Les principales sont les *Histone AcetylTransferases* (HAT), les *Histone MethylTransferases* (HMT) et les *Protein Arginine MethylTransferases* (PRMT). Elles sont en général spécifiques d'une seule modification pour un nombre restreint de résidus ciblés (ALLIS, BERGER, COTE et al. 2007). Les *erasers*, telles que les *Histones Deacetylase* (HDAC) ou les *Lysines Demethylases* (KDM) sont des enzymes à la fonction antagoniste.

Les modifications d'histones sont décrites par une nomenclature où sont indiqués dans l'ordre l'histone concernée, puis le résidu et enfin sa modification (TURNER 2005). Le résidu est noté selon la nomenclature à une lettre pour les acides aminés, accompagné de son numéro dans la chaine peptidique de l'histone. L'histone et le résidu sont notés en majuscules alors que la modification est notée selon des abréviations en minuscules (TABLE 1.1). Ceci permet de combiner sans ambiguité plusieurs modifications pour la même histone. Ainsi, H3K4me3K27ac correspond à une histone 3 dont la lysine 4 est méthylée et la lysine 27 est acétylée. Il est aussi possible de marquer l'absence de modification de certains résidus. H3K4me3K9K27acK36 correspond ainsi à l'histone 3 modifiée précédente avec l'information complémentaire explicite que ni la lysine 9, ni la lysine 36 ne sont modifiées.

Les différentes modifications post-traductionnelles d'histones peuvent être reconnues par différents domaines de protéines (FIGURE 1.18). Les protéines les contenant peuvent être impliquées dans des mécanismes tels que la transcription, la réplication, la réparation de l'ADN et la régulation du cycle cellulaire (LAWRENCE, DAUJAT et SCHNEIDER 2016). Des combinaisons de modifications d'histones peuvent générer des interactions synergiques ou antagonistes pour les différents complexes protéiques capables ou incapables de se fixer à la chromatine les possédant, entrainant ainsi des modifications structurales et fonctionnelles de la chromatine. Selon un paradigme populaire, les combinaisons de modifications d'histones parsemant le génome constituent un "code histone" formant un langage entre la chromatine et les complexes protéines pouvant interagir avec (STRAHL et ALLIS 2000).

Les six modifications les plus étudiées aux travers des différents types cellulaires correspondent à des modifications de la queue de l'histone 3 : H3K27ac, H3K27me3, H3K4me1, H3K4me3, H3K9me3 et H3K36me3. L'intérêt pour ces marques réside dans le fait que la combinaison de leur présence ou absence à un locus donné est un bon estimateur de l'état fonctionnel local de la chromatine (FIGURE 1.19 ; TABLE 1.2 ; SECTION 2.1).

résidu	abrév.	modification	résidu	abrév.	modification
K,R	ac	acétylation	K	me1	monométhylation
K,E	ar1	mono-ADP-ribosylation	K	me2	dimethylation
K,E	arn	poly-ADP-ribosylation	R	me2a	asymmetrical me2
K	bio	biotinylation	R	me2s	symmetrical me2
K	bu	butyrylation	K	me3	trimethylation
K	cit	citrullination	S,T	og	O-GlcNAcylation
K	cr	crotonylation	Y	oh	hydroxylation
K	for	formylation	Y	ox	oxidation
C	gt	gluthathionylation	S,T,Y,H	ph	phosphorylation
K	hib	2-hydroxyisobutyrylation	K	suc	succinylationa
F,P	iso	isomerization	K	su	SUMOylation
K	mal	malonylation	K	ub	ubiquitination

TABLE 1.1 – Réactions chimiques affectant des résidus d'histones et leurs abréviations. Constitué à partir de TURNER 2005, HUANG, SABARI, GARCIA et al. 2014 et ZHAO et GARCIA 2015.

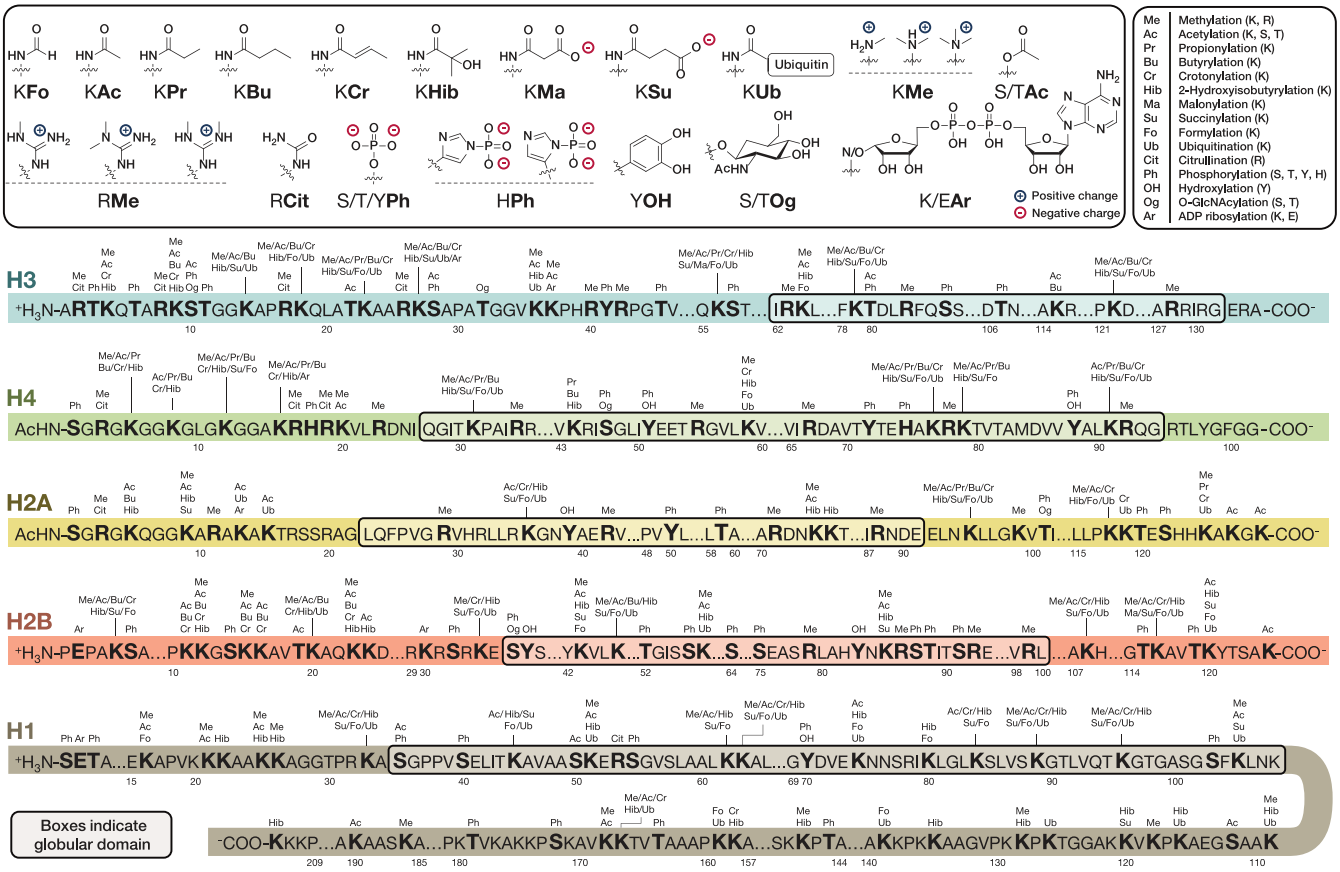


FIGURE 1.17 – Principales modifications post-traductionnelles des histones. Adapté de HUANG, SABARI, GARCIA et al. 2014.

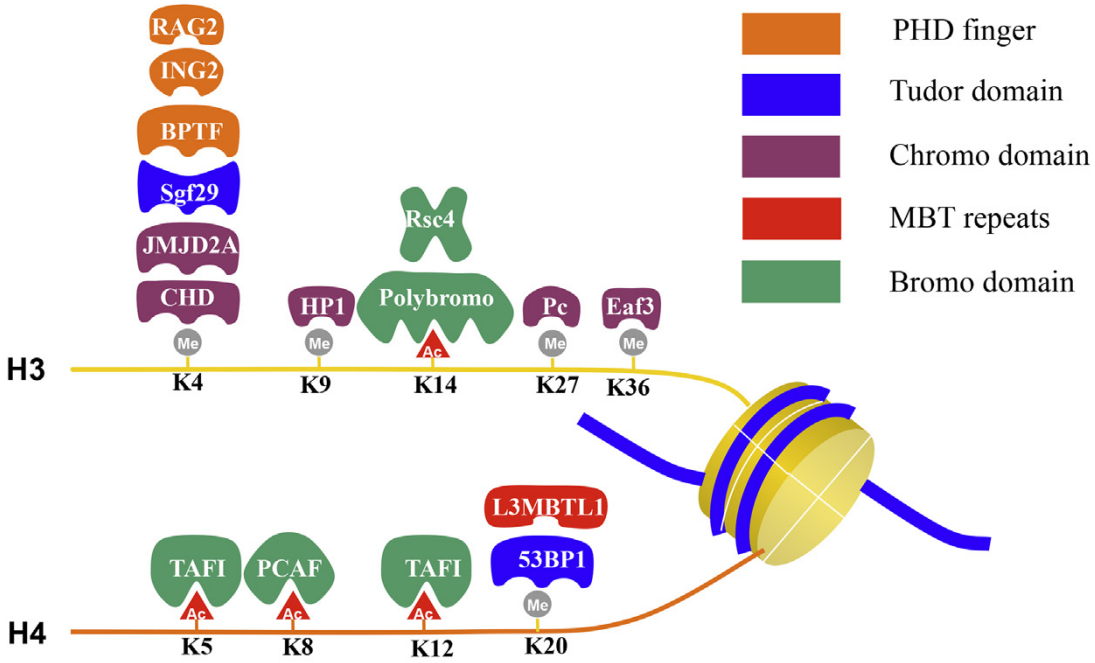


FIGURE 1.18 – Principaux domaines fixant les modifications d'histone les plus étudiées. Extrait de WANG, HAN, FAN et al. 2015.

Augmenter le nombre de marques épigénétiques permet d'atteindre une granularité plus fine dans l'estimation de l'état local de la chromatine, mais cela implique avec les technologies actuelles, un coût supplémentaire qui n'est pas nécessairement justifié, même pour des projets d'envergure (SECTION 1.3.3.5). En effet, certaines marques tendent à colocaliser très fortement ce qui réduit l'intérêt de les étudier toutes étant donnée qu'une seule suffirait en tant que marqueur. C'est notamment ce qui semble le cas pour l'acétylation et la butyrylation des lysines 5 et 8 de l'histone H4 étudiée à différents stades de la spermiogénèse (SECTION 2.2.1).

1.1.3.4 Dynamiques chromatinienne et épigénétiques

Les modifications épigénétiques sont des phénomènes dynamiques et réversibles au cours du temps à plusieurs échelles. Certaines modifications peuvent être transmises à la descendance (HEARD et MARTIENSSEN 2014 ; TREROTOLA, RELLI, SIMEONE et al. 2015), d'autres sont acquises au cours de la vie d'un individu, naturellement avec l'âge (SLIEKER, RELTON, GAUNT et al. 2018 ; CICCARONE, TAGLIATESTA, CAIAFA et al. 2018), ou en réponse à des stimuli environnementaux particuliers (TURNER 2009 ; DIAS, MADDOX, KLENGEL et al. 2015). À l'échelle cellulaire, ces stimuli peuvent entrainer des modifications épigénétiques de sorte à ce que l'accès à un épigénotype, même partiel (accessibilité chromatinienne, méthylation de l'ADN ou certains modifications d'histones), permette de caractériser sans ambiguité des sous-populations proches de cellules

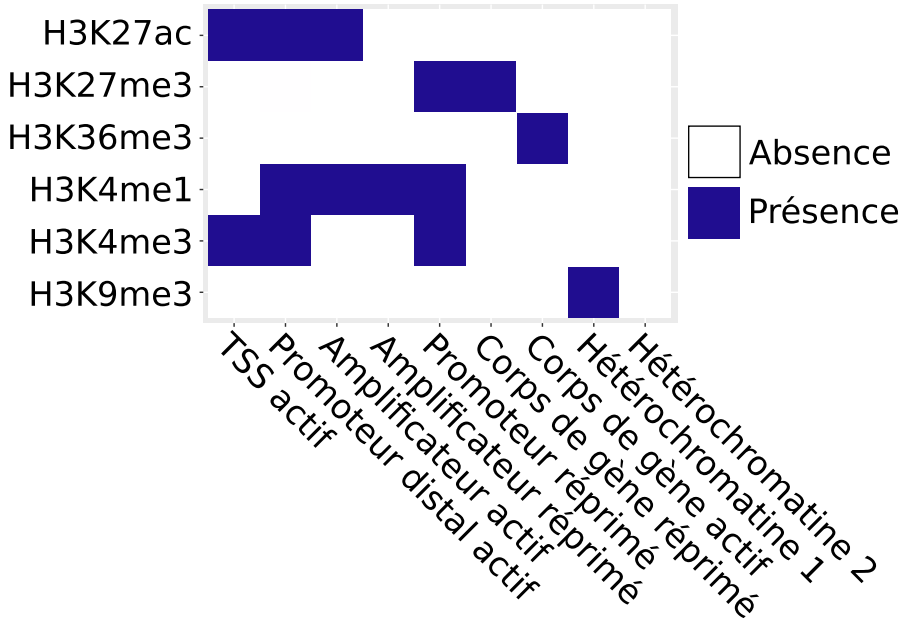
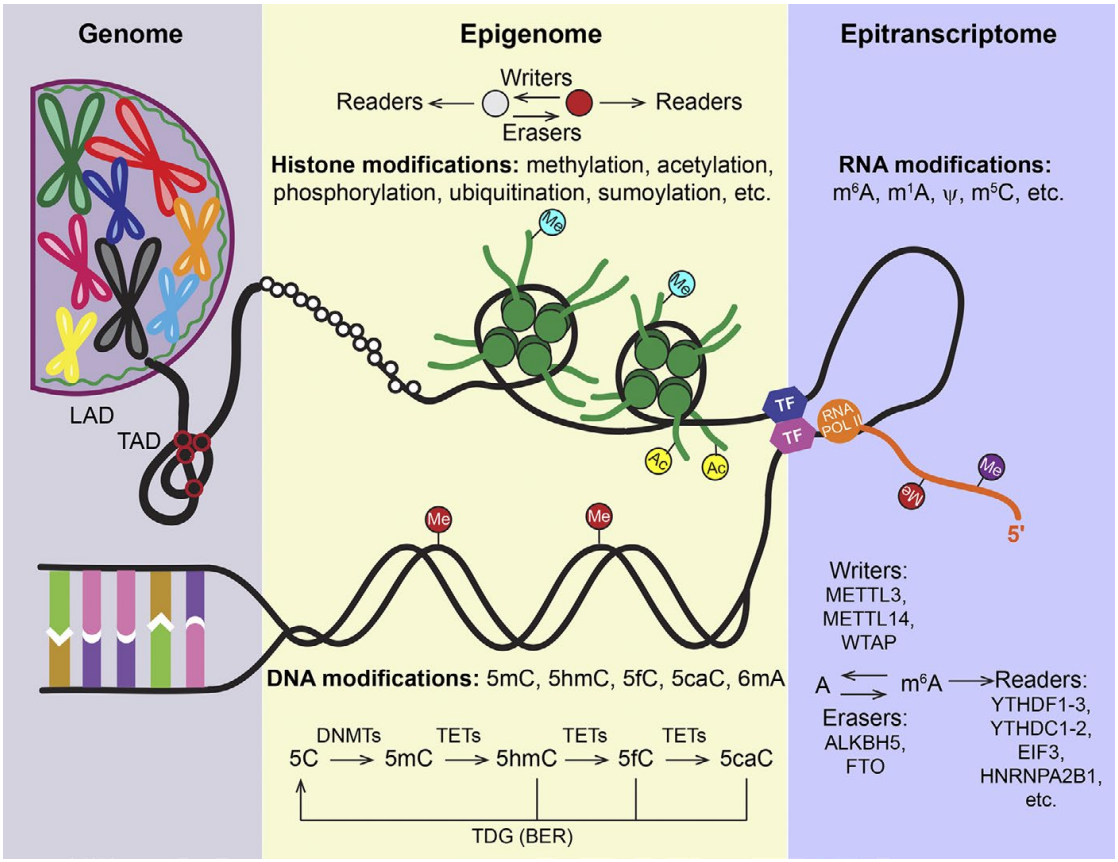


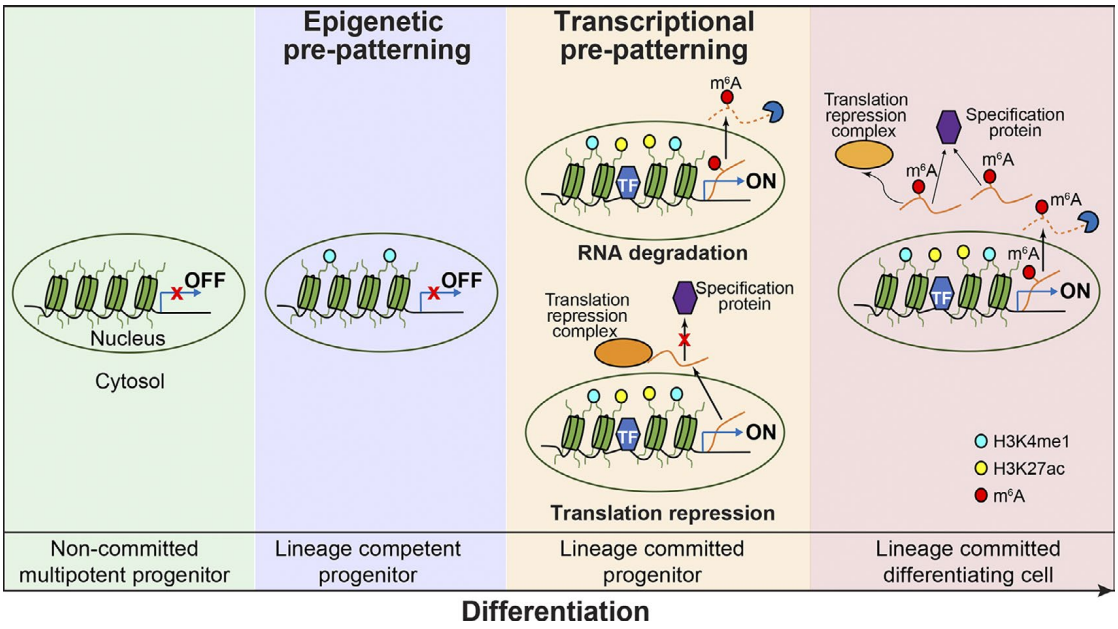
FIGURE 1.19 – Modèle simplifié de segmentation du génome basé sur la présence des six modifications d'histones principales.

en différentiation (SANTA-PAU, JUAN, PANCALDI et al. 2017), voire de prédire les devenirs divergents de cellules possédant des caractères phénotypiques identiques (ULIRSCH, LAREAU, BAO et al. 2019 ; GROSSELIN, DURAND, MARSOLIER et al. 2019).

Les grandes lignes des dynamiques sous-jacentes à la régulation de l'expression des gènes par des mécanismes d'écriture, de lecture et d'effacement des modifications épigénétiques sont connues (FIGURE 1.20a ; BELL 2016 ; YOON, VISSERS, MING et al. 2018), tout comme celles sous-jacentes aux processus d'engagement et de persistance dans une voie de différentiation (FIGURE 1.20b ; YOON, VISSERS, MING et al. 2018). Cependant, il existe une grande diversité des mécanismes et acteurs épigénétiques impliqués dans cette différentiation en fonction des types cellulaires étudiés. De ce fait de nombreuses approximations persistent dans les modèles quand ces derniers ne sont pas simplement incomplets voire inexacts. Les principaux projets de recherche auxquels j'ai pu participer pendant ma thèse ont pour objectif de faire la lumière sur certaines zones d'ombres dans le cas de deux sujets d'études : la thymopoïèse humaine et la spermatogénèse murine. Leurs spécificités sont présentées dans la section suivante.



(a) Régulation de l'expression des gènes



(b) Différentiation cellulaire

FIGURE 1.20 – Modèles de dynamiques épigénétiques. Extrait de YOON, VISSERS, MING et al. 2018.

1.2 Spécificités des modèles d'études de dynamiques épigénétiques

1.2.1 Thymopoièse humaine

1.2.1.1 Produits de la thymopoïèse : les lymphocytes T

L'appellation de lymphocyte T rassemble différents types de cellules du système immunitaire dont la maturation a lieu dans le thymus. Les lymphocytes T sont divisés en sous-groupes en fonction de leur fonction. Les lymphocytes T cytotoxiques sont capables de détruire les cellules tumorales ou infectées par des virus présentant des antigènes spécifiques à travers le Complexe Majeur d'Histocompatabilité (CMH) de classe 1. Ils sont aussi appelées plus brièvement CD8+ de par la présence de ce marqueur à leur surface. Les lymphocytes T auxiliaires sont des intermédiaires de la réponse immunitaire qui s'activent lorsqu'un antigène spécifique est reconnu par l'intermédiaire du CMH de classe 2 à la surface de cellules dendritiques, macrophages ou lymphocytes B. Ils produisent alors des cytokines qui vont induire l'activation ou la maturation d'effecteurs de la réponse immunitaire. Les lymphocytes T auxiliaires sont appelés plus brièvement CD4+ de par la présence de ce marqueur à leur surface. Les lymphocytes de ces deux types expriment un récepteur membranaire (TCR) dont la séquence est spécifique à chacun grâce à un mécanisme de recombinaison génétique dit V(D)J (SECTION 1.2.1.3) ayant lieu au cours de la thymopoïèse. C'est ce récepteur, en association avec CD4 ou CD8, qui permet de reconnaître l'antigène présenté par un CMH. Il existe également d'autres types de lymphocytes T (régulateurs, mémoire, NKT, MAIT et $\gamma\delta$) qui ne seront pas abordés en détail.

1.2.1.2 Organe de la thymopoïèse : le thymus

Le thymus est l'organe spécialisé pour la différentiation des lymphocytes T chez les mammifères. Il se situe dans la partie supérieure du thorax, au dessus du cœur dans le médiastin antéro-supérieur et antéro-moyen (FIGURE 1.21). C'est un organe entouré d'une capsule conjonctive et bilobé, où chaque lobe est divisé en lobules. Dans ces derniers, on distingue d'une part le cortex, la partie externe, qui contient les cellules lymphoïdes T immatures et en prolifération, d'autre part, la médulla, qui contient les cellules T plus matures. Les corpuscules de Hassal sont des structures arrondies et éosinophiles situées dans la médulla. Les corpuscules de Hassal produisent des cytokines pouvant influencer les populations thymiques mais leur fonction précise reste débattue.

Le thymus évolue en taille, masse et activité au cours de la vie. Il augmente de taille chez l'enfant jusqu'à la puberté, puis rétrécit graduellement avec le vieillissement. Chez l'enfant, il présente une apparence rouge sombre de par sa vascularisation puis évolue graduellement vers le gris puis le jaune chez la personne âgée par l'infilration de tissu adipeux. La thymopoïèse reste opérationnelle tout au long de la vie mais à un niveau de plus en plus faible.

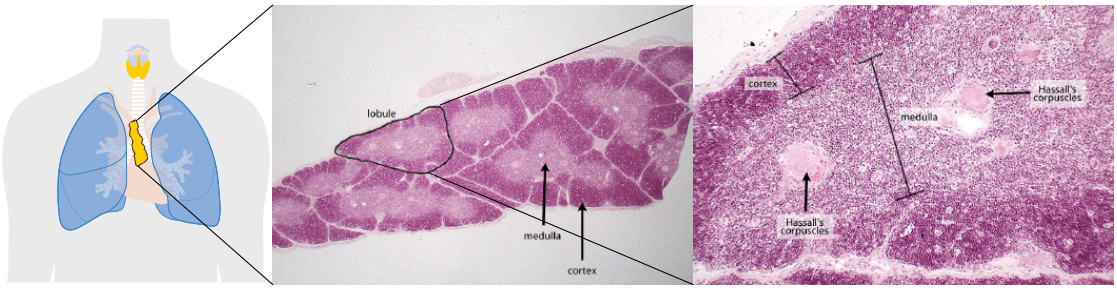


FIGURE 1.21 – Anatomie du thymus. Créé à partir du schéma de CC Wikimedia et des coupes transversales du Dartmouth College.

1.2.1.3 Étapes de la thymopoïèse

Origine des précurseurs hématopoiétiques La thymopoïèse constitue l'une des branches de l'hématopoïèse, le processus physiologique de production des cellules sanguines et éléments figurés du sang (FIGURE 1.22). Une petite population de cellules souches hématopiétiques est à l'origine de l'ensemble des cellules sanguines. Elles sont caractérisées par leur multipotence et leur capacité à s'auto-renouveler, migrer dans le sang et coloniser différents territoires hématopoiétiques (PASSEGUE, JAMIESON, AILLES et al. 2003). Situées principalement dans des niches de la moelle osseuses, elles se divisent asymétriquement de façon à produire d'une part des cellules identiques afin de maintenir un nombre constant de cellules souches, d'autres parts des progéniteurs multipotents. Ces derniers peuvent maturer en différents types de progéniteurs engagés dans une voie de différentiation hématopoiétique : lymphoïde (CLP), myéloïde (CMP) ou mixte (LMPP).

Colonisation thymique Les progéniteurs lymphoïdes dérivés de la moelle osseuse entrent dans le thymus par vagues sucessives au cours de la vie au niveau de la jonction cortico-médullaire vascularisée (FIGURE 1.23). Ils s'y différentient en progéniteurs thymiques précoces mais conservent la capacité de s'auto-renouveler et de se différencier en lymphocytes T mais aussi en cellules NK, dendritiques, monocytes et mastocytes (ROTHENBERG, MOORE et YUI 2008).

Engagement dans la voie de différentiation T Les progéniteurs thymiques précoces migrent dans le cortex jusqu'à atteindre la zone sub-capsulaire dans un processus de maturation où l'on peut distinguer plusieurs étapes phénotypiques sur la base des marqueurs de surface : DN1 ($CD34^+$, $CD38^-$, $CD1a^-$), DN2 ($CD34^+$, $CD38^+$, $CD1a^-$) et DN3 ($CD34^+$, $CD38^+$, $CD1a^+$). Les cellules DN3 sont caractérisées par leur engagement complet dans la voie de différentiation T.

Réarrangements des TCR L'initiation des réarrangements des gènes des chaîne du TCR marque l'engagement irréversible des cellules sur une voie de différentiation de lymphocytes T. Les cellules commencent par réarranger les segments

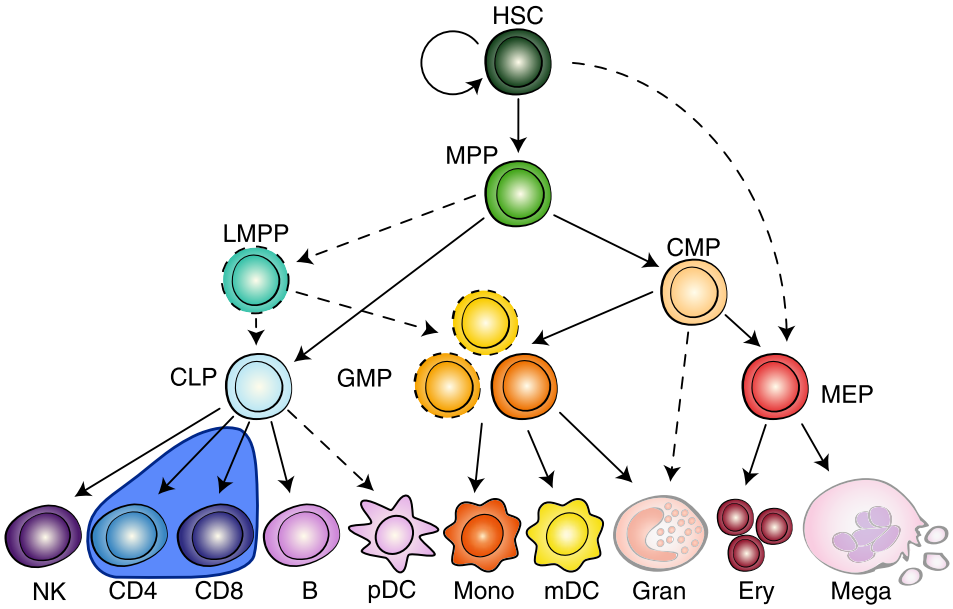


FIGURE 1.22 – Lignées hématopiétiques. La thymopoïèse est surlignée en bleue. HSC : Cellule Souche Hématopiétique. MPP : Progéniteur Multi-Potent. LMPP : Progéniteur Multi-Potent engagé dans la voie Lymphoïde. CMP : Progéniteur Myéloïde Commun. CLP : Progéniteur Lymphoïde Commun. GMP : Progéniteur Granulocyte-Macrophage. MEP : Progéniteur Mégacaryocyte-Erythroïde. Adapté de ULIRSCH, LAREAU, BAO et al. 2019.

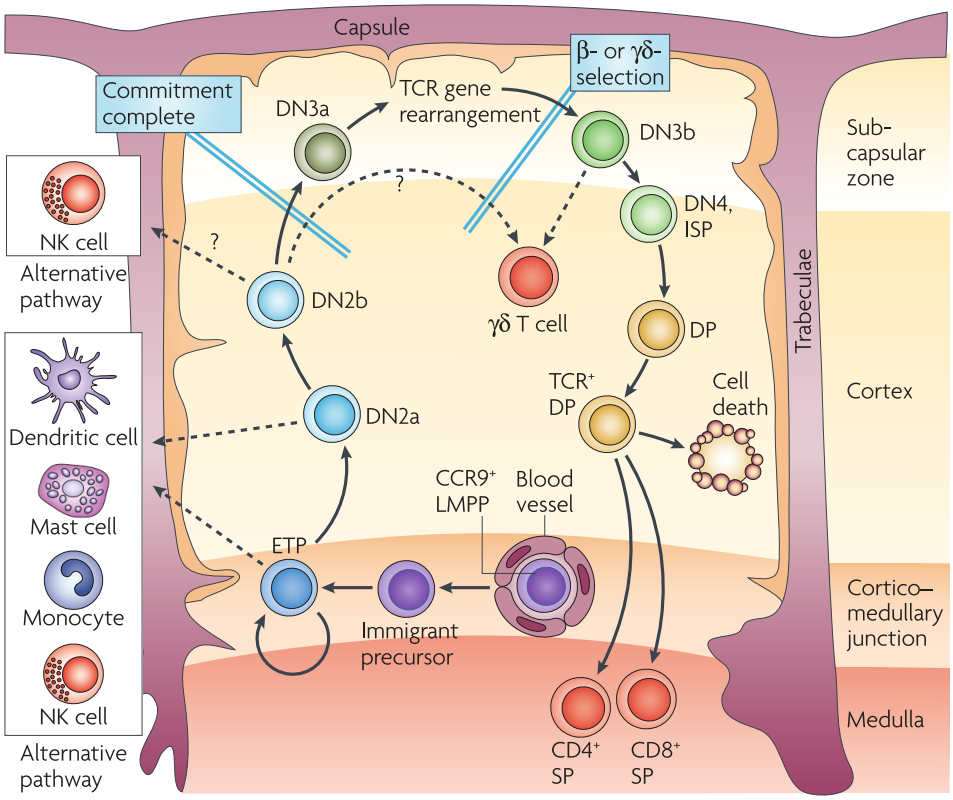


FIGURE 1.23 – Étapes de la thymopoïèse, représentées au travers d'une coupe transversale schématique d'un lobule thymique. CCR9 : CC chemokine receptor 9. Extrait de ROTHENBERG, MOORE et YUI 2008.

D et J du locus du TCRδ, puis poursuivent par le réarrangement D-J du TCRβ ou par le réarrangment V-J du TCRγ (FIGURE 1.24). Le TCRα reste invariant à ce stade. Ces réarrangements sont réalisés par un mécanisme impliquant la reconnaissance de deux Séquences Signal de Recombinaison, 12RSS et 23RSS, par un complexe d'endonucléase composé des protéines RAG-1 RAG-2 (FIGURE 1.25). Deux de ces complexes peuvent reconnaître les séquences RSS et les cliver avant d'interagir entre eux pour former un complexe protéines-ADN-clivé appelé apparié. Le brins d'ADN clivés sont ensuite ciblé par des complexes protéiques spécialisés dans le processus de réparation de l'ADN par jonction d'extrémités non homologues. Ce processus de réparation ajoute une séquence de quelques nucléotides variable entre les deux brins clivés qui contribue la grande variabilité des séquences gènes des chaînes du TCR.

β-sélection Les thymocytes au stade DN3 expriment la chaîne invariante du TCRα et la version en cours de réarrangement du TCRβ. Si le réarrangement du TCRβ permet la création d'un complexe de pré-TCR fonctionnel, des signaux sont produits pour stopper les mécanismes de réarrangement du TCRβ et inactiver l'allèle alternatif. La cellule initie alors une phase de prolifération constituée de 5 à 6 cycles de division, ainsi que le réarrangement du TCRα. Elle perd le marqueur de surface CD34 qui est remplacé par le marqueur CD4. La cellule passe alors au stade de différentiation Immature Simple Positif (ISP) correspondant à l'engagement irréversible dans la voie de différentiation lymphocytaire $\alpha\beta$. Les cellules qui ne passent pas la β-sélection peuvent passer une sélection alternative $\gamma\delta$ et se différentier en lymphocytes $\gamma\delta$.

Sélection positive Lorsque le réarrangement du TCRα se termine, les cellules expriment un TCR$\alpha\beta$ complet et le co-récepteur CD8 est exprimé conjointement au CD4. Cette étape correspond au stade de thymocytes double positif. Les cellules migrent dans cortex thymique où des antigènes du soi leurs sont présentés par l'intermédiaire de MHC à la surface des cellules corticales épithéliales. Les cellules capables d'interagir avec un des MHC vont recevoir un signal de survie alors que celles qui en sont incapable mourront par négligence. Si le thymocyte est capable d'interagir avec un MHC de classe 1, il recevra un signal entrainant la diminution de l'expression de CD4 et poursuivra sa différentiation en lymphocyte T CD8. Inversement, il se différentiera en lymphocyte T CD4 s'il est capable d'interagir avec un MHC de classe 2.

Sélection négative Les thymocytes simple positif ayant passé la sélection positive migrent vers la médulla où des antigènes du soi leurs sont de nouveau présentés par l'intermédiaire du MHC de classe 1 de cellules thymiques épithéliales (mTECs) ou du MHC de classe 2 de cellules dendritiques. Les thymocytes interagissant trop fortement avec l'antigène du soi reçoivent un signal apoptotique menant à la mort cellulaire. Une faible proportion de ceux-ci sont sélectionnés pour devenir lymphocytes Treg. Les thymocytes simple positif ayant passé avec

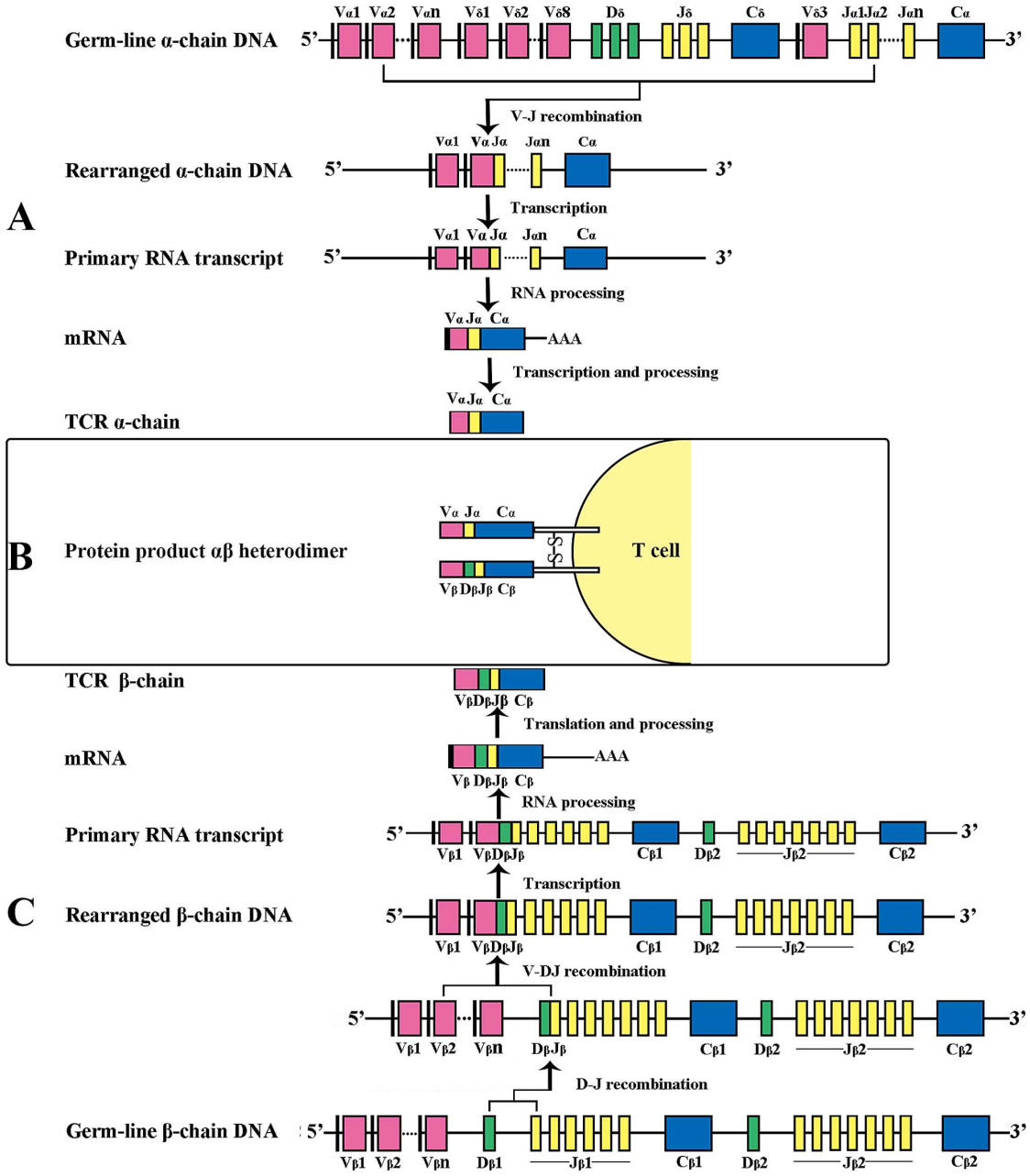


FIGURE 1.24 – Principe combinatoire des réarrangement du TCR humain formant les gènes fonctionnels encodant l'hétérodimère αβ. (A) La recombinaison V-J de la chaîne du TCRα sélectionne et relie un des 46 segments V à un des 51 segments J. Le transcrit du gène est épissé pour ne conserver que le segment V-J réarrangé et relié au segment constant C. (B) Structure de l'hétérodimère αβ. (C) La recombinaison V-D-J de la chaîne du TCRβ s'effectue de façon analogue mais en deux étapes, D-J puis V-D-J. Extrait de WANG, YU, HE et al. 2016

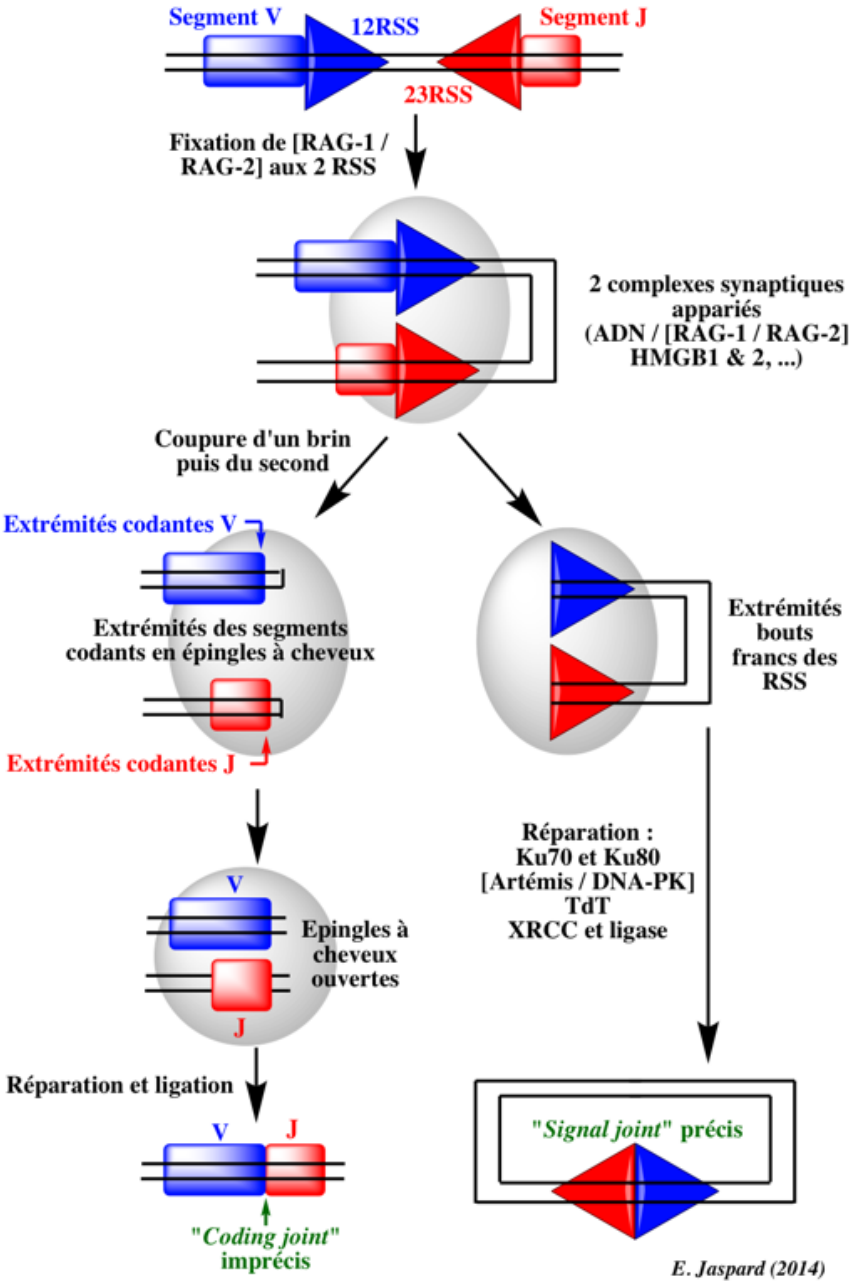


FIGURE 1.25 – Principe moléculaire des recombinaisons V(D)J. Extrait de E. Jaspard 2014

succès les sélections β, positive et négative quittent le thymus pour devenir des lymphocytes T matures. Ceux-ci ne représentent que 2% des cellules engagées dans cette voie de différentiation lymphocytaire.

1.2.1.4 Dynamiques épigénétiques au cours de la thymopoïèse

Les lymphocytes T sont issus d'un processus itératif de choix engageant progressivement les précurseurs hématopoiétiques multipotents vers la voie de différentiation T dans le thymus. Cette différentiation nécessite la mobilisation d'un programme d'expression de gènes spécifiques devant permettre notamment la production des TCR fonctionnels, la sélection des lymphocytes basées sur l'affinité de leur TCR et la tenue du rôle d'effecteur du système immunitaire.

Chez la souris, des éléments de ces programmes d'expression sont caractérisés. Ainsi, le cas des complexes médiateurs du réarrangement des gènes des chaînes du TCR a été déjà évoqué (SECTION 1.2.1.3 ; ROTHENBERG, MOORE et YUI 2008), et certains facteurs de transcription clés de différents stades de différentiation sont identifiés (FIGURE 1.26 ; ZHANG, MORTAZAVI, WILLIAMS et al. 2012 ; YUI et ROTHENBERG 2014). Des mécanismes généraux de réorganisation des compartiments chromatiniens contrôlés par ces facteurs sont proposés afin d'expliquer l'activation des programmes d'expression des stades de la lignée T et l'inactivation de ceux des stades précurseurs (FIGURE 1.27 ; HU, CUI, FANG et al. 2018).

Un de ces compartiments s'ouvrant précocement inclut le locus partagé du TCRD et TCRA. Des mécanismes de régulation plus fine semblent alors nécessaires pour expliquer la désynchronisation observée entre les réarrangements du TCRD et du TCRA (SECTION 1.2.1.3). L'implication dans cette régulation de l'unique amplificateur du locus, l'enhancer α (Eα), a été établie (BASSING, TILLMAN, WOODMAN et al. 2003), mais son mécanisme épigénétique d'activation reste non élucidé.

Plus généralement, plusieurs études ont associé la dynamique d'activation des promoteurs et amplificateurs au cours de la différentiation T murine avec la présence des marques épigénétiques H3K27ac, H3K4me1, H3K4me2 et H3K4me3, cette dernière marque étant surtout associée aux amplificateurs (PEKOWSKA, BENOUKRAF, ZACARIAS-CABEZA et al. 2011 ; ZHANG, MORTAZAVI, WILLIAMS et al. 2012). Cependant, le manque de données générées, chez la souris et encore plus chez l'homme, n'a jamais permis de réaliser une analyse croisée des dynamiques de modifications de l'ADN, de l'accessibilité de la chromatine et des marques d'histones au cours de la thymopoïèse. Une telle analyse est proposée dans ce manuscrit (SECTION 2.1).

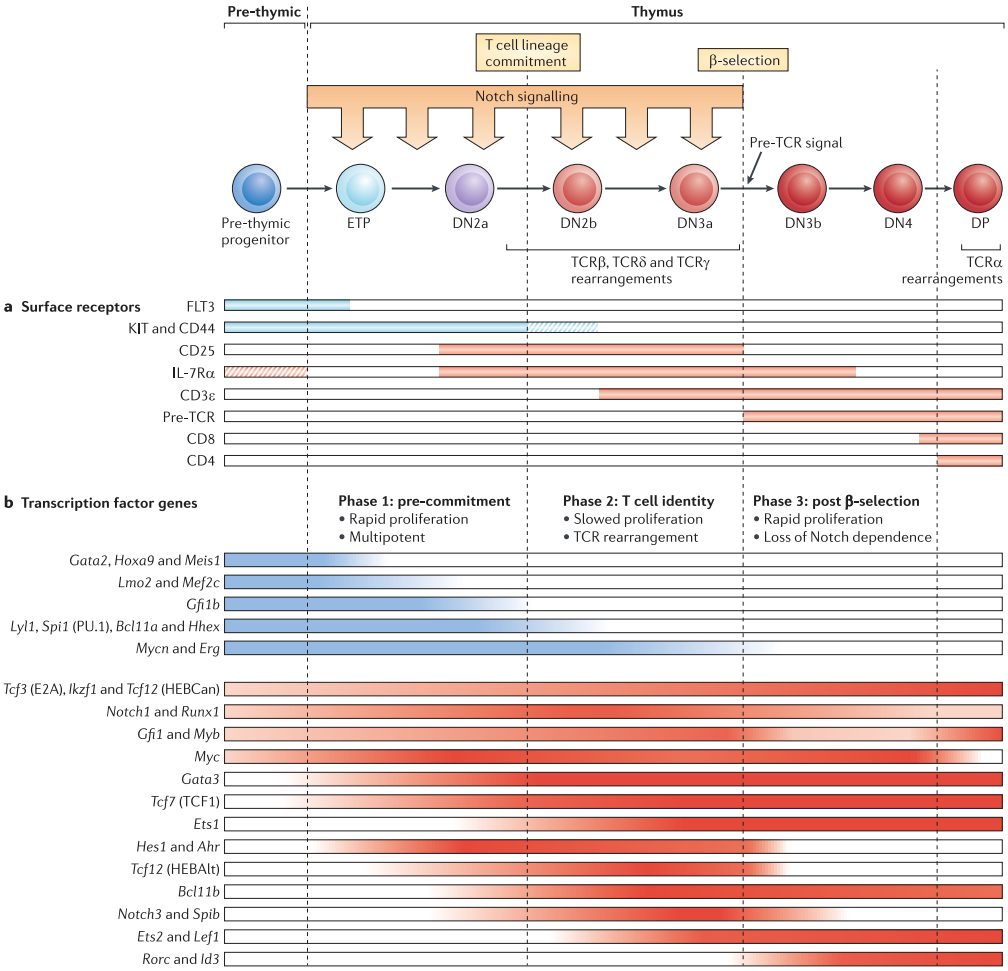


FIGURE 1.26 – Vue d'ensemble des dynamiques épigénétiques au cours de la thymopoïèse. Extrait de YUI et ROTHENBERG 2014.

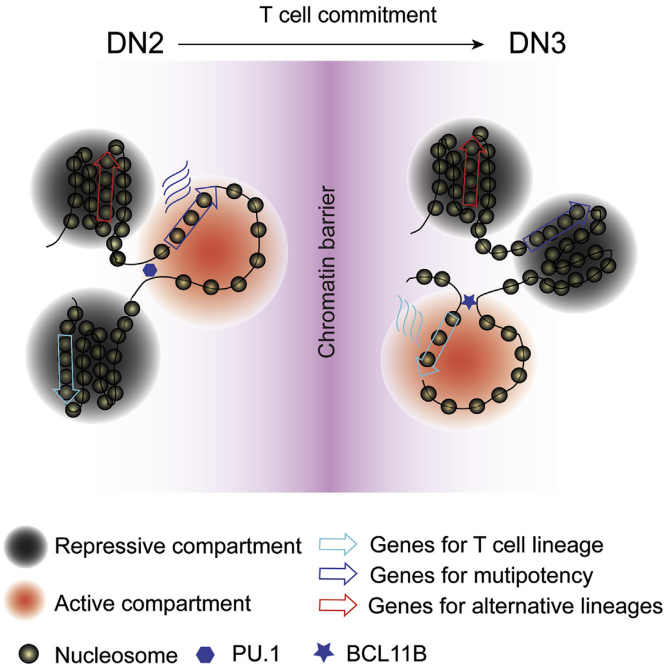


FIGURE 1.27 – Vue schématique des réorganisations des compartiments chromatiniens lors de l'engagement dans la voie de différentiation T. Extrait de HU, CUI, FANG et al. 2018.

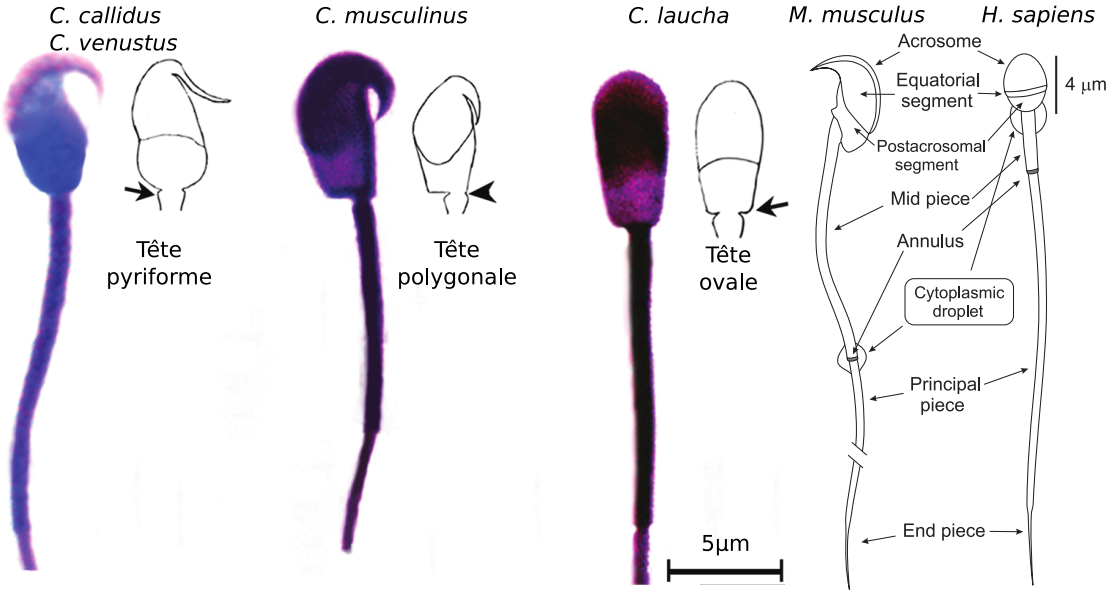
1.2.2 Spermatogénèse murine

1.2.2.1 Produits de la spermatogénèse : les spermatozoïdes

Le spermatozoïde est la cellule reproductrice, ou gamète, mâle intervenant dans la reproduction sexuée. Il comprend un noyau, un système de locomotion et une vésicule nommée acrosome contenant des enzymes libérées au contact de l'ovule afin d'y faciliter sa pénétration. Une grande diversité des formes de locomotion existe au sein du vivant : un ou plusieurs flagelles, des cils ou des lamellipodes. Un unique flagelle est la configuration standard pour les mammifères. Des différences de formes existent cependant entre espèces mêmes proches (Figure 1.28a ; Rossi, De La Sancha, Luaces et al. 2018). Contrairement à son positionnement central chez l'homme, le flagelle est excentré chez la souris (Figure 1.28b).

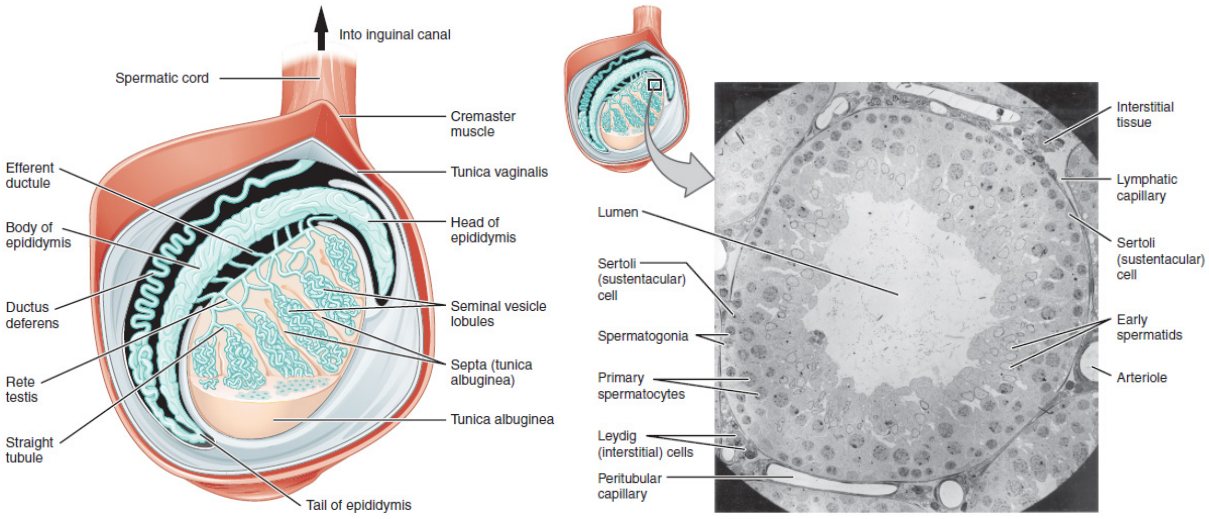
1.2.2.2 Organe de la spermatogénèse : le testicule

Chez les mammifères, la spermatogénèse se déroule dans le testicule. Cet organe ovoïde est entouré par une gaine de tissue conjonctif fibreux, l'albuginée (Figure 1.29a). Le parenchyme testiculaire est constitué de lobules séparés par des cloisons conjonctives. Au sein de ces lobules se trouvent les tubes séminifères compactés en structure hélicoïdale. Tous les stades de différentiation des gamètes mâles s'y retrouvent, ordonnés de la périphérie vers la lumière du stade



(a) Extrait de Rossi, De La Sancha, Luaces et al. 2018 (b) Extrait de Darszon, Nishigaki, Beltran et al. 2011

Figure 1.28 – Comparaison des formes de spermatozoïdes chez certains rongeurs et l'homme.



(a) Structure interne du testicule (b) Section transversale de tube séminifère

FIGURE 1.29 – Organisation du testicule. Adapté de © OpenStax College / CC BY 3.0.

le plus précoce au plus mature (FIGURE 1.29b). Les tubes séminifères sont structurés par des cellules somatiques de soutien, les cellules de Sertoli. Ils sont reliés aux deux extrémités sur le *rete testis* qui collecte les spermatozoïdes et mène à l'épididyme. Les spermatozoïdes terminent leur maturation au travers de l'épididyme puis sont complémentés d'autres sécrétions dnas le canal déférent pour constituer le sperme.

1.2.2.3 Étapes de la spermatogénèse

La spermatogénèse repose sur trois phases majeures : une phase mitotique, une phase méiotique et une phase de différentiation post-méiotique appelée spermiogénèse (FIGURE 1.30) et durent respectivement 11, 10 et 14 jours chez la souris.

Phase mitotique La phase mitotique permet la renouvellement et la prolifération des cellules souches germinales. Ce rôle est tenu par les spermatogonies de type A. Périodiquement, certaines de ces cellules initient une différentiation en spermatogonies de type B. Leur activité transcriptionnelle augmente fortement pour permettre la synthèse des protéines nécessaires pour la phase méiotique suivante.

Le déclenchement de cette différentiation est cyclique selon une période d'environ 8 jours et demi. L'organisation des tubes séminifères en quatre à cinq couches superposées de gamètes en différentiation est la résultante de la période du cycle et de la durée de différentiation. Pour chaque section de tube séminifère, il est possible d'identifier un stade parmi douze au sein d'un cycle en se basant sur les associations phénotypiques des différentes couches de gamètes observés

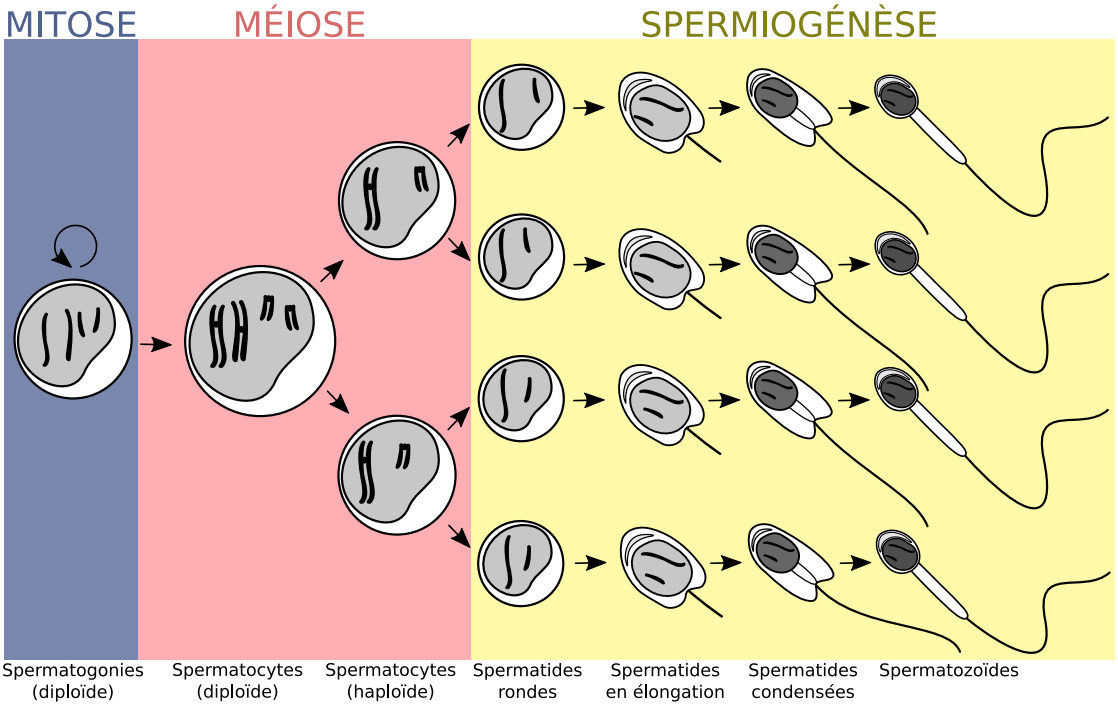
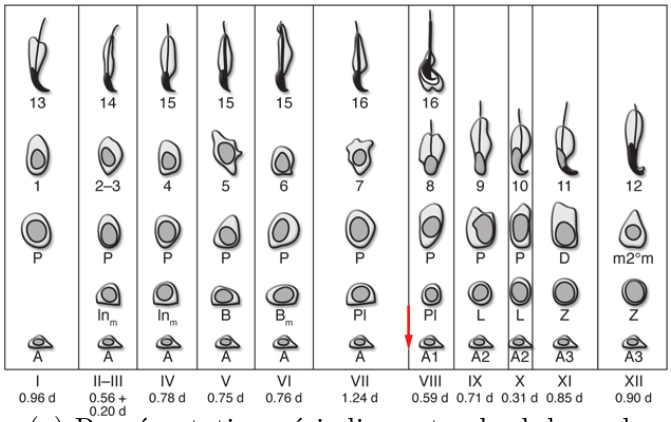


FIGURE 1.30 – Vue d'ensemble de la spermatogénèse.

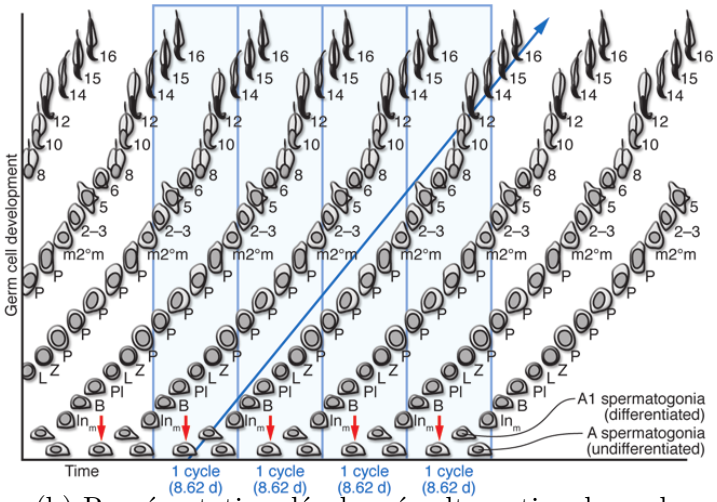
(FIGURE 1.31a). Au cours de sa différentiation, chaque gamète sera présent dans quatres itérations du cycle avant de quitter son tube séminifère (FIGURE 1.31b). Le déclenchement de la différentiation n'est pas synchrone pour l'ensemble du testicule et s'effectue plutôt séquentiellement le long des tubes séminifères. Les stades successifs sont répartis le long d'un tube séminifère (FIGURE 1.31c), ce qui assure une production continue de spermatozoïdes.

Phase méiotique La phase méiotique correspond à une recombinaison du matériel génétique suivie de deux divisions nucléaires successives permettant la production de quatre cellules haploïdes (FIGURE 1.32). Cette phase débute par une interphase durant laquelle chaque cellule germinale va dupliquer son contenu en ADN. S'ensuit les deux divisions méiotiques, constituées chacune d'une prophase, métaphase, anaphase et télophase. Les étapes de la première et deuxième division sont notées respectivement I et II.

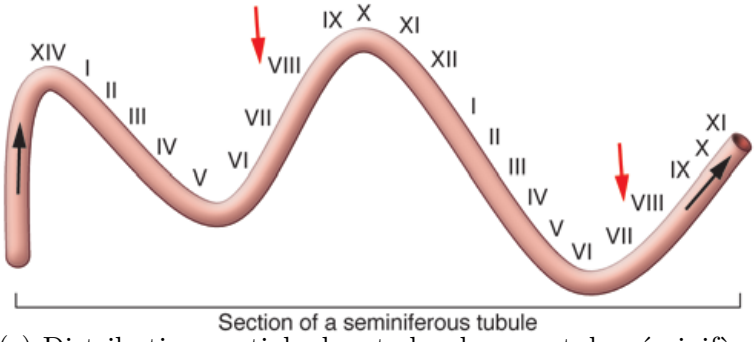
La prophase I est une étape longue, subdivisée en cinq stades successifs : leptotène, zygotène, pachytène, diplotène et diacinèse. Les spermatocytes primaires au stade leptotène condensent leurs chromosomes dupliqués lors de l'interphase. Au stade zygotène, les chromosomes homologues vont se rapprocher grâce à la formation des complexes synaptonémaux. Une fois appariés, des échanges de fragments chromosomiques s'effectuent par recombinaison homologue avec la formation de chiasmata au stade pachytène. Les complexes synaptonémaux disparaissent au stade diplotène entrainant la séparation des chromosomes homologues. La membrane nucléaire disparait en diacinèse.



(a) Représentation périodique standard du cycle



(b) Représentation développée alternative du cycle



(c) Distribution spatiale des stades dans un tube séminifère

FIGURE 1.31 – Cycle de l'épithélium séminifère. Les douzes stades sont notés de I à XII. Les flèches rouges indiquent l'engagement dans la phase méiotique. 4 cycles de 8,62 jours sont nécessaires pour passer d'une spermatogonie initiant la méiose à un spermatozoïde quittant les tubes séminifères. A : spermatogonie de type A indifférenciée ; A1 : spermatogonie de type A différenciée ; In_m : spermatogonie intermédiaire ; B : spermatogonie de type B ; Pl : spermatocyte préleptotène ; L : spermatocyte leptotène ; Z : spermatocyte zygotène ; P : spermatocyte pachytène ; D : spermatocyte diplotène ; m2°m : spermatocyte secondaire. 1 à 7 : spermatide ronde ; 8 à 11 : spermatide allongée ; 12 à 16 : spermatide condensée. Adapté de HOGARTH et GRISWOLD 2010.

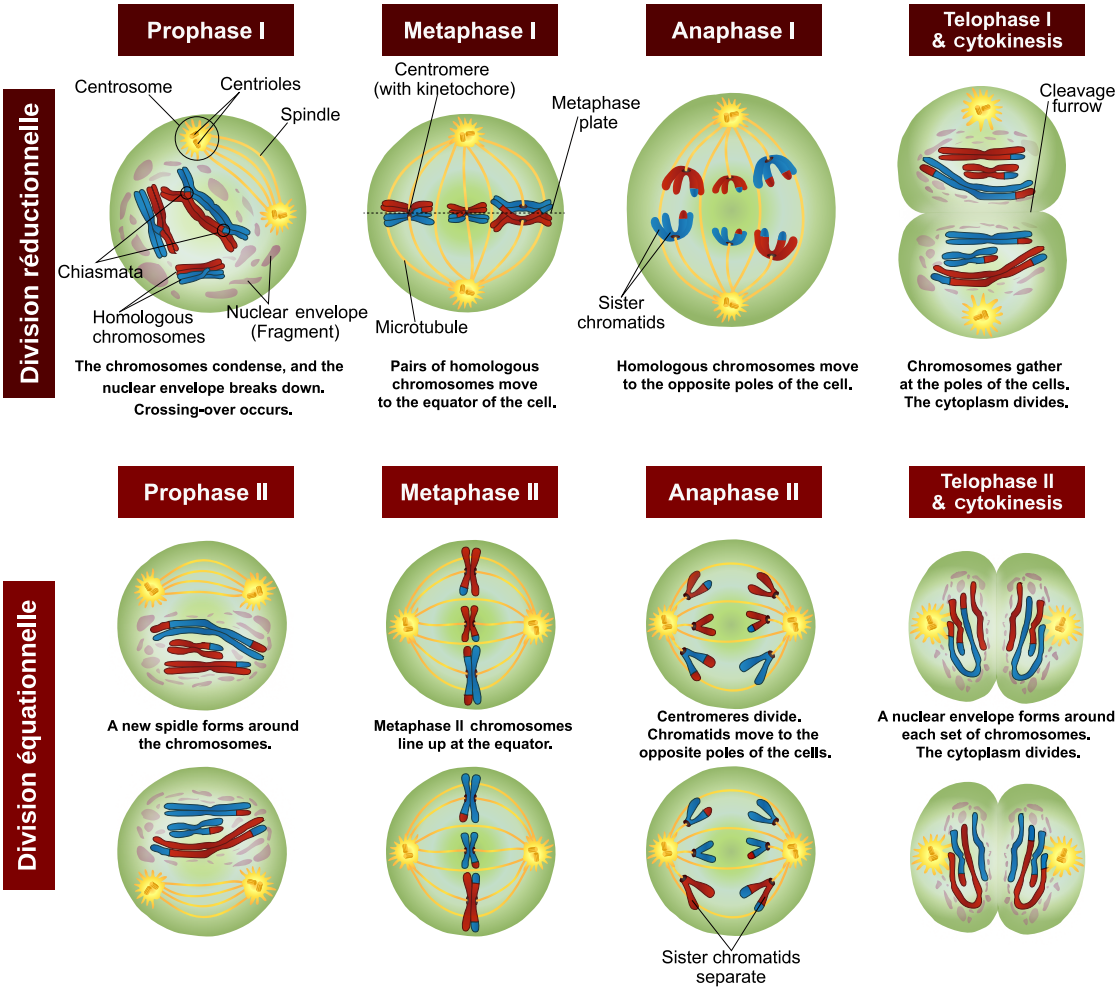


FIGURE 1.32 – Déroulement de la méiose. Adapté de © Ali Zifan / CC BY-SA 4.0.

Les chromosomes s'alignent sur la plaque équatoriale durant la métaphase I. Puis, les chromosomes homologues migrent chacun vers un pôle lors de l'anaphase I. Lorsqu'ils sont enfin rassemblés aux pôles en télophase I, le cytoplasme se scinde pour former deux spermatocytes secondaires haploïdes. Cette première division est dite réductionnelle puisqu'elle divise par deux le nombre de chromosomes par cellule.

La deuxième division méiotique se réalise sans duplication de l'ADN préalable. Cette division est dite équationnelle car le nombre de chromosmes reste inchangé. À la place, ce sont les chromatides sœurs de chaque chromosome qui sont séparées. Cette deuxième division donne naissance à de petites cellules rondes, les spermatides.

Spermiogénèse La phase post-méiotique correspond à la différentiation des spermatides rondes en spermatozoïdes. On peut distinger deux stades intermédiaires majeurs, les spermatides en élongations et les spermatides condensées, à partir de trois caractères phénotypiques principaux.

Premièrement, les spermatides vont développer un acrosome formé initialement au niveau du pôle antérieur du noyau grâce à la fusion de plusieurs vésicules issues de l'appareil de Golgi. Il s'étend et se densifie jusqu'au pôle postérieur du noyau des spermatides condensées.

Deuxièmement, les spermatides se dotent d'un flagelle permettant d'assurer la motilité du spermatozoïde dans le tractus génital femelle. La formation de ce flagelle est initiée par une migration des centrioles des spermatides rondes au niveau du pôle postérieur du noyau. Ces centrioles guident la formation des micrutubules structurant le flagelle. Un grand nombre de mitochondries se réorganisent au niveau de la région proximale du flagelle pour fournir l'énergie nécessaire au mouvement du flagelle. La maturation du flagelle est accompagnée d'une élimination de l'excès de cytoplasme, évacué sous forme de corps résiduels dans les spermatides condensées tardives.

Troisièmement, le noyau des spermatides s'allonge et la chromatine transitionne d'une structure nucléosomale de type somatique vers une structure compacte composées principalement de nucléoprotamines (FIGURE 1.33).

Après avoir été relargués dans la lumière des tubes séminifères, les spermatozoïdes vont finaliser leur maturation lors de leur progression dans l'épididyme où ils vont développer leur motilité et leur pouvoir fécondant. La chromatine des spermatozoïdes continue à se compacter, le cytoplasme diminue encore de volume et la structure de la membrane plasmique change.

1.2.2.4 Dynamiques épigénétiques au cours de la spermatogénèse

Le spermatozoïde étant la plus petite cellule nucléée chez la plupart des espèces, sa fonction de vecteur de génome, bien qu'haploïde, est rendu possible par un compactage extrême de la chromatine dans son noyau. Les mécanismes nécessaires pour réaliser cette réduction drastique du volume du noyau impliquent de nombreuses dynamiques épigénétiques à l'échelle du génome : vagues de modifications post-traductionnelles des histones, incorporation de variants d'histones, de protéines de transitions puis de protamines (FIGURE 1.34).

Deux vagues globales d'acétylations ont été mises en évidence au cours de la spermatogénèse. Les histones H2A, H2B et H4 sont globalement acétylées dans les spermatogonies et les spermatocytes pré-leptotène. Les histones sont ensuite déacétylées lors de la méiose, avant d'être réacétylées lors de la phase d'élongation des spermatides (HAZZOURI, PIVOT-PAJOT, FAURE et al. 2000). Cette deuxième vague d'acétylation est suspectée d'être un prérequis au retrait des nucléosomes médié par Brdt, une facteur à double bromodomaine spécifiquement exprimé durant la spermatogénèse et reconnaissant la chromatine acétylée (GAUCHER, BOUSSOUAR, MONTELLIER et al. 2012). Plus récemment, il a été mis en évidence que la deuxième vague d'acétylation est en réalité une vague d'acylation, avec au moins la crotonylation (TAN, LUO, LEE et al. 2011) et la butyrylation (GOUDARZI, ZHANG, HUANG et al. 2016) comme modifications d'histones détectées. Ces marques présentes en compétition sur les lysines de l'histone 4 sont

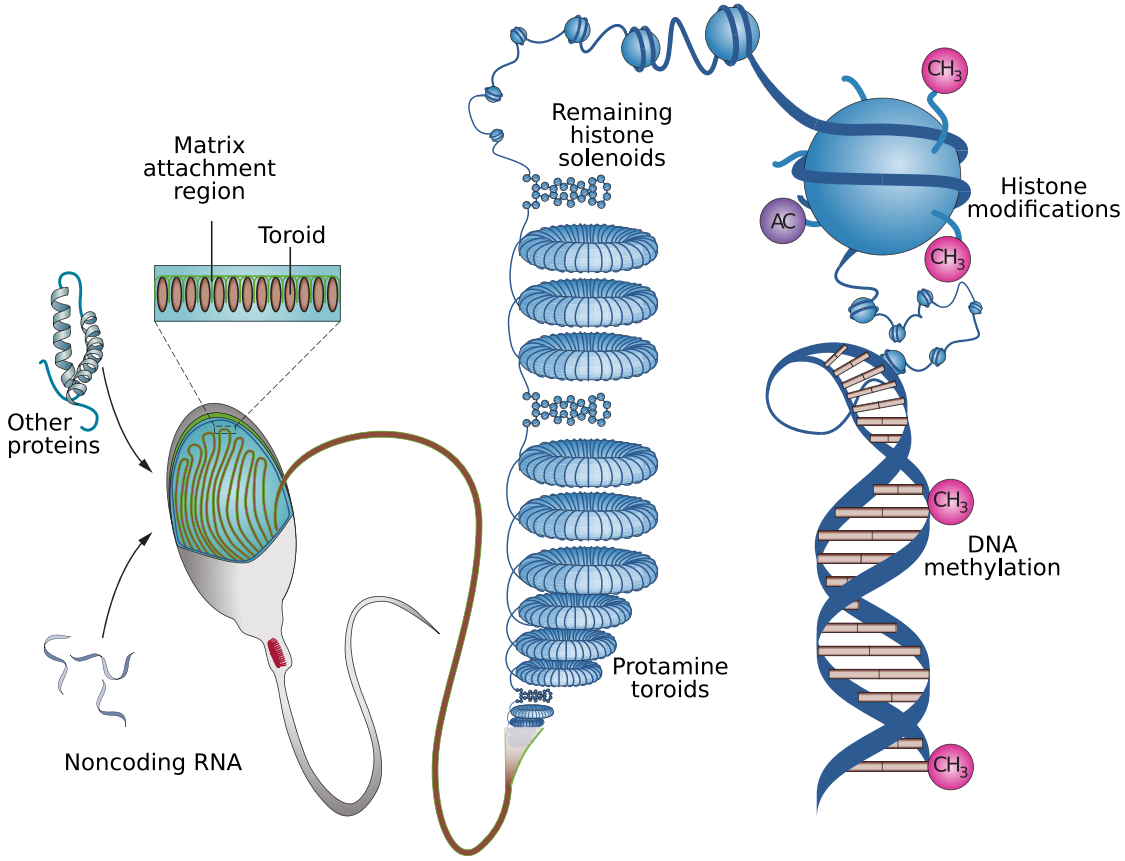


FIGURE 1.33 – Vue d'ensemble de la chromatine du spermatozoïde. Environ 85% des histones sont remplacées par des protamines au cours de la spermiogénèse. La chromatine du spermatozoïde est constituée principalement de nucléoprotamines enroulées en toroïdes attachés par endroits à la matrice nucléaire. Les nucléosomes restants sont confinés dans des régions enroulées en solénoïdes. Bien que le spermatozoïde soit une cellule transcriptionnellement inactive, les histones restantes sont acétylées, et exhibent aussi d'autres marques comme la méthylation. Adapté de SCHAGDARSURENGIN, PARADOWSKA et STEGER 2012.

enrichies dans les promoteurs et amplificateurs associés aux gènes actifs post-méiotiquement.

Les méthylations des lysines de l'histone 3 présentent des dynamiques spécifiques à chacune d'elles (FIGURE 1.34). Des expériences d'inactivation des méthyltransférases (Meisetz pour H3K4 ; HAYASHI et MATSUI 2006 ; Suv39h1 et Suv39h2 pour H3K9 ; PETERS, O'CARROLL, SCHERTHAN et al. 2001) et déméthyltransférases (JHDM2A pour H3K9me3 ; (OKADA, SCOTT, RAY et al. 2007)) responsables de certaines d'entres elles entraînent la génération de mâles stériles, ce qui souligne l'importance de ces dynamiques pour la programmation correcte du génome mâle. La connaissance de l'implication de ces marques dans des mécanismes de régulation de l'expression des gènes dans un contexte somatique (H3K4me associée aux promoteurs et amplificateurs actifs, H3K9me associée à l'hétérochromatine, H3K27me3 associée à la répression médiée par Polycomb) laisse entrevoir comment ces modifications agissent seules ou en combinaison pour moduler l'état fonctionnel de la chromatine.

Une dynamique de phosphorylation des histones, suivies d'ubiquitination et de sumoylation au cours de la méiose sont associées à l'incorporation de variants d'histones (BAARENDS, HOOGERBRUGGE, ROEST et al. 1999 ; VIGODNER et MORRIS 2005) et à la formation du *sex body*, une structure nucléaire contenant l'hétérochromatine des chromosomes X et Y. Parmi les variants incorporés au cours de la méiose, il est possible de citer H3T, TH2A, H2A.X, H2A.Z, TH2B, H3T et H3.3B. La forme ubiquitinylée de H2A disparaît en spermatide ronde pour réapparaître lors de l'élongation des spermatides, ce qui pourrait coïncider avec la deuxième phase d'intégration de variants d'histones : H1T2, HILS1, H2AL1/2, H2BL1.

Les protéines de transitions sont incorporées à la chromatine presque conjointement à ces variants tardifs. Ces protéines appartiennent à une famille de quatre membres identifiés chez les mammifères : TP1 à TP4. Les protéines TP1 et TP2 représentent respectivement 50% et 40% des protéines nucléaires des spermatides condensées. Les mécanismes par lesquels ces protéines de transition sont incorporées sont mal connus. En leur absence, l'incorporation des protamines a lieu mais semble moins structurée (BARRAL, MOROZUMI, HOGHOUGHI et al. 2017). De plus, contrairement à la protamine 1 directement traduite à partir de son ARN messager, la protamine 2 a besoin d'une série de six clivages protéolytiques après son incorporation dans la chromatine (CHAUVIÈRE, MARTINAGE, DEBARLE et al. 1992). En l'absence de protéines de transition, cette maturation n'a pas lieu (BARRAL, MOROZUMI, HOGHOUGHI et al. 2017).

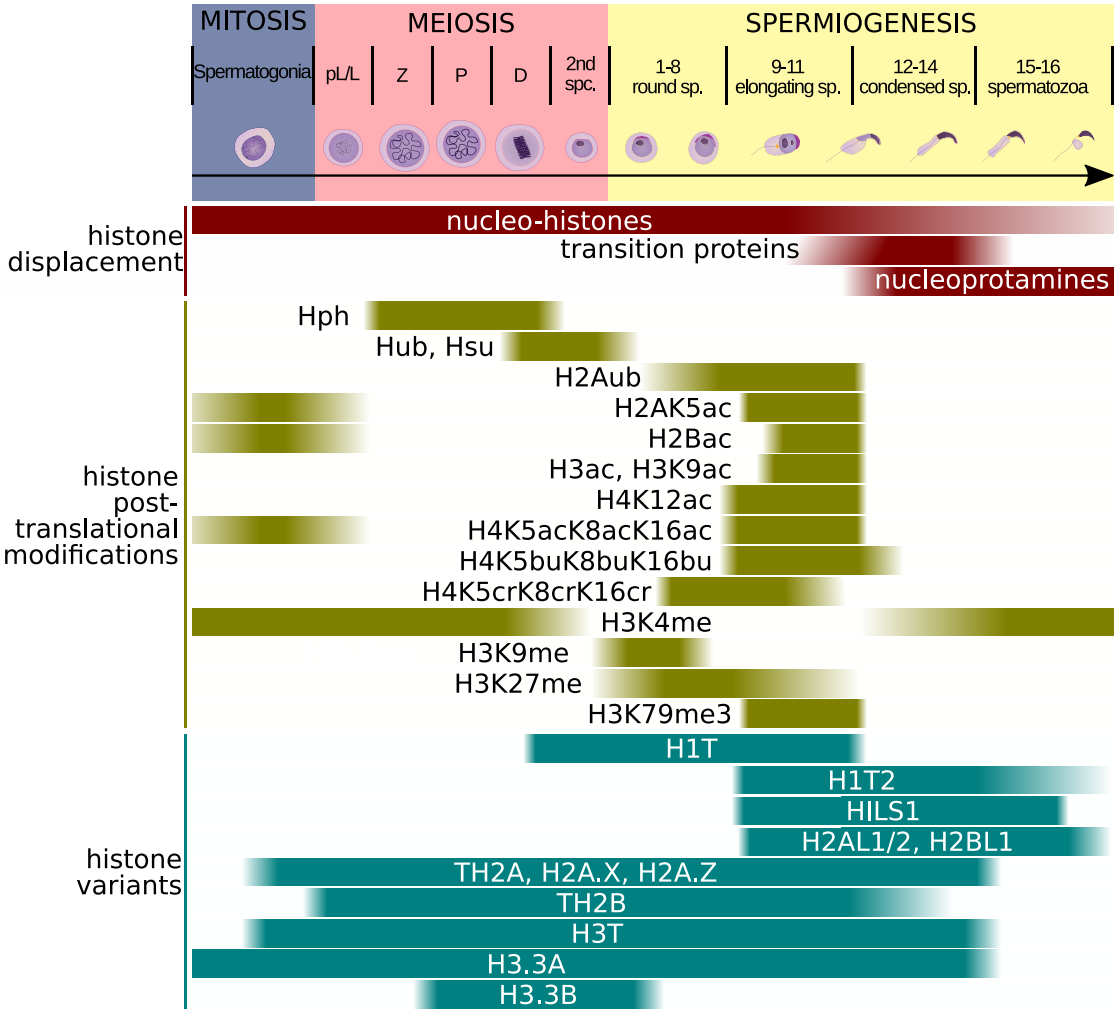


FIGURE 1.34 – Vue d'ensemble des dynamiques épigénétiques au cours de la spermatogénèse murine. Créé à partir de sources multiples. H2AK5ac, H2Bac, H3ac, H4K12ac, H4K5acK8acK16ac à partir de HAZZOURI, PIVOT-PAJOT, FAURE et al. 2000. H4K5crK8crK16cr à partir de TAN, LUO, LEE et al. 2011. H4K5KbuK8buK16bu à partir de GOUDARZI, ZHANG, HUANG et al. 2016. nucleo-histones, transitions proteins, nucleoprotamines, H3K79me3, H3K9ac, H1T, H1T2, HILS1, H2AL1/2, TH2A, TH2B, H3T, H3.3A, H3.3B à partir de BAO et BEDFORD 2016. H2Aub à partir de BAARENDS, HOOGERBRUGGE, ROEST et al. 1999. Hph, Hub, Hsu à partir de DING, LIU, LIU et al. 2015. H3K4me, H3K9me, H3K27me à partir de GE, LIN, ZHAO et al. 2017.

1.3 Contribution du séquençage haut débit à la génomique

Le séquençage à haut débit, et son pendant analytique la bioinformatique des séquences, n'est pas l'unique approche pour étudier le génome. Précédés par de nombreuses approches de biologie moléculaire (immunomarquage, électrophorèse, biopuce, *etc.*) et de physique analytique (spectroscopie de masse, imagerie, *etc.*), ils s'inscrivent en complémentarité en proposant une vision alternative du génome. Les données issues du séquençage à haut débit constituent le matériau de travail du bioinformaticien des séquences. Cette section présente comment ces données sont générées, analysées et partagées.

1.3.1 Techniques de séquençage haut débit

La réduction du coût de séquençage est un enjeu crucial crucial de ces dernières décennies pour le développement de nouvelles applications en médecine (RABBANI, NAKAOKA, AKHONDZADEH et al. 2016), biotechnologie et recherche fondamentale. Cet enjeu a fortement stimulé le développement de techniques originales et toujours plus efficientes. Ainsi se sont succédées les approches de séquençage par terminaison de chaîne (Sanger), pyroséquençage (454), ion semiconducteur (Ion Torrent), ligation (SOLiD) que l'on peut considérer aujourd'hui comme obsolètes. Le séquençage par synthèse (Illumina) qui s'est répandu de façon prédominante pourrait être dans les années à venir être dépassé par, ou au moins coexister avec, de nouvelles technologies aux propriétées analytiques intéressantes pour certaines problématiques (FIGURE 1.35). On peut ainsi penser aux technologies "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, SMITH, COUPLAND et al. 2012). Les données exploitées dans ce manuscrit proviennent principalement de séquenceurs Illumina.

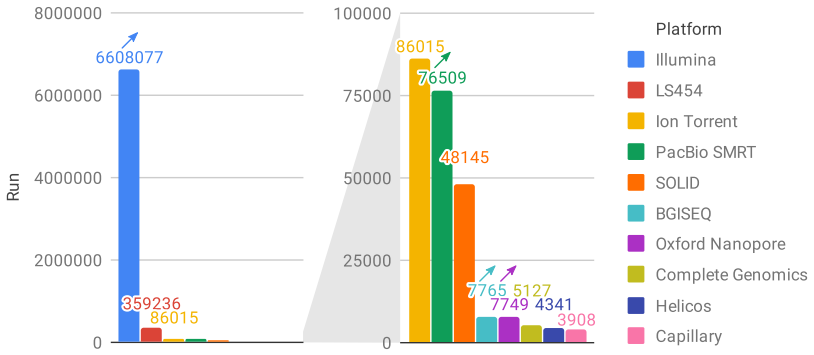


FIGURE 1.35 – Nombre de runs de séquençage disponibles dans l'ENA par plateforme de séquençage. Les flèches indiquent les technologies activement utilisées de nos jours. Requêtes réalisées le 4 juillet 2019.

1.3.2 Méthodes d'analyse du génome basées sur le séquençage haut débit

Au delà des méthodes d'exploitation directe des séquences génomiques issues des organismes permettant notamment l'assemblage de génome, la détection de variants et de réarrangements chromosomiques ainsi que l'étude de la diversité génétique de populations (Davey et Blaxter 2010), se sont développées de nombreuses techniques explorant diverses caractéristiques du génome. 265 méthodes sont ainsi référencées en Annexe A. Le principe de celles dont j'ai été amené à analyser des données est abordé ci-dessous.

1.3.2.1 RNA-seq et analyse du transcriptome

Le RNA-seq est une technique de quantification du niveau d'expression des gènes en terme de transcrits ARN (Morin, Bainbridge, Fejes et al. 2008). C'est la technique basée sur le séquençage à haut débit la plus répandue, avec plus de 17 109 articles la référençant dans Pubmed, loin devant la deuxième technique la plus référencée, le ChIP-Seq, avec 3 833 articles (Figure 1.36 ; Données consultées le 10 mai 2019).

Cette technique repose sur des étapes successives d'extraction de l'ARN des cellules, de sélection pour des familles d'ARN d'intérêt, de fragmentation et de rétrotranscription en ADN avant séquençage à haut débit (Figure 1.37). Les ARN ribosomiques étant très majoritaires, la sélection est nécessaire pour pouvoir quantifier avec efficience les autres types d'ARN. Si l'intérêt est porté exclusivement sur les transcrits matures messagers, la sélection peut être réalisé par sélection basée sur leur queue polyadénylée. Sinon, un enrichissement relatif pour les ARN non ribosomiques peut être obtenu par ribo-déplétion. L'étape de rétrotranscription, nécessaire pour un séquençage avec une technologie par

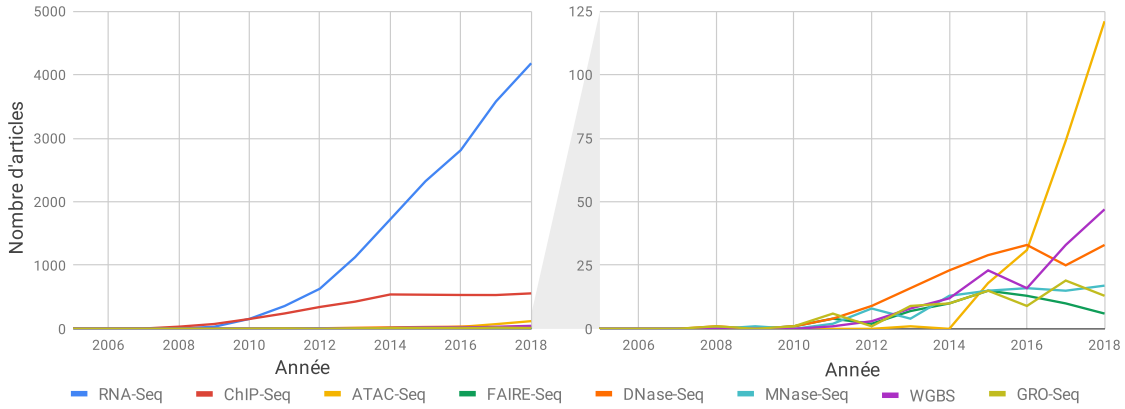


Figure 1.36 – Tendances Pubmed des principales techniques basées sur le séquençage à haut débit. Les valeurs correspondent au nombre d'articles contenant un mot clé donné par année. Tendances extraites à partir de Medline.

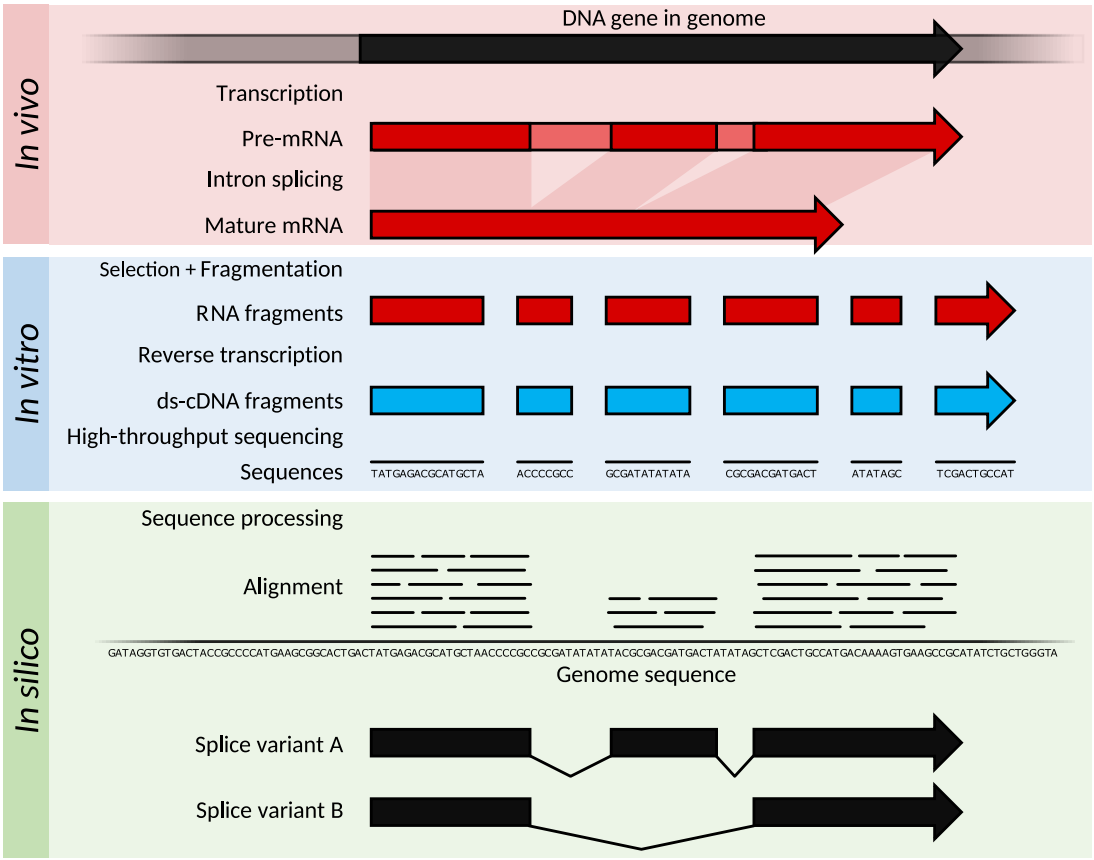


FIGURE 1.37 – Principe du RNA-Seq. © Thomas Shafee / CC BY-SA 4.0

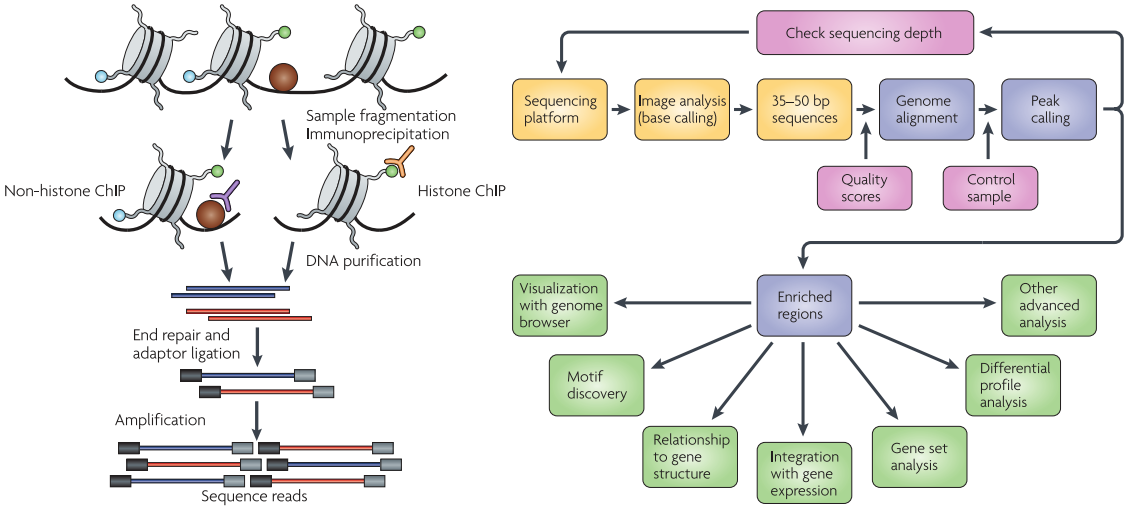
synthèse (Illumina), n'est plus nécessaire avec celle par nanopore (DEPLEDGE, SRINIVAS, SADAOKA et al. 2019).

La quantification des transcrits peut se faire après alignement sur le génome (DOBIN, DAVIS, SCHLESINGER et al. 2013) ou sur des transcrits de référence (BRAY, PIMENTEL, MELSTED et al. 2016). Le RNA-seq permet de détecter de nouveaux gènes, des transcrits alternatifs et des gènes de fusion (VU, DENG, TRAC et al. 2018 ; HEYER, DEVESON, WOOI et al. 2019).

1.3.2.2 ChIP-seq et analyse des interactions protéines-ADN

Le ChIP-seq (*Chromatin ImmunoPrecipitation followed by sequencing*) est une technique d'analyse des interactions entre les protéines et l'ADN (BARSKI, CUDDAPAH, CUI et al. 2007). Elle permet de détecter et d'identifier les séquences dADN 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 aux modifications d'histone.

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. C'est



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

FIGURE 1.38 – Principe du ChIP-seq. Adapté de PARK 2009.

une adaptation de l'immunoprécipitation de la chromatine sur biopuce (ChIP-on-chip ; APARICIO, GEISBERG et STRUHL 2004) à la technologie du séquençage à haut débit. Elle apporte une couverture du génome plus complète et une plus haute résolution dans le positionnement des sites de fixation identifiés (BARSKI, CUDDAPAH, CUI 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 immunoprécipitée par un anticorps spécifique de la protéine d'intérêt (FIGURE 1.38a). Une meilleure résolution peut être atteinte si l'étape de fragmentation initialement basée sur le principe de sonication de l'ADN est remplacée ou complémentée par une digestion enzymatique : MNase ou lambda exo-nucléase (RHEE et 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 *peak-calling*. Un échantillon contrôle, nommé *input*, contenant de la chromatine fragmentée mais non immunoprécipitée sert de référence pour discriminer les enrichissements locaux de la chromatine des artefacts, principalement liés au protocole de fragmentation (MEYER et LIU 2014), d'amplification (BENJAMINI et SPEED 2012) et à l'incomplétude des génomes de références (MIGA, EISENHART et KENT 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.* (FIGURE 1.38b).

1.3.2.3 MNase-seq et analyse du paysage nucléosomal

Le MNase-seq est une technique d'isolation des structures protégeant l'ADN contre la dégradation par une endo-exonucléase, la nucléase micrococcale. Cette technique permet principalement d'isoler les fragments d'ADN de 147 bp enroulés autour des mononucléosomes, mais peut aussi permettre d'observer, de façon artefactuelle ou souhaitée, différents polymères d'histones ainsi que d'autres structures bloquant la progression de la nucléase tels que les G-quadruplex. La proportion obtenue de chacun de ces éléments, ainsi que leur provenance dans le génome en terme de régions plus ou moins ouvertes de la chromatine, dépend de la concentration en nucléase et de la durée de digestion. Cette durée est contrôlé par ajout d'EDTA qui stoppe l'activité enzymatique. Les différentes fractions chromatiniennes peuvent être séparées par gradient de sucrose et ultracentrifugation. Les protéines sont ensuite éliminées des fractions d'intérêt par ajout de protéinase K (Barral, Morozumi, Tanaka et al. 2017). Le séquençage et l'analyse des données est similaire à ceux pour le ChIP-Seq à quelques exceptions près : d'une part l'étape de détection de pics requiert des algorithmes spécifiques de par l'organisation rapprochée régulière des nucléosomes (Figure 1.39 ; Chen, Xi, Pan et al. 2013 ; Polishko, Bunnik, Le Roch et al. 2014)

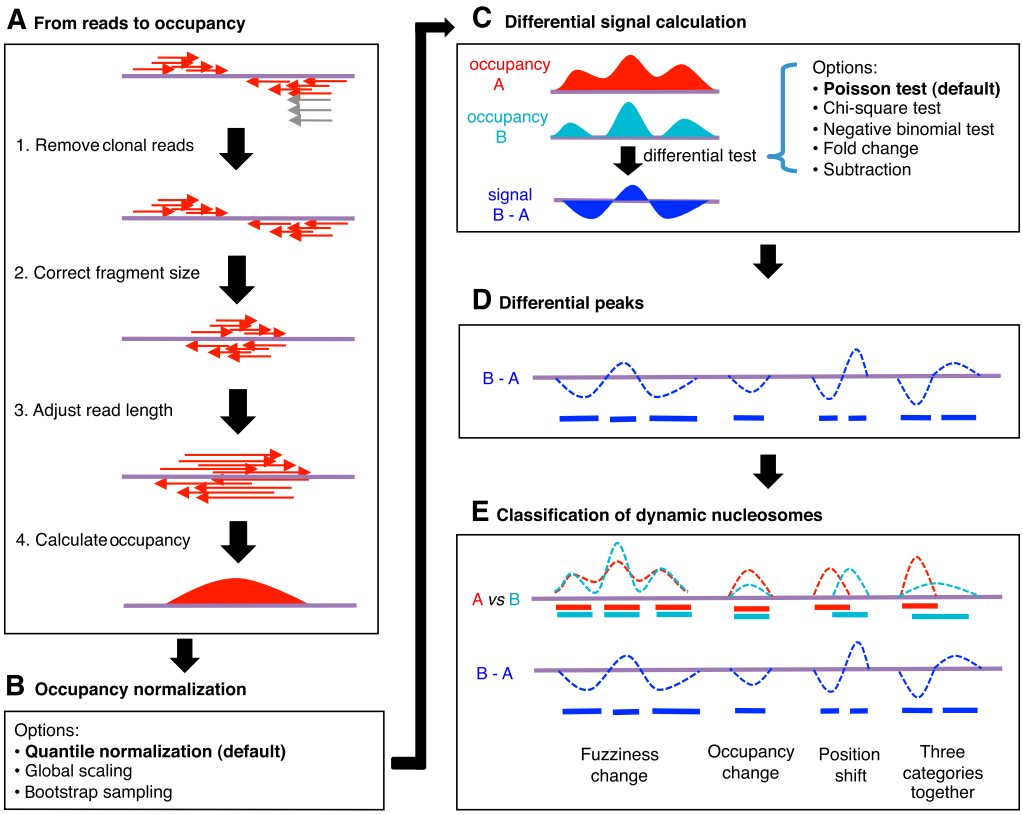
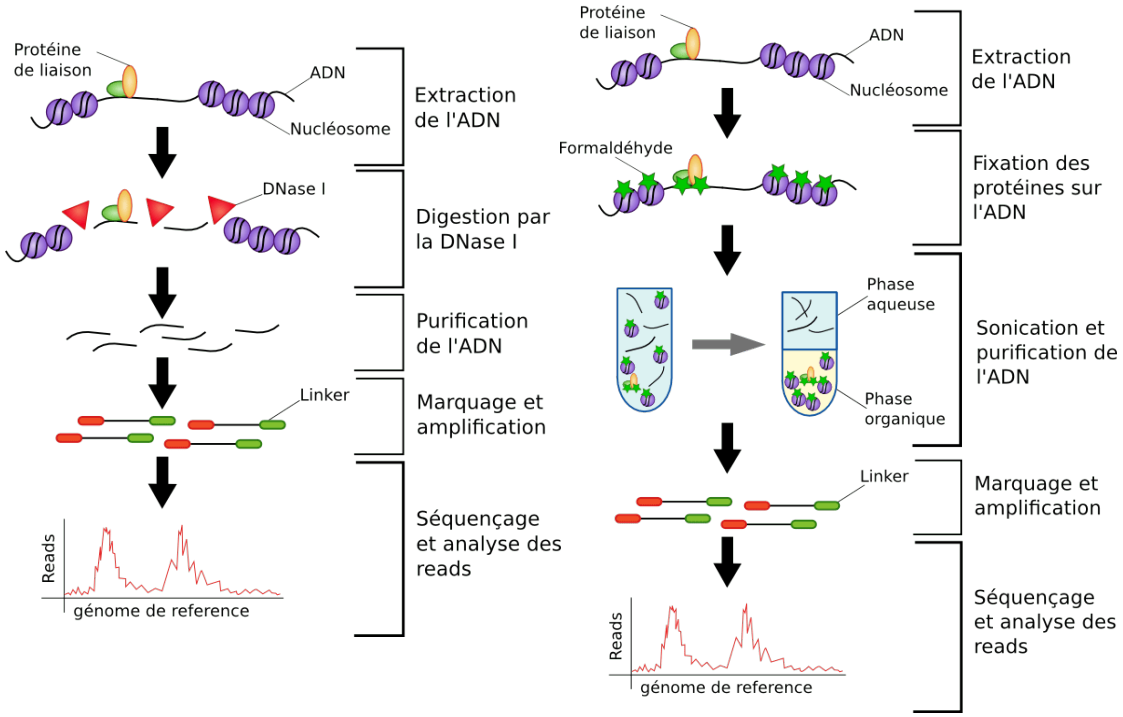


Figure 1.39 – Pipeline d'analyse standard de la dynamique nucléosomale entre conditions à partir de données MNase-Seq. Extrait de Chen, Xi, Pan et al. 2013.

1.3.2.4 DNase-seq, FAIRE-seq, ATAC-seq et analyse de l'accessibilité chromatinienne

Le DNase-seq (Boyle, Davis, Shulha et al. 2008), le *Formaldehyde-Assisted Isolation of Regulatory Elements followed by sequencing* ou FAIRE-seq (Giresi, Kim, McDaniell et al. 2007) et l'*Assay for Transposase-Accessible Chromatin followed by sequencing* ou ATAC-seq (Buenrostro, Giresi, Zaba et al. 2013) sont trois techniques appliquées pour l'étude de l'accessiblité 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 1.40a). Le FAIRE-seq repose sur le principe que la chromatine fixée par du formaldéhyde peut être extraite de la chromatine composée d'ADN nu après fragmentation (Figure 1.40b). 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 1.41a).

Les protocoles expérimentaux des trois techniques diffèrent à l'avantage de la technique la plus récente, plus rapide et moins exigeante en quantité de matériel biologique (Figure 1.41b). Les étapes d'analyse des données restent similaires à celles du ChIP-seq et les profils obtenus par les trois approches demeurent sensiblement identiques (Figure 1.41b). L'ATAC-seq ayant tendance à isoler éga-



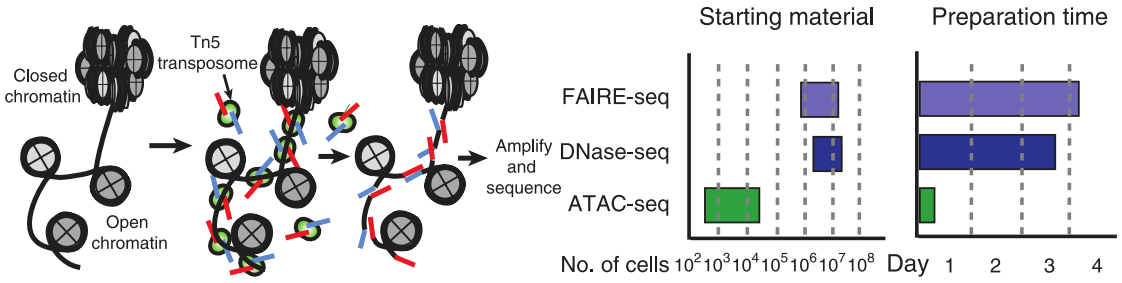
(a) Principe du DNase-seq (b) Principe du FAIRE-seq

Figure 1.40 – Principe du DNase-seq et du FAIRE-seq.

lement les fragments nucléosomaux des régions ouvertes, il peut être utile de filtrer les fragments de taille voisine à celle d'un nucléosome afin d'obtenir le paysage nucléosomal local en alternative de données MNase-seq (FIGURE 1.42 ; BUENROSTRO, GIRESI, ZABA et al. 2013).

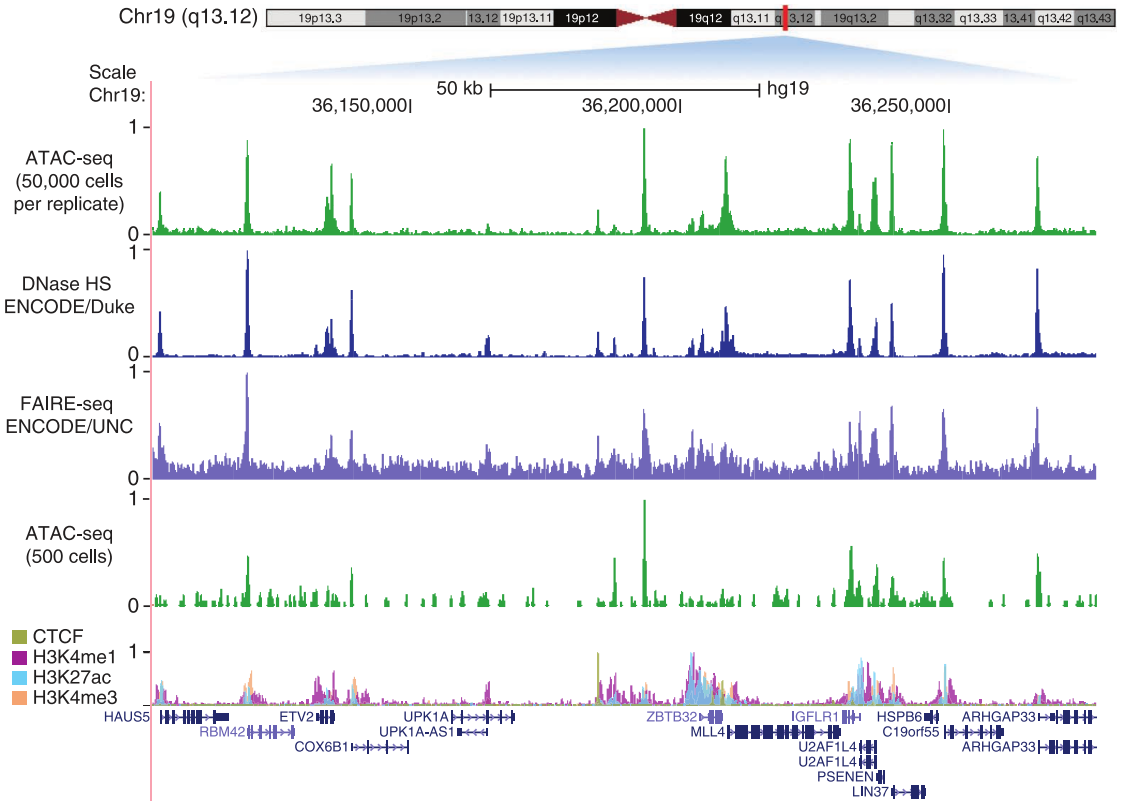
1.3.2.5 Bisulfite-seq et analyse du méthylome

Le bisulfite-seq est une approche basée sur la conversion différentielle des cytosines en fonction de leur niveau de méthylation par un traitement chimique au bisulfite (FROMMER, MCDONALD, MILLAR et al. 1992). Les cytosines non méthylées sont converties en uraciles, elles mêmes amplifiées en thymines avant séquençage à haut débit. Les cytosines méthylées ne sont pas converties par ce traitement et sont détectées comme cytosines lors du séquençage. La comparaison des séquences génomiques traitées et non traitées par le bisulfite permet d'inférer si une thymine détectée correspond à une vraie thymine ou à une cytosine non méthylée (FIGURE 1.43). Une limitation de l'approche réside dans l'impossibilité de discriminer des cytosines méthylées de celles hydroxyméthylées (HUANG, PASTOR, SHEN et al. 2010), alors que leurs propriétés épigénétiques diffèrent (SECTION 1.1.3.1). Des variations du protocole original, tel que celui du *Tet-Assisted Bisulfite Sequencing* (YU, HAN, HON et al. 2018) ou l'*Oxidative Bisulfite Sequencing* (BOOTH, OST, BERALDI et al. 2013) entraînent des combinaisons de conversions nucléotidiques différentes, ce qui peut permettre de discriminer les différentes modifications des cytosines.



(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 1.41 – Principe de l'ATAC-seq et ses améliorations par rapport au DNase-seq et FAIRE-seq. Extrait de BUENROSTRO, GIRESI, ZABA et al. 2013.

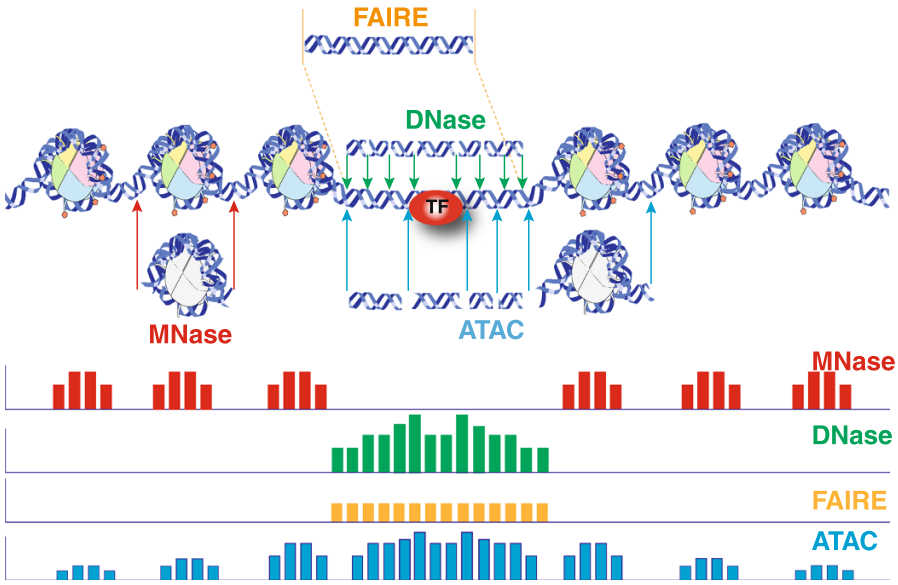


FIGURE 1.42 – Comparatif schématique des résultats obtenus par les différentes techniques d'analyse de l'accessibilité de la chromatine. Extrait de TSOMPANA, BUCK, LUGER et al. 2014.

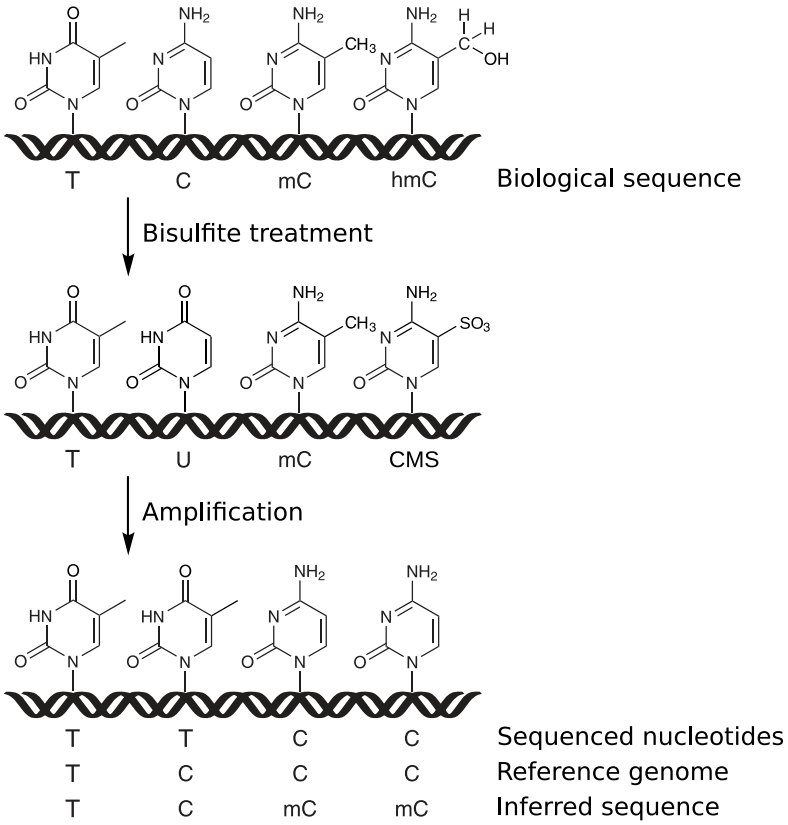


FIGURE 1.43 – Principe du bisulfite-seq. La connaissance du génome de référence de l'individu traité permet de déduire après séquençage une thymine d'une cytosine non méthylée. Il n'est pas possible de discriminer une cytosine méthylée d'une cytosine hydroxyméthylée par bisulfite-seq.

1.3.3 Consortiums et grands projets pour la génomique

À l'achèvement du Projet Génome Humain, différents consortiums et grands projets internationaux ont vu le jour afin de compléter notre compréhension de la génomique. Au delà des projets visant à étendre le nombre d'espèces dont le génome est séquencé (MISRA, LANGEFELD, OLIVIER et al. 2018), ainsi que ceux visant à cataloguer les variations génomiques au sein des populations (AUTON, ABECASIS, ALTSHULER et al. 2015 ; SUDMANT, RAUSCH, GARDNER et al. 2015 ; TURNBULL, SCOTT, THOMAS et al. 2018), se sont développées différentes initiatives pour structurer l'ensemble des données générées par la communauté scientifique. Cette section présente les différents projets qui ont rendu possible les analyses bioinformatiques réalisées durant cette thèse.

1.3.3.1 Archivage de séquences

L'*International Nucleotide Sequence Database Collaboration* (INSDC) est une initiative entre les principales institutions de biologie moléculaire

— européenne, le *European Bioinformatics Institute* (EBI),

— états-unienne, le *National Center for Biotechnology Information* (NCBI) et

— japonaise, le *DNA Data Bank of Japan* (DDBJ)

afin de constituer une archive de l'ensemble des données brutes et métadonnées issues d'expériences basées sur le séquençage à haut débit. En pratique, chaque institut héberge une version de l'archive synchronisée commune et propose ses propres interfaces pour accéder à son contenu. Le NCBI propose la Sequence Read Archive (SRA) dont les fichiers peuvent être récupérés grâce à l'outil en ligne de commande SRA toolkit en plus de l'accès manuel par l'interface web. Les équivalents de l'EBI et du DDBJ sont respectivement l'*European Nucleotide Archive* (ENA) et la *DDBJ Sequence Read Archive* (DRA). Tout effort de recherche scientifique sérieuse se doit de mettre à disposition de la communauté les données brutes de résultats publiés en les téléversant auprès de l'un des instituts qui fournit en retour des identifiants uniques à renseigner dans les publications. Malheureusement, cela est impossible pour certaines données humaines dont l'utilisation a été consentie uniquement pour des applications spécifiques et non pour un accès libre. L'*European Genome-phenome Archive* (EGA) est une alternative pour ce cas de figure (LAPPALAINEN, ALMEIDA-KING, KUMANDURI et al. 2015).

1.3.3.2 Assemblage de génomes

Les ébauches du génome humain ont été publiées en février 2001. Le séquençage complet du génome a été annoncé en 2004 à la fin du projet Génome Humain. Au delà de l'impact sociétal de cette annonce se cache une réalité technique différente : certaines régions, principalement télomériques et centromériques, sont absentes du génome de référence le plus récent. Du fait de leur constitution riche en régions répétées, les technologies de séquençage classiques ne permettent pas de reconstituer les contigs d'où proviennent les fragments

séquencés. Grâce aux nouvelles approches *long reads*, le séquençage et l'assemblage de ces régions semblent désormais à portée (LARSEN, HARRIS, LIU et al. 2017 ; AMEUR, CHE, MARTIN et al. 2018). De plus, les génomes de référence actuels ne représentent pas nécessairement les allèles et haplotypes les plus communs, ni les plus longs. Le *Genome Reference Consortium* GRC, constitué de l'EBI, du NCBI, du *Wellcome Sanger Institute*, du *McDonnel Genome Institute* et du *Zebrafish Information Network* a pour rôle de coordonner les efforts de correction des génomes de références actuels de l'homme (GRCh38), de la souris (GRCm38), du poulet (GRCg6a) et du poisson-zèbre (GRCz11). Des patchs sont ainsi proposés régulièrement corrigeant les erreurs d'assemblage et ajoutant des haplotypes alternatifs. Les corrections de type insertions ou délétions ne peuvent, sauf cas particuliers, être corrigées sans changer les coordonnées d'un génome de référence, ce qui implique un changement de génome de référence. L'itération suivante du génome murin (GRCm39) prévue pour l'été 2019 pourrait potentiellement relancer un projet d'étude des régions péricentromériques mené au cours de cette thèse et qui n'avait jusque là donné aucun résultat intéressant (SECTION 2.2.2).

1.3.3.3 Cartographie des gènes et transcrits

Plusieurs grands projets ont été menés afin de produire des annotations les plus exhaustives possibles sur la position des gènes, exons et introns dans le génome. Initié dès 2000, *Functional ANnoTation Of the Mammalian genome* (FANTOM) a développé différentes techniques focalisées sur l'étude du transcriptome. Celles-ci ont permis d'apporter la première estimation du nombre de gènes dans le génome (FANTOM CONSORTIUM 2001). Plus récemment, le développement et l'application à large échelle de la technique *Cap Analysis Gene Expression* (CAGE) consistant à analyser uniquement l'extrémité 5' des ARN a permis de produire la plus grande collection de promoteurs et TSSs chez l'homme et la souris (FORREST, KAWAJI, REHLI et al. 2014) et d'enhancers transcrits chez l'humain (ANDERSSON, GEBHARD, MIGUEL-ESCALADA et al. 2014).

En continuation du projet Génome Humain a été lancé un vaste projet de création d'une encyclopédie des éléments du génome, ENCODE (DUNHAM, KUNDAJE, ALDRED et al. 2012). GENCODE est un sous-projet d'ENCODE ayant pour mission initiale d'identifier et de cartographier tous les gènes humains codant pour des protéines (HARROW, FRANKISH, GONZALEZ et al. 2006). Le projet se poursuit depuis, avec notamment l'ajout régulier d'annotations pour des ARN longs non codants et de transcrits d'épissage alternatif (HARROW, FRANKISH, GONZALEZ et al. 2012), ainsi que le support des mêmes types d'annotations chez la souris (FRANKISH, DIEKHANS, FERREIRA et al. 2019). Les annotations de GENCODE proviennent de deux sources principales : le projet d'annotation automatique Ensembl et le projet d'annotation manuelle HAVANA (FIGURE 1.44). De nos jours, Ensembl a dépassé son cadre initial en proposant des fichiers d'annotations sensiblement identiques à ceux de GENCODE mais avec un cycle de mise à jour plus

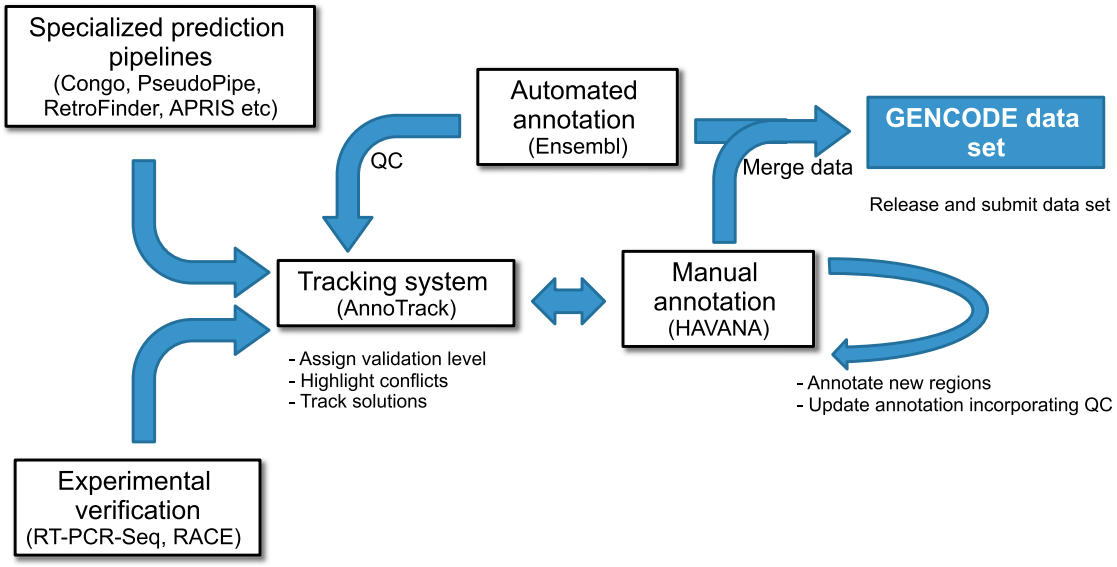


FIGURE 1.44 – Flux de données entre les différents groupes du consortium GENCODE. Extrait de HARROW, FRANKISH, GONZALEZ et al. 2012.

court. Dans le cadre de ce travail de thèse, ce sont ces annotations Ensembl qui sont utilisées.

1.3.3.4 Classification des gènes en ontologies

Alors que le génome de l'Homme et ceux d'autres espèces étaient encore en cours de séquençage, il est apparu qu'un grand nombre de fonctions biologiques étaient tenues par le produit de gènes similaires pour l'ensemble des eucaryotes. L'identification du rôle biologique d'un de ces produits chez un organisme pouvait raisonnablement être transféré aux autres organismes. En 1998, le consortium Gene Ontology (GO) s'est constitué afin de développer une nomenclature pouvant être appliquée chez tous les eucaryotes pour décrire le rôle des gènes (GENE ONTOLOGY CONSORTIUM 2000). Cette nomenclature est organisée sous la forme d'un graphe acyclique orienté composé de trois ontologies indépendantes : processus biologique, fonction moléculaire et composant cellulaire (FIGURE 1.45). Les efforts de recherche coordonnées par le consortium GO ainsi que l'adoption de la nomenclature par d'autres projets, *e.g.* FANTOM (FANTOM CONSORTIUM 2001), ont permis de constituer la plus grande base de données structurées d'annotations de gènes. Plus de vingt ans plus tard, cette dernière continue d'être mise à jour mensuellement (CARBON, DOUGLASS, DUNN et al. 2019).

D'autres ontologies de gènes ont été parallèlement développées avec des objectifs différents. On peut citer de manière non exhaustive KEGG, Reactome Pathways, Treefam (LI 2005), BioCyc et les collections MSigDB (SUBRAMANIAN, TAMAYO, MOOTHA et al. 2005 ; LIBERZON, SUBRAMANIAN, PINCHBACK et al. 2011 ; LIBERZON, BIRGER, THORVALDSDÓTTIR et al. 2015). Conçues pour être utilisées

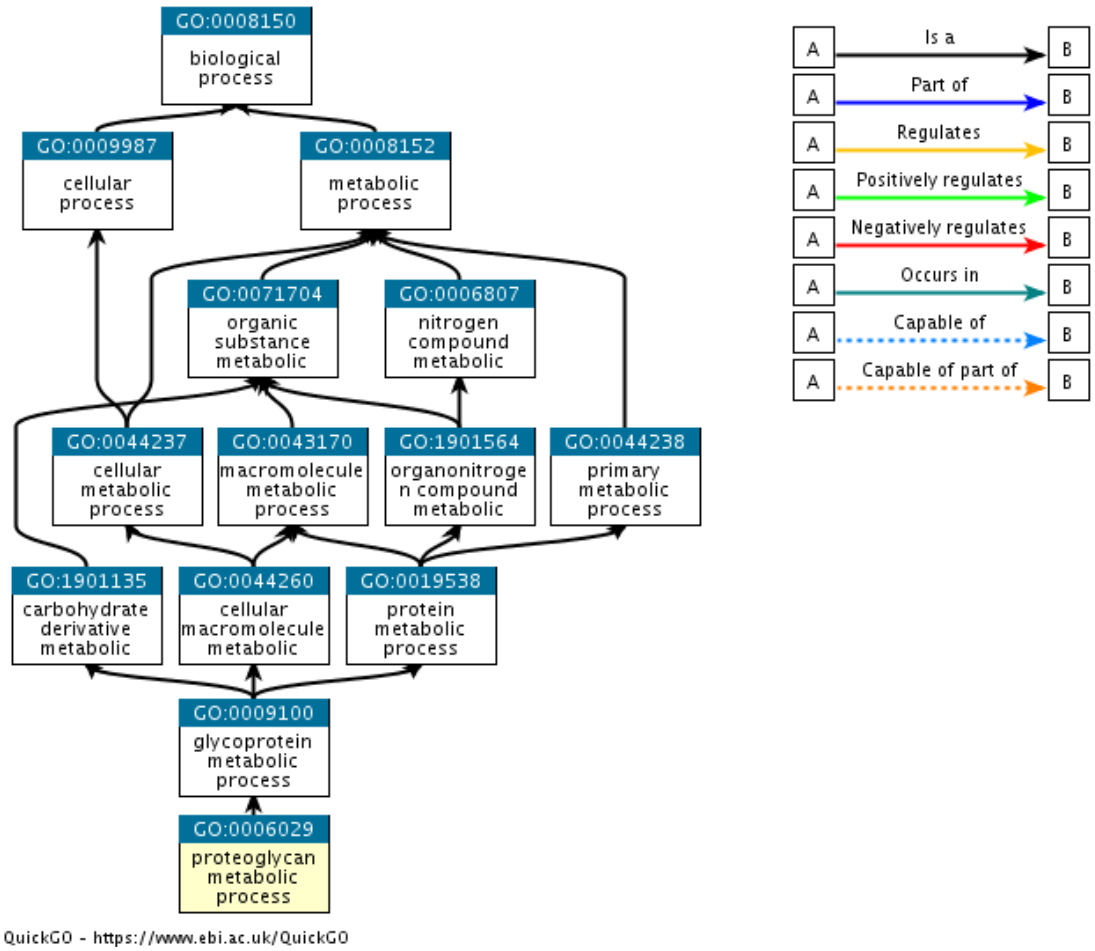


FIGURE 1.45 – Graphe orienté acyclique représentant les ancêtres d'un terme GO : GO:0006029

par défaut dans l'outil d'anlyse d'enrichissement fonctionnel GSEA, ces collections sont également pratiques pour d'autres types d'analyse d'enrichissements fonctionnels de par la spécificité des termes constituants les ontologies, en comparaison relative de la redondance des termes GO intrinsèque à leur organisation hiérarchique en graphe orienté acyclique (SECTION 3.4).

1.3.3.5 Cartographies épigénomiques

En 1999, alors que le projet de séquençage du génome humain poursuivait son cours, s'est constitué le *Human Epigenome Consortium*, le premier consortium visant à fédérer un projet encore plus ambitieux : cartographier différents épigénomes de référence (BECK, OLEK et WALTER 1999). Devant l'ampleur de la tâche, de l'intérêt scientifique et sociétal croissant pour une discipline promettant d'expliquer de nombreuses pathologies modernes et grâce aux développements successifs de méthodes d'analyse de l'épigénome toujours moins onéreuses et plus précises, se sont succédées différentes itérations du projet (JONES et MARTIENSSEN 2005 ; JONES, BAYLIN, BERNSTEIN et al. 2008). L'*International Human Epigenome Consortium* est créé en 2010 avec pour objectif la production de 1000 épigénomes de référence en 7 à 10 ans (BAE 2013). Il coordonne et rassemble les résultats des projets

— canadien (CEEHRC),
— européen (BLUEPRINT),
— états-uniens (ENCODE et NIH Roadmap),
— allemand (DEEP),
— britannique (HipSci),
— japonais (AMED-CREST),
— coréen (KNIH),
— singapourien (GIS) et
— chinois (EpiHK).

au sein de l'*Epigenome Reference Registry* (EpiRR). Si l'on s'en tient aux stricts critères actuels de définition d'un "épigénome de référence" (TABLE 1.2), le projet atteindra difficilement son objectif initial avec seulement 361 épigénomes complets presque neuf ans après son lancement (TABLE 1.3). Néanmoins, en addition des 3068 épigénomes partiels actuellement disponibles et des nombreuses publications qui en découlent, le projet est surtout considéré comme un succès (STUNNENBERG, HIRST, ABRIGNANI et al. 2016 ; NATURE METHODS EDITORIAL 2018). Si l'ensemble des données processées est accessible librement au travers de multiples interfaces simples d'utilisation, on peut regretter en tant que bioinformaticien que la plupart des données brutes soient accessibles uniquement sur procédure administrative dépendante du consortium ayant généré les données (DYKE, CHEUNG, JOLY et al. 2015). Des conséquences de cette contrainte sont abordées en SECTION 1.4.3.1.

	Epigénome de référence minimal selon	
	IHEC	NIH Roadmap
Bisulfite-Seq	requis	
MeDIP-Seq MRE-Seq RRBS MethylCap-Seq		1/4 requis
RNA-Seq Array based	requis	1/2 requis
smRNA-Seq	optionnel	optionnel
ChIP-Seq input	requis	requis
H3K27me3	requis	requis
H3K36me3	requis	requis
H3K4me1	requis	requis
H3K4me3	requis	requis
H3K27ac	requis	optionnel
H3K9me3	requis	requis
DNase-Seq DGF FAIRE-Seq	1/3 optionnel	1/3 optionnel

TABLE 1.2 – Critères de définition d'un épigénome de référence complet. Adapté de Reference Epigenome Standards.

	Nombre d'épigénomes		
Projet	Complet	Partiel	Total
AMED-CREST	20	355	375
BLUEPRINT	112	1137	1249
CEEHRC	113	335	448
DEEP	54	103	157
ENCODE	0	121	121
GIS	0	291	291
HipSci	0	640	640
KNIH	0	37	37
NIH Roadmap	62	49	111
Total	361	3068	3429

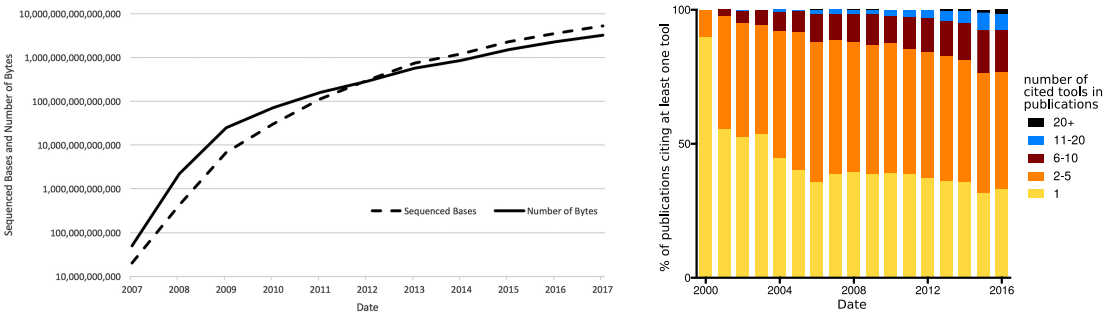
TABLE 1.3 – Comptage des épigénomes générés dans le cadre d'IHEC. Dernière mise à jour le 14 mai 2019. Une version actualisée de ces données est accessible ici.

1.3.4 Nouveau paradigme pour l'analyse des données

La baisse du coût de séquençage, l'augmentation du nombre de méthodes expérimentales s'y reposant pour explorer différentes facettes du génome et l'augmentation des données générées (FIGURE 1.46a) exercent une pression accrue sur les ressources informatiques et humaines nécessaires pour les analyser (MUIR, LI, LOU et al. 2016). Pour le bioinformaticien en charge de l'analyse de données d'un projet biologique, cela se traduit par une complexification des analyses nécessaires, visible par une augmentation du nombre d'outils bioinformatiques utilisés dans les articles publiés (FIGURE 1.46b).

La racine de cette complexification tient dans la nécessité de ne plus seulement analyser les données d'un type d'expérience, mais aussi de les croiser et de les intégrer avec l'ensemble des données pertinentes déjà générées qu'il faudra potentiellement réanalyser avec la nouvelle problématique biologique en considération. Ceci implique d'avoir à traiter des données massives, hétérogènes et fragmentées sur des supports de stockage publics ou privés dispersés. Leur hétérogénéité est liée au type d'expérience auxquelles elles se rapportent, mais aussi aux évolutions technologiques au sein même d'un type d'expérience, *e.g.* transition *single-end* vers *paired-end*, *long reads*, *single-cell*, *etc*.

La deuxième raison est liée à l'explosion du nombre d'outils et de protocoles bioinformatiques publiés. Le service Omictools recense ainsi 377 outils d'alignement de reads, 534 protocoles d'analyse différents rien que pour le ChIP-Seq, et 356 pour le RNA-Seq (HENRY, BANDROWSKI, PEPIN et al. 2014 ; CLÉMENT, EMERIC, GONZALEZ et al. 2018). Disposer d'une méthodologie simple et pratique permettant d'une part de déterminer et d'appliquer le protocole le plus approprié pour chaque type de données, d'autre part d'intégrer ces différents types de données à différentes échelles, est d'un intérêt majeur pour la recherche. Cet enjeu m'a poussé à m'intéresser particulièrement aux notions gravitant autour du concept de reproductibilité abordé dans la section suivante.



(a) Accumulation des données dans SRA. Extrait de KARSCH-MIZRACHI, TAKAGI et COCHRANE 2017

(b) Complexification des protocoles. Extrait de CLÉMENT, EMERIC, GONZALEZ et al. 2018.

FIGURE 1.46 – Tendances clés illustrant l'entrée de la génomique dans un nouveau paradigme de recherche.

1.4 Reproductibilité en bioinformatique

1.4.1 Définitions

Un des piliers de la méthode scientifique consiste en la possibilité d'évaluer l'exactitude des raisonnements et conclusions établis par d'autres scientifiques. Ceci requiert une démarche décrivant avec précision les procédures expérimentales et l'analyse des données qui en découlent. Idéalement, toute expérience et son analyse devraient être décrites avec suffisamment de détails pour permettre à d'autres scientifiques possédant les compétences et les moyens de suivre les différentes étapes du travail publié et d'obtenir les mêmes résultats.

En pratique, la plupart des approches expérimentales sont soumises à des biais techniques et humains rendant souvent difficile la reproduction d'une expérience. Depuis plusieurs siècles, la méthode scientifique s'accomode de cette limitation en accordant un crédit supérieur à un résultat lorsque celui-ci est reproduit à plusieurs reprises par des chercheurs indépendants. La science avance par corroboration, et si un résultat ne peut être reproduit alors il est remis en question ou tombe dans l'oubli. Un problème majeur ayant émergé ces dernières décennies sous l'appellation de "crise de la reproductibilité" est la recrudescence de publications non reproductibles dans la plupart des domaines scientifiques (BAKER et PENNY 2016). Si l'approche corroborative permettra sûrement de séparer le bon grain de l'ivraie, il n'est pas satisfaisant à l'échelle humaine de consacrer du temps à corroborer ou invalider des résultats plutôt qu'à explorer de nouveaux mécanismes inconnus.

Parallèlement, l'avènement de l'informatique moderne et son utilisation pour simuler ou analyser des données expérimentales laissait supposer une disparition de cette limitation dans les domaines où il est possible de l'appliquer. En effet, l'ordinateur est un calculateur parfait et tout résultat issu d'algorithmes et de méthodes décrites devrait pouvoir être reproduit à l'identique. Dans la pratique, la façon d'implémenter les algorithmes décrits, les variations introduites par l'environnement matériel et logiciel, ainsi que de potentielles erreurs humaines rendent la reproduction d'analyses incertaine.

Jon Claerbout fut l'un des premiers à s'attaquer à ce problème (CLAERBOUT et KARRENBACH 1992) en décrivant une méthodologie de recherche reproductible basée sur un ensemble de solutions logicielles permettant de communiquer au lecteur d'un article l'intégralité des processus des données brutes jusqu'aux figures et tables. Dans cette approche la reproductibilité prend un sens différent de l'usage scientifique jusqu'alors. Pour éviter toute confusion, Goodman suggère d'être plus explicite et de parler de *reproductibilité de méthodes* (GOODMAN, FANELLI et IOANNIDIS 2016). Sa définition serait de fournir suffisamment de détails sur les procédures et les données de façon à ce que les mêmes procédures puissent être exactement répétées sur les mêmes données. On fait alors une distinction avec la *reproductibilité de résultats* qui consiste à obtenir les mêmes résultats à partir d'une étude indépendante visant à reproduire au plus près les

procédures de l'étude originale, et la *reproductibilité inférentielle* qui consiste à établir les mêmes conclusions à partir de procédures indépendantes et des données de l'étude originale ou d'une autre.

En réintégrant cette triple distinction, on peut noter qu'une méthodologie de recherche reproductible s'intéresse en priorité à la reproductibilité de méthodes. Les reproductibilités de résultats et inférentielle se vérifient par application de la méthode scientifique générale. Mon travail de thèse ne fait pas exception et par la suite l'emploi du terme reproductibilité est toujours associé à la reproductibilité de méthodes. Additionnellement, un *flux de travail* est défini comme étant la suite complète d'étapes, manuelles ou automatisées, permettant à partir de données brutes d'obtenir des résultats d'intérêt scientifique : figures d'article ou fichiers de données processées. Un flux de travail est alors reproductible si toutes ses étapes sont correctement documentées et il est automatisé si toutes ses étapes le sont.

Une source majeure de confusion et de déception relative à ce concept réside dans la diversité des sensibilités, expertises et attentes pratiques propres à chaque individu par rapport à une expérience ou un résultat présenté comme reproductible. Des connaissances techniques peuvent être nécessaires pour pouvoir comprendre un flux de travail reproductible et ce résultat n'est alors effectivement reproductible que pour les individus possédant ces connaissances ou disposant du temps nécessaire pour les acquérir. La meilleure façon de s'assurer une reproductibilité effective pour une large audience consiste à automatiser autant que possible un flux de travail.

De plus, un flux de travail reproductible et automatisé peut ne pas correspondre à l'attente de certains individus car étant non-réutilisables pour des applications à d'autres échantillons et projets plus ou moins similaires. La reproductiblité et la réutilisabilité sont deux concepts distincts, mais c'est un souhait bien légitime d'espérer que l'un vienne avec l'autre. C'est pourquoi un objectif secondaire d'une méthodologie de recherche reproductible consiste à simplifier la réutilisabilité en appliquant des modèles généralisables.

1.4.2 Enjeux

Plusieurs motivations pour faire de la recherche reproductible sont détaillées ci-dessous. Si la première raison est intrinsèquement liée aux nouvelles exigences des journaux scientifiques concernant la mise à disposition d'analyses reproductibles pour toute publication, les suivantes sont plus personnelles.

Transparence et inspection La première motivation est d'être en mesure de vérifier l'exactitude des résultats produits. Les descriptions contenues dans les articles publiés sont rarement suffisantes pour convaincre les lecteurs sceptiques de la fiabilité du travail réalisé. De nos jours, la complexité des analyses publiées ne peut être retranscrite dans le format "publicitaire" contraignant d'article à nombre de caractères et de figures limités. Il faut faire de nombreux choix simplificateurs et le lecteur peut avoir l'impression de devoir faire aveuglément

confiance dans ces choix non détaillés par les auteurs de l'article original. Adopter un flux de travail reproductible signifie fournir au lectorat le code et les données qui détaillent l'ensemble des décisions prises pour la génération des résultats. Cela permet aux autres et à nous-même de s'assurer plus facilement de la fiabilité des résultats.

Traçabilité et simplicité dans l'organisation du travail Cette possibilité d'inspection du travail accompli permet également d'être en mesure de retrouver les détails de l'ensemble des analyses réalisées sur des durées dépassant les capacités mémorielles humaines. Cela simplifie le travail de rédaction des méthodes pour les articles dont les analyses peuvent durer plusieurs années et assure la possibilité d'interrompre puis de reprendre des analyses après plusieurs mois.

Sérennité face à la complexité En tant que bioinformaticiens, nous sommes amenés quotidiennement à faire des choix d'outils, d'algorithmes et de paramètres qui peuvent être débatables. Un flux de travail reproductible permet de ne pas se retrouver paralysé par l'infinité des choix possibles. Il devient possible d'appliquer initialement la combinaison de choix qui semble à ce moment la plus pertinente en sachant qu'il sera toujours possible de revenir sur d'autres combinaisons afin de comparer si les résultats convergent ou divergent vers la même interprétation biologique.

Réutilisation et partage Une autre raison d'aspirer à la reproductibilité est de permettre à d'autres scientifiques d'utiliser nos méthodes et nos résultats. Équipés uniquement de notre article publié, nos collaborateurs ou confrères pourraient avoir du mal à reconstruire notre méthode avec suffisamment de détails pour l'appliquer à leurs propres données. Adopter un flux de travail reproductible adapté doit permettre d'étendre notre approche à de nouvelles applications avec un minimum d'effort. Cela peut constituer un gain de temps potentiel énorme pour les équipes et la communauté scientifique.

1.4.3 Éventail des pratiques et solutions existantes

La mise en place d'une méthodologie de recherche reproductible est un processus long, complexe mais subjectivement intéressant. En l'absence de méthodologie consensus convenant à toutes les disciplines et du fait de l'évolution constante des technologies informatiques, une méthodologie valide un jour peut ne plus l'être une décennie plus tard. Implémenter la méthodologie originelle de Claerbout aujourd'hui serait risible, à moins d'être un fan inconditionnel du partage de données par pile de CD-ROMs.

Paradoxalement, la complexité liée à la mise en place d'une telle approche n'est pas l'absence de solutions existantes, mais plutôt l'existence d'un trop grand nombre de solutions développées en parallèle pour répondre aux besoins. Le jeune chercheur doit investir un temps conséquent pour choisir, apprendre et

appliquer une méthodologie entièrement productive. Je présente dans cette section ma vision des différents pratiques et solutions existantes qui concernent la traçabilité des données, le développement logiciel, l'automatisation des interactions entre données et logiciels, et enfin le partage des résultats. Cette vision est limitée par les applications que j'ai rencontré au cours de ma thèse et je suis intéressé par toute suggestion d'amélioration et de mises à jour.

1.4.3.1 Traçabilité des données

Les données sont au cœur de la reproductibilité. Leur collecte, leur stockage, leur transformation, le contrôle de leur qualité, leur partage et leur identification sont des préoccupations majeures. Afin de maintenir leur traçabilité tout au long du processus de recherche, plusieurs solutions peuvent être envisagées, l'idée maîtresse étant qu'il soit possible de remonter à l'ensemble des étapes ayant abouties à la production de chaque fichier produit dans une analyse. On peut distinguer plusieurs types de données : les données d'entrées privées, les données d'entrées publiques, les données processées et les données hybrides. Ce type détermine les approches possibles pour maintenir la traçabilité.

Données d'entrées publiques Les données d'entrées publiques correspondent aux données accessibles librement par Internet nécessaires pour réaliser une analyse. Dans le cadre de la bioinformatique génomique, on peut citer dans cette catégorie l'ensemble des ressources contenues dans les archives de séquences (SRA, ENA, DRA) et dans les bases de données d'annotations. Ces ressources bénéficient déjà généralement d'un système de traçabilité qu'il suffit de noter : identifiant de chaque échantillon, version de release, *etc*. En l'absence ou en complément de tels systèmes, conserver le *Digital Object Identifier* (DOI), ou le cas échéant l'*Uniform Resource Locator* (URL), ainsi que la date d'accession à une ressource est une bonne pratique qui peut être appliquée soit en organisant rigoureusement les fichiers dans l'arborescence, soit en y ajoutant des métadonnées avec un outil comme exiftools, soit avec une base de données.

Données d'entrées privées Les données d'entrées privées correspondent aux données générées dans le cadre d'un projet de recherches accessibles à un groupe restreint de collaborateurs. En bioinformatique génomique, cela correspond principalement aux données issues de séquenceurs. Une attention particulière doit être portée à leur préservation, sauvegarde et réplication sur plusieurs supports de stockage. La traçabilité de ces données doit avant tout être assurée par les collaborateurs qui les ont générées. En tant que bioinformaticien, il convient de conserver ces fichiers intacts sous la forme sous laquelle nous les recevons. Ceci implique de ne pas renommer manuellement ces fichiers. Pour garantir le respect de cette procédure, il est possible de conserver ces fichiers dans un dossier isolé avec un accès en lecture seule.

Une méthodologie de recherche reproductible s'appliquant à des données d'entrées privées doit y accorder une attention particulière au moment de la publi-

cation de résultats. Si à ce moment les données peuvent être rendues publiques, alors il faut ajuster les méthodes de façon à ce que les analyses puissent être reproduites à partir des données hébergées publiquement. Si les données ne peuvent être rendues publiques car elles appartiennent à une entité en limitant l'accès (BLUEPRINT et RepBase pour ma part), il faut renseigner l'entité et les identifiants des données.

Données processées Les données processées correspondent aux données issues de l'application d'outils logiciels dans le cadre des analyses. L'enchaînement des outils et opérations ainsi que leur version et paramètres doivent être repertoriés et accessibles en explorant les données processées. Effectuer cette tâche manuellement est fastidieux et source d'erreurs. Il est préférable de se reposer sur un gestionnaire de flux de travail (SECTION 1.4.3.3) qui va permettre d'automatiser l'ensemble des analyses et de spécifier les environnements logiciels et les opérations de chaque étape de façon à ce que l'ensemble des données puissent être regénérées à partir d'une réexécution du flux de travail. Dans le cadre d'une méthodologie de recherche reproductible traitant des données massives, il est intéressant d'isoler dans un dossier ce type de données. En effet, le stockage est une ressource précieuse sur les serveurs et il peut être souhaitable à la fin d'une étude de pouvoir, en une opération de dossier, supprimer tous les fichiers processés puisqu'il est de toute façon possible de les regénérer.

Données hybrides Enfin, les données hybrides rassemblent les différents cas de figures où l'attribution à un des précédents types est discutable. Ainsi, des données processées dont la génération requiert un temps de calcul suffisamment important pour que l'on ne souhaite pas les regénérer ne devraient pas être considérées comme supprimables à tout moment mais plutôt comme des données d'entrées de deuxième ordre. De même, il est possible que certaines étapes d'analyses ne puissent être trivialement automatisées et requièrent une action manuelle. Dans ce cas, le produit de cette étape est une donnée processée mais préférablement protégée comme une donnée d'input. Enfin, en automatisant des téléchargements de données publiques de sources fiables et pérennes, on peut les considérer conceptuellement comme des données processées et ainsi faire l'économie de leur stockage local.

1.4.3.2 Développement logiciel contrôlé

En informatique, un logiciel est un ensemble de séquences d'instructions interprétables par une machine et d'un jeu de données nécessaires à ces opérations. Un flux de travail automatisé présente de grandes similarités. En conséquence, de nombreuses pratiques de développement logiciel peuvent être récupérées et appliquées pour garantir la précision et la fiabilité du flux de travail automatisé.

Contrôle de version Le moindre changement apporté à un programme peut avoir d'importantes conséquences désirées ou indésirées. Une amélioration ap-

portée à un script pour une nouvelle analyse peut être synonyme de régression, bug ou altération du résultat pour une analyse antérieure. La première étape pour pouvoir aborder ce genre de conflits est l'utilisation d'un logiciel de gestion de versions tel que Git, Subversion ou Mercurial. Additionnellement, ces outils peuvent être utilisés pour collaborer à plusieurs sur des projets de développement logiciel ou d'analyses de données. Les plateformes en ligne Github, GitLab et Bitbucket proposent des solutions d'hébergement et de collaboration gratuites pour la plupart des besoins en recherche.

Tests et intégration continue La deuxième étape pour aborder ce genre de conflits est d'appliquer des principes d'intégration continue. L'objectif est de détecter à chaque modification de code que le résultat des modification n'entraîne pas d'altération dans les analyses déjà effectuées. Idéalement, des tests unitaires devraient être écrits pour chaque fonction et des tests intégratifs pour l'ensemble des flux de travail. L'exécution de ces tests peut être automatisée à chaque soumission de modification de code (commit) par exemple via des crochets git locaux ou bien par le biais de services en ligne tel que circleci ou travis. Ces tests doivent correspondre à des exemples sur des fichiers légers de ce que devrait faire le code du flux de travail.

Alternativement, si les ressources de calculs ne sont pas un facteur limitant, il est possible de déplacer le dossier de données processées d'un flux de travail, puis de le réexécuter ce dernier. La comparaison des fichiers déplacés et des fichiers regénérés informe sur l'aspect reproductible de l'analyse. Selon les types de données, la comparaison peut s'effectuer avec un outil de comparaison de lignes comme diff ou un outil de comparaison de hash.

Ces tests sont utiles pour identifier les étapes dépendantes de la génération de nombres pseudo-aléatoires. Ceux-ci généreront à chaque exécution des résultats différents que l'on peut espérer très similaires. Il est en général possible de fixer une graine aléatoire pour obtenir des résultats déterministes. Néanmoins, c'est parfoit insuffisant, notamment pour l'application de certains algorithmes parallélisés ou s'exécutant sur processeurs graphiques. Fixer systématiquement une graine aléatoire peut aussi cacher la fluctuation d'un algorithme derrière l'apparente robustesse. Dans tous ces cas, il peut être plus intéressant d'évaluer la reproductibilité inférentielle des données plus que leur exacte reproduction. En d'autres termes, si la structure exacte des données peut varier, les figures finales doivent toujours permettre de mener à la même interprétation. Cette opération doit être réalisée manuellement bien que l'on puisse citer pour l'anecdote des approches telles que pytest-mpl pour automatiser ce processus.

Déploiement contrôlé Lorsqu'un flux de travail devient raisonnablement reproductible pour soi, il est important de s'assurer qu'il le soit aussi pour d'autres utilisateurs ayant un environnement logiciel différent. Lister la version des programmes utilisés dans un protocole est une façon très insuffisante de permettre la reproductibilité. D'une part parce que derrière une version d'un logiciel listé

peut se cacher un long et fastidieux travail d'installation. D'autre part parce cette version peut ne plus être disponible au moment où l'on souhaite reproduire l'analyse. De plus, chaque programme possède sa propre liste de dépendances ce qui multiplie les risques d'être dans l'incapacité de refaire fonctionner un flux de travail.

Plusieurs solutions sont possibles pour résoudre ce type de problématique, chacune avec ses forces et faiblesses. Historiquement, l'usage de machine virtuelle a été la première solution. Le principe consiste à développer un flux de travail à l'intérieur d'une machine entièrement virtualisée, système d'exploitation et environnement logiciel complet inclus. Partager le flux de travail revient alors à partager avec d'autres utilisateurs la machine virtuelle qui a permis de l'exécuter. Cette solution présente l'avantage de la simplicité pour un utilisateur sans grande connaissance informatique car il n'est pas nécessaire d'avoir conscience des versions des outils et dépendances requises. Il suffit juste d'arriver, par miracle ou persévérance, à un résultat fonctionnel et celui-ci pourra être répliqué sur d'autres machines aux performances matérielles équivalentes. L'inconvénient de cette solution est sa lenteur. En effet la virtualisation d'un système d'exploitation complet ralentit la vitesse d'exécution des tâches. De plus, il n'est pas possible de faire du calcul distribué sur cluster à plusieurs nœuds de calculs par cette approche. Cette solution est donc inadaptée pour les flux de travail traitant une quantité importante de données.

Une alternative plus récente repose sur l'isolation de l'environnement logiciel avec des containers. Une première approche consiste à utiliser un container de la même façon qu'une machine virtuelle. Cependant, comme un container se passe de la simulation d'un système d'exploitation, il est plus léger et réactif qu'une machine virtuelle. Une autre configuration plus pratique est alors applicable : chaque outil possède son propre container (FIGURE 1.47) et à chaque fois qu'une étape d'un flux de travail doit être réalisée, une instance du container est créée pour les outils impliqués. L'avantage de cette configuration est qu'un flux de travail peut être déployé sur des infrastructures matérielles à taille variable (ordinateur, cluster de calcul, cloud) et est donc adapté pour traiter une quantité de données importante. Docker est la solution de containers la plus populaire dans les milieux informatiques. Singularity est une alternative conçue spécialement pour le calcul scientifique et présentant des atouts pour l'exécution en cluster de calculs (KURTZER, SOCHAT et BAUER 2017). Kubernetes est une solution populaire pour l'orchestration de containers dans le cloud.

Ces approches nécessitent un prérequis pour fonctionner : les outils de containerisation doivent être obligatoirement installés par l'administrateur de la machine utilisée. Ceci peut ne pas être le cas pour des clusters de calcul. En particulier, celui que j'ai principalement utilisé pour ma thèse n'en disposait pas. Dans ce cas, une troisième approche est possible. Celle-ci repose sur la gestion des environnements logiciels par un gestionnaire de paquets multi-plateformes multi-langages. Dans ce cas, chaque outil peut posséder son environnement logiciel mais il n'est pas physiquement isolé du reste du système comme c'est le

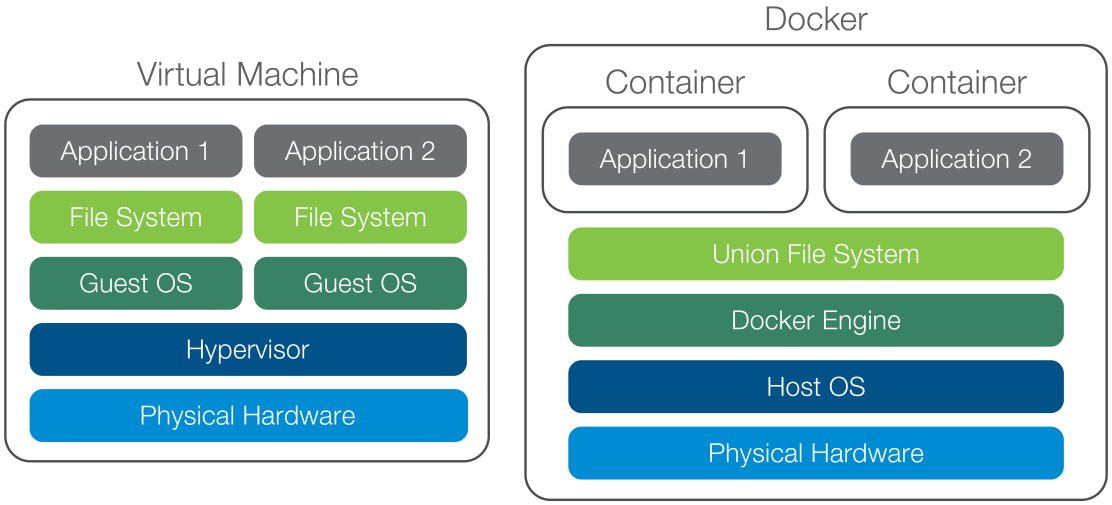


FIGURE 1.47 – Différences entre machine virtuelle et container. Un environnement applicatif déployé sur des machines virtuelles nécessite de virtualiser un système d'exploitation complet et son système de fichiers pour chaque application. Le déploiement de ces mêmes applications avec des containers permet d'utiliser les ressources du système d'exploitation physique tout en isolant les fichiers nécesaires pour chaque application dans son container. Cette distinction permet d'atteindre avec des containers une utilisation plus efficiente des ressources disponibles.

cas pour un container. L'avantage de ces approches est de pouvoir fonctionner pour tout utilisateur indépendamment d'un accès administrateur à la machine. L'inconvénient est qu'il devient plus complexe de certifier qu'aucun problème de dépendances ne puisse survenir. En pratique, le risque est très limité quand les paquets sont correctement construits. Deux gestionnaires de paquets de ce type sont Nix et Conda. C'est vers ce dernier que je me suis tourné pour mon travail de thèse et mes développements logiciels. J'ai ainsi empaqueté plusieurs logiciels (disponibles ici) et contribué au projet Bioconda (DALE, GRÜNING, SJÖDIN et al. 2018) qui empaquète et rend disponible plus de 6325 logiciels d'intérêt pour la bioinformatique (Nombre mesuré le 22 mai 2019).

1.4.3.3 Automatisation des interactions entre logiciels et données

L'automatisation est la pierre angulaire de toute méthodologie de recherche reproductible. Son objectif est d'assurer autant que possible les interactions entre logiciels et données. Un flux de travail reproductible peut ne pas être automatisé, mais toute personne intéressée par la reproduction d'une analyse appréciera de n'avoir qu'une commande à exécuter, plutôt que de suivre des instructions manuelles. Différentes approches complémentaires existent.

L'approche la plus basique consiste en l'écriture d'un script, habituellement en langage shell sous Linux, dont l'objectif va être d'orchestrer l'ensemble des étapes du flux de travail. C'est à la fois la solution la plus simple pour s'initier à la méthodologie de recherche reproductible automatisée, mais aussi la plus complexe pour arriver à parfaire ses différents concepts :

— Scalabilité, *i.e.* possibilité d'éxecution du flux de travail sur des infrastructures de tailles différentes pour des données de volumes variables.

— Portabilité, *i.e.* possibilité de déployer le flux de travail sous différentes solutions déjà évoquées : dans un machine virtuelle, dans des containers, grâce à un gestionnaire de paquets multi-plateformes.

— Réentrabilité, *i.e.* possibilité d'interrompre le flux de travail et de le reprendre à l'endroit interrompu au lieu du début.

— Archivage, *i.e.* disposer d'une procédure pour la conservation des données d'entrées et de l'ensemble des environnements logiciels utilisés.

— Accessibilité, *i.e.* un collaborateur externe peut comprendre le fonctionnement du flux de travail sans aide de celui qui l'a rédigé, en se basant uniquement sur sa documentation.

Heureusement, différentes infrastructures logicielles existent pour fournir un cadre permettant plus simplement de satisfaire à tous ces concepts.

Programmation lettrée et document computationnel Le document computationnel est une bonne première approche pour permettre aux débutants d'assurer l'accessibilité de leur travail. C'est une application pour des analyses de données du principe de programmation lettrée (KNUTH 1984). Celui-ci repose sur

l'idée qu'un programme informatique est plus facilement documenté, et donc accessible aux autres, s'il est considéré comme un récit ou rapport dans laquelle la logique de l'enchaînement des étapes est présentée à la manière d'un essai dans lequel sont inclus des blocs de code qui masquent les abstractions et la complexité. Après l'original langage WEB sont apparues des solutions analogues dans les langages usuellement utilisés pour de l'analyse de données : Sweave, puis knitr pour R, et Jupyter pour Python. Les adeptes d'Emacs apprécieront Org-mode. Bien qu'originellement pensées pour exécuter du code du langage dans lequel elles ont été développées, ces solutions permettent aujourd'hui d'intégrer tout type de code et d'arriver à un résultat similaire. Par conséquent, le choix de l'une d'entre elles repose sur des critères personnels ou consensuels entre collaborateurs.

Gestionnaire de flux de travail La structure du document computationnel qui fait tout son intérêt est aussi sa limite. Elle incite à une structuration de la pensée linéaire et chronologique. Un flux de travail pour le traitement de données complexes est plus facilement structuré sous la forme de graphe orienté acyclique. Cette structuration peut se faire grâce à un gestionnaire de flux de travail. Avec au moins 117 outils référencés, il est difficile de faire un choix entièrement éclairé pour en choisir un. Pour ma part, j'ai été introduit par mon co-directeur de thèse Denis Puthier à Galaxy puis à Snakemake. Si Galaxy présente de grandes vertues pédagogiques avec une interface graphique très accessible pour un utilisateur débutant, j'ai préféré développer l'ensemble de mes analyses grâce à Snakemake qui propose une interface en ligne de commande plus pratique pour un bioinformaticien. Quoi qu'il en soit, un gestionnaire de flux de travail permettant scalabilité, portabilité, réentrabilité, archivage et accessibilité est un bon choix.

1.4.3.4 Partage pérenne des produits de la recherche

Le partage pérenne des produits de la recherche avec le public se pense à plusieurs niveaux : articles, données, flux de travail et logiciels. Si les flux de travail pourraient être vus comme des logiciels, on fait ici une distinction basée sur le fait qu'un flux de travail est appliqué à des données brutes pour produire des données traitées et des figures d'articles, alors qu'un logiciel a une portée qui s'étend au delà d'une application pour un seul article. Cette distinction apporte des nuances sur la marche à suivre pour le partage de l'un et de l'autre.

Idéalement, les différents produits de la recherche doivent être partagés sous une forme autosuffisante qui permette de se passer du ou des chercheurs qui les ont produits. De ce fait, les présentations orales et leurs supports, diapositives ou posters, sont écartés puisqu'ils nécessitent la présence du chercheur pour une diffusion correcte du savoir.

Articles Les articles sont le premier vecteur de dissémination du savoir scientifique. Les chercheurs produisent de la recherche financée principalement sur fonds publics, écrivent des articles, évaluent gratuitement d'autres travaux au

travers des comités de lecture et fournissent le plus gros du travail de mise en forme. Pourtant, jusqu'à récemment, la prévalence quasi-monopolistique de journaux proposant la publication des articles derrière des coûts unitaires ou d'abonnements payants empêchait l'accès à la recherche, et donc par extension, à la reproductibilité, à une grande partie de la population.

Cette épineuse problématique tend cependant à s'estomper ces dernières années. En effet, certains financeurs exigent désormais que les travaux produits par leurs fonds soient publiés dans des revues à accès libre ou dans des archives ouvertes (HAL, arXiv et variantes) sous peine de sanction. De plus, le développement discutable de solutions pirates (Sci-Hub et Libgen) de mise à disposition d'articles scientifiques permet à n'importe qui d'accéder à une littérature plus complète que la plupart des combinaisons d'abonnements auquel un chercheur peut prétendre.

Données Si le cas de la publication des données issues du séquençage a déjà été évoqué (SECTION 1.3.3.1), d'autres types de données peuvent être publiées. Pour ces cas, Cyverse, Figshare et Open Science Framework proposent des solutions d'hébergement de données avec une fonctionnalité intéressante d'association des jeux de données à un DOI pour permettre un référencement théoriquement plus robuste à l'épreuve du temps qu'un simple URL. En effet, l'URL identifie une localisation sur Internet alors qu'un DOI identifie la ressource de façon persistante. Dans ce cas, la localisation n'est qu'une information contenue dans un DOI et elle peut être mise à jour si nécessaire sans que les références au DOI ne soient affectées.

Donner accès aux données générées est important pour la communauté scientifique mais n'est pas suffisant sans une documentation claire des méthodologies employées pour produire les données. Les données sont généralement référencées à partir de l'article qui les a générées, mais il peut également être profitable d'inclure des métadonnées détaillées directement avec les données afin d'améliorer leur référencement dans les moteurs de recherche globaux ou dédiés. L'ensemble des solutions citées précédemment permettent d'associer ces métadonnées au DOI. Ces dernières annéees, le concept de *FAIR data* s'est développé et rassemble des principes érigés afin d'assurer quatre objectifs : que les données publiées soit trouvables, accessibles, interopérables et réutilisables (WILKINSON, DUMONTIER, AALBERSBERG et al. 2016).

Analyses de données Les analyses constituées des flux de travail, documents computationnels et rapports générés automatiquement peuvent être publiés dans un dépôt de code. Zenodo est une initiative intéressante pour pouvoir fournir un DOI à la place d'un URL pour un dépôt d'analyses de données. L'article publié référençant le DOI, il peut alors être possible de changer d'hébergeur sans conséquence pour le lecteur. Pour les utilisateurs de Jupyter, Binder est un service intéressant qui transforme directement un dépôt Git en une collection de notebooks interactifs exécutés dans le cloud, rendant ainsi le code directement

reproductible par n'importe qui, n'importe où. Finalement, la nature multiple des produits de la recherche moderne peut amener à réfléchir au format élémentaire de la publication. Plutôt que d'avoir article, données et flux de travail séparés et reliés par de simples références croisées, plusieurs initiatives envisagent des solutions pour empaqueter l'ensemble de ces produits dans un même objet sous la dénomination de *Research Objects* (BECHHOFER, BUCHAN, DE ROURE et al. 2013). Ces solutions s'inscrivent en continuation des principes FAIR appliqués aux données.

Logiciels Comme pour le flux de travail, un minimum pour le partage d'un logiciel est la publication de son dépôt de code source. Pour faciliter la communication avec des utilisateurs ayant des habitudes diverses, il est préférable de proposer une documentation détaillée et un empaquetage dans plusieurs systèmes de gestion de paquets :

— langage-spécifique (CRAN/Bioconductor, pip ,CTAN, CPAN, npm, *etc.*),

— système-spécifique (apt, yum, pacman, *etc.*) et

— multi-plateformes (conda, nix, *etc.*),

— ainsi qu'en application isolée (Snap, Flatpak ou Appimage).

Toutefois, maintenir des paquets à jour pour l'ensemble de ces solutions est trop coûteux en temps pour être réalisé. Il faut alors privilégier les solutions apportant le plus de visibilité et couvrant le plus d'utilisateurs potentiels. En bioinformatique génomique, conda semble être la solution la plus populaire.

Si les logiciels et flux de travail sont mis à disposition du public, une dernière considération est à prendre en compte pour assurer leur exécution par d'autres : les licences (STODDEN 2009 ; MORIN, URBAN et SLIZ 2012). Un logiciel, même dont le code est publiquement accessible, ne peut être réutilisé en l'absence d'une licence autorisant expressément cet usage. De nombreuses licences plus ou moins permissives existent et peuvent être simplifiées en trois grandes catégories : les licences propriétaires, permissives (BSD, MIT, *etc.*) et copyleft (GPL, LGPL, *etc.*). Une licence est définie comme propriétaire si elle entrave au moins un de ces droits : utiliser, modifier, dupliquer ou diffuser le logiciel sur lequel porte la licence. Ce type de licence n'est pas compatible avec une méthodologie de recherche reproductible. À l'inverse, les licences permissives et copyleft sont libres et permettent une diffusion de recherche reproductible. La nuance entre ces deux types tient dans une limitation différente du champ des réutilisations possibles du code. Les licences permissives autorisent la réutilisation pour des applications sous tous types de licences, alors que les licences copyleft imposent que tout code réutilisant des composants sous licence copyleft le soit également. Ceci exclut donc la réutilisation pour des logiciels propriétaires mais aussi pour des logiciels à licence permissive (FIGURE 1.48). Par conséquent le choix d'un type de licence libre pour un logiciel n'est pas toujours une volonté de ses développeurs mais peut aussi être imposé si un des composants réutilisés est sous licence copyleft.

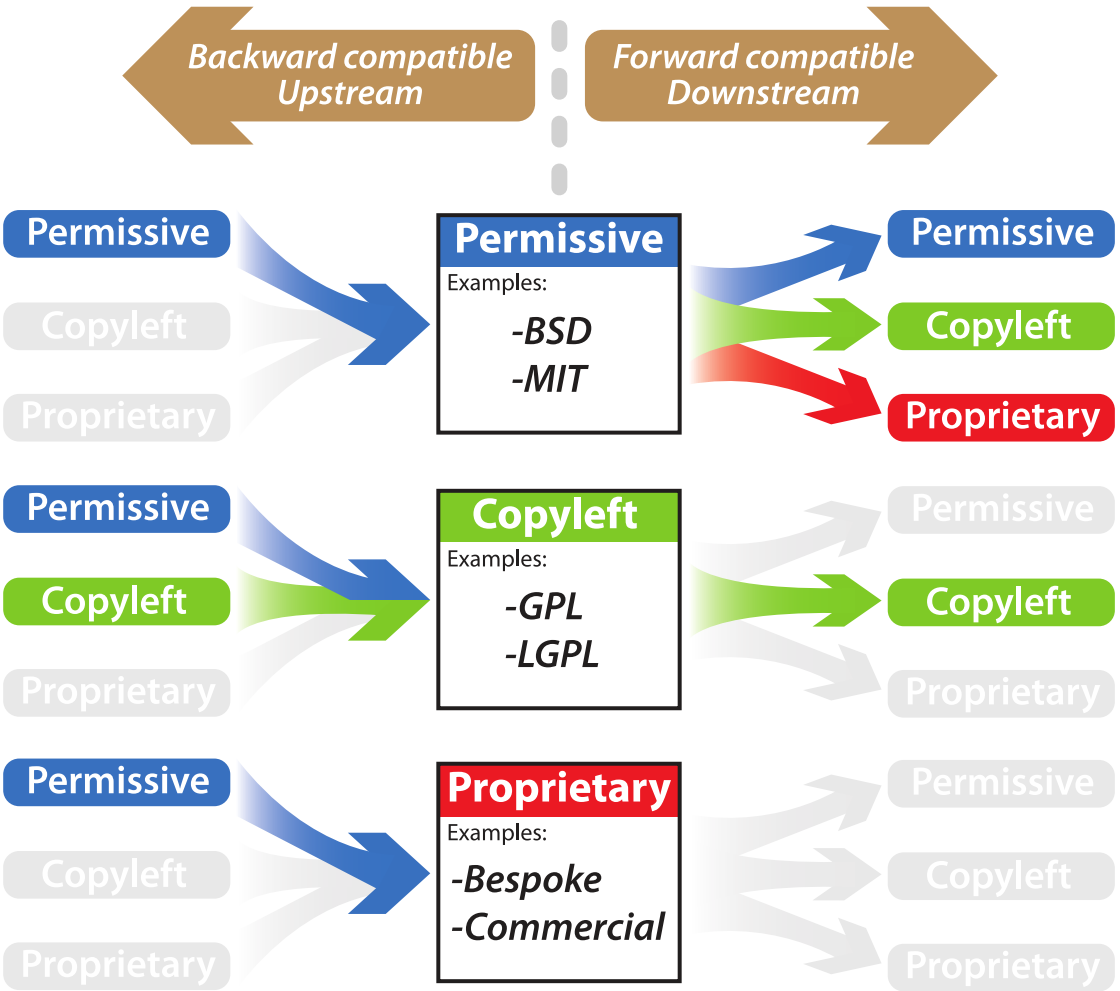


FIGURE 1.48 – Vision schématique de la directionnalité des principales licences. Extrait de MORIN, URBAN et SLIZ 2012.

1.5 Objectifs de thèse

Depuis l'avènement du séquençage à haut débit dans les années 2000, les possibilités d'exploration du fonctionnement du génome des organismes n'ont cessé de s'accroître. La production de données volumineuses est accessible en routine et plusieurs initiatives internationales permettent de mettre en commun et valoriser les données ainsi générées. La mise à disposition de ces données devrait en théorie permettre à n'importe quel chercheur d'être en mesure de reproduire les analyses publiées dans la littérature et d'interroger ces données pour ses propres problématiques biologiques. Cependant, la pléthore de procédures d'analyses disponibles et leur complexité en terme de nombre d'étapes, outils et paramètres, rendent ces tâches ardues. Cette complexité s'accroît inexorablement au fur et à mesure que de nouvelles technologies, méthodes, outils et données s'ajoutent à ceux déjà disponibles, et que les problématiques biologiques s'orientent de plus en plus vers le croisement et l'intégration de davantage de ces données hétérogènes et de provenances diverses.

Des pratiques et solutions pour apporter des éléments de réponse à cette problématique ont été développées de manière fragmentée par les communautés scientifiques et informatiques. Cependant, la non-existence d'une implémentation parfaite, universelle et consensuelle, ainsi que l'abondance de solutions semblant convenir mais qui ne survivent pas l'épreuve de certains cas pratiques maintiennent ce sujet dans l'actualité bioinformatique.

Sur le plan technique, l'objectif principal de ma thèse consistait à déterminer un ensemble d'outils et de méthodes permettant d'assurer la reproductibilité, l'automatisation, le partage, l'archivage, le développement simplifié, l'optimisation et la réutilisation d'analyses complexes de données volumineuses. Cet objectif générique a été couplé à des objectifs spécifiques de recherche fondamentale dans le domaine de l'épigénétique, et plus particulièrement dans l'étude de dynamiques épigénétiques de deux modèles biologiques : la thymopoïèse humaine et la spermatogénèse murine.

Les dynamiques épigénétiques étudiées au cours de la thymopoïèse s'inscrivent dans la poursuite des efforts de recherche dirigés par les docteurs Salvatore Spicuglia et Vahid Asnafi dans le cadre du projet BLUEPRINT afin de générer les cartographies épigénétiques des précurseurs lymphocytaires thymiques. L'analyse des données ChIP-Seq, bisulfite-Seq, RNA-Seq et ATAC-Seq produites pour différents stades de différentiation des thymocytes avait deux objectifs principaux. Premièrement, de replacer les paysages épigénétiques de ces thymocytes par rapport à toutes les populations issues de lignées hématopoiétiques disponibles dans les bases de données, ceci afin de s'assurer de la qualité et de la cohérence des données générées. Deuxièmement, de caractériser, à partir des données ainsi validées, les dynamiques épigénétiques au cours de la thymopoïèse en se concentrant sur les éléments régulateurs distaux (SECTION 2.1).

Les dynamiques épigénétiques étudiées au cours de la spermatogénèse ont pu être réalisées dans le cadre de projets de recherche dirigés par les docteurs Saadi

Khochbin et Sophie Rousseaux. L'objectif du premier projet était d'étudier un nouveau langage de modifications d'histones potentiel grâce à des ChIP-Seq des marques H4K5ac, H4K5bu, H4K8ac et H4K8bu réalisés aux stades pachytène et spermatides ronde (SECTION 2.2.1). Le deuxième projet avait pour objectif de lever le voile sur la dynamique d'éviction, de remplacement et de repositionnement des nucléosomes au cours de la spermiogénèse grâce à des données MNase-Seq générées aux différents stades de différentiation. La comparaison de ces données avec d'autres données générées dans deux modèles invalidant les gènes *Nut* ou *H2AL2* devait permettre de préciser le lien entre les deux protéines éponymes et la réorganisation de la chromatine (2.2.4).

Les besoins en analyses bioinformatiques pour ces différents projets biologiques, ainsi que pour d'autres projets auxquels j'ai participé pendant ma thèse (ANNEXE B), ont permis d'affiner le produit de la recherche méthodologique entreprise et documenté dans ce manuscrit (SECTION 3.1). Ils ont également nécessité le développement logiciel de certaines approches analytiques pouvant dorénavant être appliqués à d'autres problématiques (SECTIONS 3.2, 3.3 et 3.4).

2 Décryptage de mécanismes épigénétiques

2.1 Dynamiques épigénétiques des précurseurs lymphocytaires thymiques humains

La coordination de la différentiation des précurseurs thymiques humains par les différents mécanismes épigénétiques est un processus peu compris (SECTION 1.2.1.4). Afin de poser les bases du décryptage de ces mécanismes, des épigénomes de référence complets (TABLE 1.2) de sous-populations thymiques saines ont été produits à partir du tri de cellules thymiques obtenues à l'hôpital Necker-Enfants-Malades. Des données ChIP-Seq, RNA-Seq et Bisulfite-Seq ont ainsi été générées pour cinq sous-populations thymiques : *early immature* (CD34+), *early cortical* (CD4+ CD8+ CD3low), *late cortical* (CD4+ CD8+ CD3high), et *single positive* (CD4+ ou CD8+). Les analyses produites à partir de ces nouvelles données font l'objet d'un article en préparation (ARTICLE 1 ; CIESLAK, CHARBONNIER, TESIO et al. 2019).

En appliquant à ces données une approche d'analyse des lignées hématopoïétiques des grands consortiums réalisée par SANTA-PAU, JUAN, PANCALDI et al. 2017, mon analyse intégrative a permis de montrer que les paysages épigénétiques des sous-populations thymiques sont caractéristiques de leur identité biologique, avec une position intermédiaire entre cellules souches hématopoïétiques et lymphocytes T périphériques (FIGURE 1 de l'ARTICLE 1). L'étude des dynamiques épigénétiques au cours de la différentiation T a révélé que la déméthylation de l'ADN est un marqueur des régions régulatrices distales majoritairement indépendamment du statut d'activation de ces régions (FIGURE 2 de l'ARTICLE 1). Certaines régions distales hypométhylées et ouvertes de façon constitutive au cours des stades précoces de la différentiation (précurseurs immatures et stades corticaux) sont pourtant associées à des gènes différentiellement exprimés entre ces stades. L'enhancer α correspond au plus marqué de ces cas de figures (FIGURE 3 de l'ARTICLE 1). L'étude des mécanismes de régulations qui permettent d'expliquer pourquoi ce locus ne s'active que tardivement malgré son état ouvert et déméthylé a été réalisée par d'autres collaborateurs.

Epigenomic dynamics of human T-cell precursors reveal *TCRA* enhancer silencing by HOXA

Agata Cieslak,[1,8] Guillaume Charbonnier,[2,3,8] Melania Tesio,[1] Eve-Lyne Mathieu,[2,3] Mohamed Belhocine,[2,3] Aurore Touzart,[1,4] Guillaume Hypolite,[1] Guillaume Andrieu,[1] Joost H Martens,[5] Eva Janssen-Megens,[5] Marta Gut,[6] Ivo Gut,[6] Nicolas Boissel,[7] Arnaut Petit,[8] Denis Puthier,[2,3] Elizabeth Macintyre,[1] Hendrik G Stunnenberg,[5] Salvatore Spicuglia,[2,3,*] and Vahid Asnafi[1,9,*]

[1]Université Paris Descartes Sorbonne Cité, Institut Necker-Enfants Malades (INEM), Institut national de la santé et de la recherche médicale (Inserm) U1151, and Laboratory of Onco-Hematology, Assistance Publique-Hôpitaux de Paris, Hôpital Necker Enfants-Malades, 75015 Paris, France

[2]Aix-Marseille University, Inserm, Theories and Approaches of Genomic Complexity (TAGC), UMR1090, 13288 Marseille, France

[3]Equipe Labellisée Ligue Contre le Cancer

[4]Division of Cancer Epigenomics, German Cancer Research Center, Heidelberg

[5]Department of Molecular Biology, Faculties of Science and Medicine, Radboud Institute for Molecular Life Sciences, Radboud University, 6500 HB Nijmegen, Netherlands

[6]Centro Nacional de Análisis Genómico (CNAG), 08028 Barcelona, Spain

[7]Université Paris Diderot, Institut Universitaire d'Hématologie, EA-3518, Assistance Publique-Hôpitaux de Paris, University Hospital Saint-Louis, Paris, France.

[8]Department of Pediatric Haematology and Oncology, Assistance Publique-Hôpitaux de Paris, Hôpital Armand Trousseau, Paris, France.

[8]These authors contributed equally

[9]Lead contact

*Correspondence: salvatore.spicuglia@inserm.fr (SS), vahid.asnafi@nck.aphp.fr (VA)

Title: 10-12 words, should be no more than 85 characters (including spaces)

We: 12 mots and 85 characters

Article: the total character count under 45,000* (including spaces and main figure legends but excluding STAR Methods text, supplemental item legends, and References section)

Figures: no more than seven figures and/or tables.

SUMMARY

150 words. Currently : 149 words.

How epigenetic dynamics coordinates differentiation of human T cell precursors is not well understood. Here we present a comprehensive resource of epigenomics data and carried-out an integrative analysis along the other hematopoietic populations available from the BLEUPRINT consortium. Although T cell commitment is accompanied with a large scale of epigenetic changes, the majority of distal regulatory elements are constitutively unmethylated throughout T cell differentiation irrespectively of their activation status. Among these, we identified the enhancer of the *TCRA* gene (Eα) as being in an open and unmethylated chromatin structure before activation. Thorough gene expression analysis revealed that the HOXA5-9 transcription factors are down-regulated concomitantly to Eα activation. Furthermore, by interacting with ETS1, HOXA proteins bind to Eα and repress its activity. Accordingly, HOXA9 overexpression prevents *TCRA* rearrangements and imposes developmental bias towards γδ T cell lineage, highlighting the key role of HOXA proteins in the epigenetic control of *TCRA* rearrangements.

Keywords:

(up to 10)

T cell differentiation, epigenetics, DNA methylation, enhancer, *TCRA* gene rearrangements, *HOXA* genes, T-ALL

INTRODUCTION

T lymphocytes develop from a stepwise process of cell fate choices whereby distinct signaling pathways in the thymus cause hematopoietic precursors to become committed to the T cell fate, while mobilizing a T cell gene expression program that prepares the cells for T cell antigen receptor (TCR) expression, TCR-based repertoire selection, and long, versatile careers as immune effectors (Dik et al., 2005a; Spits, 2002). However major questions about the molecular mechanisms involved in this process and stage-specific regulation of T cell genes are not yet well defined. Moreover, despite numerous reports on the dynamic of epigenetics marks during mouse T cell differentiation (Pekowska et al., 2011; Zhang et al., 2012; Hu et al., 2018) we have a limited understanding of epigenetic mechanisms controlling human T cell differentiation. This is of crucial importance given the potential relevance for immune-related diseases as well as oncogenic transformation of T cell precursors.

Human T lymphocyte ontogeny in the thymus requires the ordered somatic recombination of V, D and J gene segments at the *TCR* loci to determine the development into either γδ or αβ T cell lineages (Dik et al., 2005a; Spits, 2002). While the *TCRD*, *TCRG* and *TCRB* loci rearrange at the early tCD34 stage, the induction of germline expression and rearrangements of the *TCRA* locus start at the early cortical (EC) stage and achieve high level of mature TCRA expression in late cortical (LC) and subsequent simple positive (SP) stages (Dik et al., 2005a). It has been previously shown that the Eα enhancer plays a primary and essential role in the tightly regulation of *TCRA* rearrangements and expression during αβ T cell differentiation (Bassing et al., 2003; Sleckman et al., 1997). The Eα enhancer is the only known enhancer associated with *TCRA* loci and has been shown to be both necessary and sufficient to provide lineage and stage specific *TCRA* rearrangements and expression. Namely, Eα activates transcription originating from the T early alpha (*TEA*) promoter located upstream of the Jα gene segments (100kb upstream Eα), which is essential to modify 5' Jα chromatin state for initial recombination events (Abarrategui and Krangel, 2009; Hawwari and Krangel, 2005; Villey et al., 1996). The Eα enhancer represents a paradigm of gene regulation whereby a nucleoprotein complex is assembled during T cell differentiation in order to drive timely and stage specific rearrangements of the *TCRA* locus (Giese et al., 1992, 1995; Roberts et al., 1997). However the factor(s) (either activator or repressor) governing the stage specific activation of Eα have remained elusive.

To investigate the epigenetic mechanisms guiding T cell identity, we analyzed epigenomic and transcriptomic data from the BLUEPRINT Epigenome Project (Stunnenberg and Hirst, 2016), including whole-genome bisulfite sequencing (WGBS) DNA methylation maps, as well as genome-wide maps for six histone modifications and RNA expression for the key thymic subpopulations of human T cell precursors. Integration with other available epigenomic data covering the development of the major branches of the hematopoietic system provided a compressive view of the epigenomic dynamics in play at crucial transitions of T cell commitment and development. We observed that the majority of distal regulatory elements are constitutively unmethylated throughout T cell differentiation irrespectively of their activation status. This was best exemplified by the enhancer of the *TCRA* gene (Eα) which was found in an open and unmethylated chromatin structure before activation.

Thorough gene expression analysis revealed that the HOXA5-9 transcription factors are down-regulated concomitantly to Eα activation. Furthermore, by interacting with ETS1, HOXA proteins bind to Eα and repress its activity, thereby preventing premature activation of Eα. Accordingly, HOXA9 overexpression prevents *TCRA* rearrangements and imposes developmental bias towards γδ T cell lineage, highlighting the key role of HOXA proteins in the epigenetic control of *TCRA* rearrangements. Abnormal expression of *TCRA* in early T cell precursors block differentiation and bias cell lineage choice towards T γδ lineage (ref), while epigenetic repression of Eα and subsequent inhibition of *TCRA* rearrangements by HOXA-like oncogenes contributes to leukemogenesis (Dadi et al., 2012). Thus, fine tune regulation of Eα activity by HOXA proteins appears essential for both normal T cell differentiation and prevention of oncogenic transformation.

RESULTS

A reference epigenome of human early T cell differentiation

As part of the BLEUPRINT project we generated reference epigenomes of sorted human thymocytes encompassing the main differentiation stages **(Figure S1A, S1B**), including immature double negative $CD34^+$ (tCD34: $CD34^+/CD3^-/CD4^-/CD8^-$), early cortical (EC: $TCR\alpha\beta^-/CD3^-/CD4^+/CD8^+$), late cortical (LC: $TCR\alpha\beta^+/CD3^{+low}/CD4^+/CD8^+$), single positive CD4+ (SP4: $TCR\alpha\beta^+/CD3^+/CD4^+/CD8^-$), single positive CD8+ (SP8: $TCR\alpha\beta^+/CD3^+/CD4^-/CD8^+$). Reference epigenomes comprised ChIP-seq for three histones marks positively associated with gene expression (H3K4me1, H3K4me3 and H3K27ac) and two with silent chromatin (H3K27me3 and H3K9me3), whole genome bisulfite sequencing (WGBS) and RNA-seq (**Table S1 to do**). Replicates from different individuals where merged and normalized to provide a consistent set of reference epigenome tracks for each differentiation stage.

To obtain a global view of the epigenomic landscape of human hematopoiesis we integrated the ChIP-seq data of each thymic subpopulations into 11 chromatin states derived from the 6 histone modifications (**Figure S1C**) and compared them with chromatin states of other hematopoietic populations generated by the BLEUPRINT and Roadmap consortia. Multiple Correspondence Analysis (MCA) based on reduced chromatin states (Carrillo-de-Santa-Pau et al., 2017) allowed us to create a multi-dimensional space were the different hematopoietic samples were placed based on their vectors of chromatin states across the genome. Clustering based on the first two components grouped samples from the same cell type while the main hematopoietic populations were clearly separated from each other, in agreement with previous results (Carrillo-de-Santa-Pau et al., 2017) (**Figure 1A**). Consistently, the first component reveals two main branches of differentiation, representing the myeloid and lymphoid cell lineages, while Hematopoietic Stem Cells (HSCs) localized in a central position. The second component appears to reflect the differentiation stage and environmental context of each cell type with respect to the HSCs. Interestingly, thymocyte populations clustered tight to peripheral T cells, except for the most immature T cell precursor (tCD34), which clustered close to HSCs and are not fully committed to the T cell lineage (Dik et al., 2005a) (**Figure 1A**).

To assess whether the epigenomic states were consistent with the known underlying biological function of each cell type, we analyzed the enrichment on biological process of

active enhancer states (**Figure S1C**). Enhancer regions found in each population were significantly enriched in gene pathways associated with the corresponding lineage (**Figure 1B**). For instances, thymic and peripheral T cells were associated with TCR signaling, B cells were associated with BCR signaling and myeloid lineages were associated with phagocytosis and innate immunity.

Lineage specific genes displayed the expected chromatin dynamics across hematopoietic populations (**Figure 1C**). The stem cell marker *CD34* was open in HSC and tCD34 early T cell precursor, the CD3 cluster of T cell receptors where open in T cell samples, the B cell master transcription factor *PAX5* was associated with open chromatin in B cells; the myeloid marker *CD33* was associated with open chromatin in HSC, monocytes and neutrophils but only transcribed in the myeloid lineage. Thus, the epigenomic states, including those from the newly generated thymocyte populations, were able to capture the main biological differences between cell types and were fully consistent with the known underlying biological processes.

We next investigated more in detail the epigenetics dynamics during human T cell differentiation. Chromatin states clearly separated the main stages of thymic T cell differentiation from HSC and peripheral T cells (**Figure 1D**), suggesting a specific chromatin signature of thymic T cell precursors. Indeed, while the first dimension (separating T cells from HSC and tCD34 precursors) was enriched on cell homeostasis and activation functions, the second dimension (separating the thymocyte populations from peripheral T cells and HSC) was specifically enriched on V(D)J recombination (**Figure S1D**). To have a global view of main epigenetics changes during T cell differentiation, we represented the chromatin state transitions among HSC, thymic subpopulations and peripheral T cells using a Sankey diagram (**Figure 1E**). We observed a progressive increase of heterochromatin regions along with a decrease of active enhancers and transcribed regions. This suggests a progressive loss of plasticity from HSC to mature T cells. Finally, master transcription factors that are either repressed (*SPI1*/PU.1, *TAL1*) or activated (*BCL11B*, *TCF1*/TCF7) during T cell commitment or required for specific CD4+ (*ZBTB7B*/THPOK) or CD8+ (*RUNX3*) T cell differentiation, displayed consistent epigenomic changes (**Figure 1F**).

Overall our analysis of hematopoietic reference epigenomes fairly recapitulates hematopoietic chromatin signatures and accurately captures the main biological identities of thymic T cell precursors showing that our epigenomic data provide an excellent resource to study the regulatory networks underlying early T cell differentiation in humans.

DNA demethylation is a hallmark of distal regulatory elements irrespective of their activation status in T cells

We next analyzed DNA methylation of the 5 thymic subpopulations along with the different hematopoietic populations using WGBS data from both BLEUPRINT consortium. Similar to the chromatin state classification, dimensional reduction based on DNA methylation allows to group the different hematopoietic samples according to their cell lineage, while thymocyte populations where clearly separated from peripheral T cells (**Figure S2xx**).

Gene distal DNA hypomethylation has been shown to be a hallmark of enhancer regions, providing an accurate resolution of the actual regulatory sequences (Stadler et al., 2011). Therefore, to assess the dynamic of enhancer activities during early T cell differentiation, we retrieved all distal regions found hypomethylated in at least one thymic subpopulation and clustered the regions based on the dynamic of DNA methylation (**Figure 2A**). The first four clusters corresponded to constitutively hypomethylated regions and represented the vast majority of distal hypomethylated regions. Cluster 5 corresponded to regions preferentially hypomethylated in tCD34 precursors. Regions of clusters 1 to 5 were poorly enriched in H3K27ac (**Figure 2B**), expressed at low levels (**Figure 2C**) and were either associated with genes related to hematopoietic lineages other than T cells or not specifically enriched in relevant biological processes (**Figure 2D**). In contrast, cluster 6 to 8 corresponded to regions hypermethylated in tCD34 thymocytes that get progressive demethylation at different T cell differentiation stages. Consistently, clusters 6 to 8 where associated with high levels of H3K27ac correlated with the hypomethylation status (**Figure 2B**) as well as genes displaying high expression levels in thymocytes (**Figure 2C**) and related to T cell specific functions (**Figure 2D**).

We were surprised by the high proportion of constitutively hypomethylated regions, which represented 77% of all distal hypomethylated regions, and decided to explore more in detail the epigenomic dynamics of these potential regulatory regions. Clustering of distal constitutively hypomethylated DNA regions based on H3K27ac levels (**Figure S2xx**) revealed 5 clusters with dynamic H3K27 acetylation (**Figure 2E-F**), representing 17% of the total set of distal hypomethylated regions. Changes of H3K27ac levels roughly follow the expression dynamics of neighbor genes (**Figure 2G**). Consistent H3K27ac at promoter and distal constitutively hypomethylated regions were observed for early (*RAG1-RAG2*) and late (*CD44*) expressed loci. The *Il7R* locus was associated with distal enhancers activated either at

the early (downstream) or the late (upstream) stages, in both cases the regulated enhancers were constitutively demethylated (**Figure 2H**).

These observations strongly indicate that DNA hypomethylation marks distal regulatory elements independently of their activation status, both preceding and remaining after enhancer activation.

Chromatin opening precedes enhancer activation

The above analyses suggested that the activation onset of distal regulatory elements might be preceded by DNA demethylation. This raised the question of whether pre-marked (i.e., DNA hypomethylated) enhancers are already in an open chromatin configuration (i.e., nucleosome free). To address this issue, we performed Assay for Transposase-Accessible Chromatin (ATAC) experiments in tCD34 and EC stages, as theses stages reflected the major epigenetic transitions in early T cell differentiation (**Figures 1 & 2**). We identified 2084 and 2905 distal ATAC peaks in tCD34 and EC, respectively (**Figure 3A-B**). Around one third of the distal ATAC peaks (1108) where shared between the two thymic populations. Distal ATAC peaks present in both tCD34 and EC stages and specific to EC where in proximity of genes associated with T cell differentiation and function, while tCD34 specific peaks where associated with more general terms (**Figure 3C**). Consistently, analysis of conserved transcription factor binding sites shows a preferential association of hematopoietic factors such as TAL1 and AML1 (RUNX1) in tCD34 associated peaks, while transcription factors involved in T cell differentiation and activation such as E47, STAT and ETS were preferentially found in EC associated peaks (**Figure 3D**).

As predicted, ATAC peak regions were found constitutively demethylated in tCD34 and EC stages independently of whether the region was specifically open in tCD34 or EC thymocytes (**Figure 3E**). This suggests that DNA demethylation both precedes enhancer opening and remains after enhancer closing. Thus, DNA demethylation likely represents a primary marking of distal regulatory regions independently of the actual enhancer activity or chromatin accessibility.

Next, we asked whether constitutively open and demethylated regions could be associated with changes in enhancer activity. For this purpose, shared ATAC peaks were ordered in function of the H3K27ac ratio between tCD34 and EC thymocytes (**Figure 3F**). Interestingly, ATAC regions with high levels of H3K27ac in tCD34 or EC (**Figure 3G**) were associated with significant decreased and increased expression of neighbor genes, respectively (**Figure**

3H; paired Wilcoxon test) and concordant proportion of differentially regulated genes (**Figure 3I**). This suggests that constitutively open and demethylated distal regions can be associated with locus activation or repression.

To identify potential interesting loci, we searched for genes strongly induced between tCD34 and EC/LC thymocytes and associated with distal open regions displaying a gain of H3K27ac (**Figure 3G**). Strikingly, we identified the *TRAC* transcript, associated to the *TCRA,* locus as the highest induced gene associated with an H3K27ac enhancer gain (**Figure 3J**), which corresponded to the known Eα enhancer (Hawwari and Krangel, 2005; Ho et al., 1989; Winoto and Baltimore, 1989) (**Figure 4A**).

The *TCRA* enhancer (Eα) is in an open but epigenetically silent configuration in immature thymocytes.

To better understand the tight regulation of *TCRA* rearrangements, we visualized the epigenetic modulations at the *TCRA* locus during human thymopoiesis (**Figure 4A**). In tCD34 thymocytes, the *TCRA* locus is in a relative close configuration with low levels of H3K4me1, H3K4me3 and H3K27ac and relative high levels of H3K27me3. In contrast, EC and LC cortical thymocytes displayed overall high levels of H3K4me1, H3K4me3, H3K27ac and H3K36me3, whereas H3K27me3 was strongly reduced. Note that active histone marks were shifted from the 5' to the 3' side of the locus between EC and LC stages, in agreement with the progressive Vα-to-Jα rearrangements (Hawwari and Krangel, 2005; Huang et al., 2002; Krangel et al., 2004). Note also that the adjacent *TCRD* locus displayed the opposite epigenetics pattern, as expected (**Figure 4A**). In contrast, the Eα enhancer was found to be in an open chromatin configuration, even at the very immature stages of thymic maturation (ATAC), fully demethylated (BS-seq) and enriched in H3K4me1 (a mark of poised enhancers) (**Figure 4A-B**). Whereas in EC and LC stages, the Eα region gained substantial levels of H3K27ac. Consistently, H3K4me3, previously shown to be associated with highly active enhancers (Pekowska et al., 2011), was found preferentially enriched at Eα in the later thymic populations.

Our findings are consistent with previous work in mouse showing that Eα is occupied by transcription factors from the DN stage of thymocyte development, thus preceding transcriptional activation of *TCRA* locus taking place after β-selection (Hernandez-Munain et al., 1999; Spicuglia et al., 2000). We have previously suggested that Eα enhancer activation

occurs via a conformational change of a pre-assembled nucleo-protein complex (Cauchy et al., 2016; Spicuglia et al., 2000). Analysis of published ChIP-seq experiments for lymphoid transcription factors ETS1, RUNX1, E47, GATA3 and Ikaros confirmed that Eα is already occupied by these activating transcription factors in murine DN thymocytes (**Figure S4A**). Importantly, a short non-coding enhancer RNAs (eRNA), the hallmark of active enhancers was only observed in murine DP thymocytes (**Figure S4B**). Additionally, as shown by ChIP-qPCR assays (**Figure 4C-D**). Eα is already bound by its activating transcription factors ETS1 and RUNX1 in the human immature tCD34 thymocytes. This was also confirmed for RUNX1 by ChIP-seq experiment (**Figure 4E**).

Overall, these data suggest that Eα is found in an open chromatin state and is already bound by its activating transcription factors from the very immature stages of human and murine thymic maturation. As Eα remains inactive until the cortical stages, this suggests repressive mechanism acting on its activity in early stages of thymic maturation in mouse and human. While, the function of Eα has been intensively studied (Carey, 1998; Hawwari and Krangel, 2005; Ho et al., 1989, 1990; McMurry and Krangel, 2000), the molecular mechanism conveying stage specific activation of Eα have remained elusive. Thus, Eα control of *TCRA* locus represents an unsolved paradigm of tissue specific gene regulation by distal regulatory elements.

The *HOXA* locus is progressively repressed during early T cell differentiation.

Given that Eα is already bound by key transcription factors required for its activation, we hypothesized that loss of a transcriptional repressor might explain the stage specific activation of Eα and subsequent *TCRA* rearrangements. To explore this possibility of repressive mechanism acting on Eα we performed RNA-seq of main human thymic subpopulations. Unsupervised gene expression analysis identified 19 clusters, each of them containing transcripts with different types of gene expression modulation during thymic maturation (**Figure 5A**). Cluster C13, containing 893 different transcripts, presented the expected profile of gene expression modulation with high expression at early stages and gradual reduction during maturation until the extinction of the expression after β-selection stage (ISP > EC > LC) (**Figure 5A**). Interestingly, this cluster contained some members of HOX-like (HOXL) subclass of homeodomain (Hox) transcription factors (**Figure 5B, S5A, and S5B**) As ectopic expression of HOX-family of transcription factors (TLX1 and TLX3 from NKL subclass) was reported to exert repressive activity on Eα in T-acute lymphoblastic leukemia (T-ALL) (Dadi

et al., 2012), we focused on HOXL genes normally expressed during thymopoieses. TaqMan Low-Density Array (TLDA) was designed to evaluate the expression of all members of HOXL subclass in an independent set of human thymic subpopulations (**Table S1**). Unsupervised clustering of TLDA array showed that *HOXA5*, *HOXA6*, *HOXA7*, and *HOXA9* genes best fit into the expected expression profiles (**Figure 5C, S5C-E**). Of note, HOXA9 appears to be a good candidate as its downregulation perfectly coincides with Cα upregulation, presence of TEA-Cα and beginning of *TCRA* rearrangements (**Figure 5D-H**).

HOXA5-9 proteins repress Eα activity via their homeodomain.

To test the possibility that HOXA5-9 proteins interfere with Eα transcriptional activity, we used a gene reporter assay in which the expression of chloramphenicol acetyltransferase (CAT) is under the control of Eα (Eα-CAT) (Giese et al., 1995). The ectopic expression of all HOXA-5, -6, -7 and -9 repressed Eα-regulated CAT expression by approximately 5-fold (**Figure 5I**). Next, we evaluated the level Eα-CAT repression of the truncated forms of HOXA proteins, without their homeodomains (ΔHD). As shown in Figure 5I, all truncated HOXAs forms (ΔHD) exerted a significantly reduced repressive activity compared to their respective full-length (FL) proteins. The reduced activity of HOXA5-9 ΔHD proteins was not due to their lack of nuclear localization because, as their FL counterpart, they were mainly localized in the nucleus (**Figure SF**). We conclude that HOXA5-9 proteins repress *in vitro* Eα transcriptional activity in a homeodomain-dependent manner.

Chromatin immunoprecipitation (ChIP) assays using Loucy (HOXA overexpressing) or HeLa (HOXA negative) cell line nucleotransfected with Flag-tagged expressing vectors for full-length *HOXA5, -6, -7, -9*, showed significantly enriched Eα DNA (**Figure 5J, 5K**). Importantly, HOXA proteins without their homeodomain (ΔHD) show reduced binding to Eα. As Eα lacks AT-rich DNA motif TAATNA characteristic for HOX homeodomain binding, we hypothesized, that HOXA proteins exert their Eα repressive activity by interacting with ETS1, as it was showed for TLX proteins, members of NKL subclass of homeodomain proteins (Dadi et al., 2012). Co-immunoprecipitation assays using human thymocytes and anti-HOXA9 antibody recovered both ETS1 and HOXA9 in the precipitated material (**Figure 5L** upper panel). Co-IP assays using Loucy cell line, which co-express HOXA9 and ETS1, showed also that immunoprecipitation performed with anti-ETS1 antibody results in the co-immunoprecipitation of HOXA9 (**Figure 5L** lower panel). Additionally, Proximity Ligation

Assay (PLA) confirmed HOXA9-ETS1 interaction in Loucy cell line and immature $CD34^+/CD3^-$thymocytes (**Figure 5M**). More mature $CD34^-/CD3^+$ thymocytes, as expected, showed no HOXA9-ETS1 (**Figure 5M**).

Taken together, these data suggest that HOXA9 transcription factor could be recruited to Eα and repress Eα activity in a homeodomain-dependent manner.

HOXA9 overexpression imposes developmental bias towards γδ T cell lineage.

To functionally evaluate the role of HOXA homeoproteins in the *TCRA* rearrangement and T cell development we transduced human CD34+ umbilical cord blood cells (UCB CD34+) with either a HOXA9 over-expressing vector (HOXA9-GFP) or a control vector (GFP) and co-cultured *in vitro* on OP9-DL1 stroma under conditions for T cell differentiation. After 7 weeks of co-culture we found more TCRγδ expressing cells when HOXA9 was over-expressed (22.0±11.6%) compared to control cells (10.8±4.6) (**Figure 6A, 6B**). The difference was significantly pronounced after 8 weeks of co-culture (HOXA9-GFP 26.8±15.3 *vs* control GFP 6.8±3.1%) (**Figure 6A, 6B**). Next, by multiplex fluorescent PCR we analyzed the *TCRA* rearrangements at different time points of co-culture. In control conditions we detected the first *TCRA* rearrangements at 4^{th} week of co-culture with *TCRA* polyclonal profile present at 5^{th} week of culture (**Figure 6C**). However, in HOXA9 overexpressing conditions we observe much less of *TCRA* rearrangements, with the difference clearly observed at 5^{th} week of culture (**Figure 6C**).

To confirm the role of HOXA9 protein in *TCRA* rearrangements and T cell development we performed *in vitro* and *in vivo* experiments in mice. To this purpose, murine DN thymocytes were transduced with either a control vector (GFP) or a HOXA9 over-expressing vector (HOXA9-GFP) and differentiated *in vitro* on OP9-DL1 stromal cells (**Figure 6D**). In control conditions and following 9 days of co-cultures, 44.4±6.8 of the immature thymocytes developed into TCRαβ expressing cells. When HOXA9 was overexpressed, however, only few thymocytes (14±7%) differentiated into TCRαβ cells and they develop instead into TCRγδ lineage (**Figure 6E-F**). To confirm these data *in vivo*, we transplanted HOXA9 overexpressing DN thymocytes and their control counterparts into sub-lethally irradiated $Rag2^{-/-}\gamma c^{-/-}$ mice (**Figure 6G**). Whereas following 3 weeks post-transplants control cells became TCRαβ positive, HOXA9 thymocytes repressed TCRαβ expression and acquired the TCRγδ on their cell surface (**Figure 6H-I**).

Taken together, these results indicated that the enforced expression of HOXA9 blocks *TCRA* rearrangements, thus diverting T cell development from the αβ to the γδ lineage.

Homeodomain protein deregulation leads to TCRγδ bias in human T-ALL.

In an effort to evaluate the implication of the homeoprotein mediated inhibition of Eα activity in the αβ *vs* γδ lineage bias, we then take advantage of the T cell acute lymphoblastic leukemia (T-ALLs) model. T-ALL is a rare disease resulting from the leukemic transformation of thymic cell precursors and their arrest at specific stages of differentiation (Asnafi et al., 2003; Ferrando et al., 2002). Approximately 30% of T-ALLs have a "mature" phenotype expressing a sCD3/TCR. Intriguingly among these T-ALLs with the expression of a surface TCR an unexpected fraction of cases (≅50%) expresses a TCRγδ and molecular mechanism of this bias remain unidentified (Asnafi et al., 2004; Trinquand et al., 2016). We analyzed a series of 150 TCR-expressing T-ALLs (TCRαβ T-ALLs, n=45 *vs* TCRγδ T-ALLs, n=95) and evaluated HOXA9 expression. The TCRγδ T-ALLs show higher HOXA9 expression compared to TCRαβ T-ALLs (**Figure 6J**). 39/95 (41%) TCRγδ expressing T-ALLs demonstrated HOXA9-overexpressing. Additionally, 32/95 (34%) TCRγδ T-ALLs overexpress homeodomain genes from the NKL subclass (ie *TLX1* or *TLX3*). In total, 71 out of 95 (75%) TCRγδ T-ALLs demonstrate deregulation of homeodomain genes (HD-deregulated), compared to 5/45 TCRαβ T-ALLs ($p<0.0001$). We then hypothesized that HD-deregulation could impose TCRγδ expressing in αβ-lineage oriented precursors by preventing TCRαβ expression. Supporting this, pre-T cell receptor alpha (pTα) was significantly more expressed in HD-deregulated TCRγδ T-ALLs as compared to TCRγδ T-ALLs without HD deregulation ($p=0.04$) (**Figure 6K**), contrary to *SOX13* which is less expressed in the HD-deregulated TCRγδ T-ALLs ($p=0.003$) (**Figure 6L**). Next, we analyzed *TCRB* rearrangements in TCRγδ-expressing T-ALLs and demonstrated an unexpected complete VDJ *TCRB* rearrangement in the group presenting HD deregulation (45/71 *vs* 3/24; $p<0.0001$) (**Table 1**). These data suggest that the molecular basis for the bias to TCRγδ expression in T-ALL could be related to homeodomain proteins (HOXA, TLX1/3) deregulation preventing Eα activation and *TCRA* rearrangements.

DISCUSSION

Here, we have taken the advantage of the generation of a comprehensive epigenomic resource encompassing the main thymic populations of human T cell precursors to shed light on the epigenomic dynamics of a well-defined developmental program. First, we found that the epigenetic landscape positions thymocytes populations between HSC and mature T cells within the lymphoid lineage. T cell differentiation was accompanied by huge remodeling of histone modifications involving a progressive closing of the chromatin landscape clearly reflecting the Waddington model of cellular differentiation (Waddington, 1957). Second, our analysis revealed an unexpected consistency/stability of DNA demethylation at distal regulatory regions. Indeed a high proportion of distal regulatory regions were found to be constitutively hypomethylated irrespectively of their activation status in T cells, thus suggesting that DNA hypomethylation might work as a main epigenetic hallmark of enhancers pre-assembly and memory. Finally, we disentangle the regulatory mechanisms leading to the stage-specific activation of the *TCRA* enhancer, Eα, a paradigm of enhancer organization and function (Carey, 1998). The Eα enhancer is found in a hypomethylated and open chromatin configuration prior the activation of the *TCRA* locus. We demonstrated that the HOXA proteins function as developmentally-regulated repressors of Eα, whereby contributing to the γδ *versus* αβ T cell lineage choice.

During T cell thymopoiesis, Eα regulates chromatin structure of *TCRA* gene segments by its ability to recruit a unique combination of transcription factors and induction of modification of histone marks (McMurry and Krangel, 2000). Active Eα at DP stage is characterized by the presence of known histone marks of active enhancers as H3K27ac, K3K4me1, H3K4me3 and the presence of histone variant H2AZ (Barski et al., 2007; Creyghton et al., 2010; Pekowska et al., 2011). Here we demonstrate that Eα is already bound by its activating transcription factors, but stays non-functional at the earliest stages of thymic maturation. Epigenetically, Eα presents characteristics of poised enhancers defined by enrichment of H3K4me1, depletion of H2K3me3 and H3K27ac. Additionally, Eα shows the presence of H3K27me3, suggesting a repressive mechanism acting on its activity. H3K27me3 and to a lesser extent H2K9me3 were found at some poised enhancers in embryonic stem cells (Rada-Iglesias et al., 2011; Zentner et al., 2011). However, some studies show modest correlation between poised enhancers and the presence of H3K27me3 (Ernst et al., 2011; Pekowska et al.,

2011). It is possible, that repressive histone marks are specific to a subset of poised enhancers required during development and/or cell differentiation.

HOXA proteins belong to the family of genes sharing a characteristic homeodomain protein fold consisting of 60-amino acids helix-turn-helix structure responsible for DNA binding and interactions with other proteins. HOXA functions are well established in antero-posterior axis definition of body segment identity specification during embryogenesis (Carroll, 1995; Goodman, 2002; Lewis, 1978). Importantly, they also play a key role in controlling cell identity and differentiation of hematopoietic stem cells and progenitors (Lawrence et al., 1996; Magli et al., 1997). *HOX* genes are highly expressed in HSCs and progenitors and their expression is silenced as cells become fully mature. Additionally, progressive down-modulation of *HOXA* transcripts was also reported during thymic cell maturation (Taghon et al., 2003). Here we reported that HOXA proteins repress the Eα activity in homeodomain-dependent manner. All HOX homeodomains bind highly similar AT-rich DNA motifs (Berger et al., 2008; Mann et al., 2009; Noyes et al., 2008). This TAATNA recognition sequence seems critically important for DNA binding as genome-wide HOXA9 binding sites in transformed myeloblasts found that >98% of sites contain a HOX motif (Huang et al., 2012). The additional specificity of HOXA proteins is achieved through a combination of motif affinity, interaction with cofactors and context-specific chromatin accessibility (Choo et al., 2011; Slattery et al., 2011). However, HOX motif is absent in the core Eα suggesting that HOXA binding to Eα is mediated by interactions with components of Eα enhanceosome. Suppressive function for HOXA proteins on Eα activity could be mediated by ETS1 as it was shown for TLX1/TLX3 proteins (Dadi et al., 2012), although we do not exclude protein-proteins interaction with and mediation of a suppressive effect by other transcription factors as RUNX1 (Hollenhorst et al., 2009).

HOXA9 is the most preferentially expressed gene in $CD34^+$ HSCs and early hematopoietic progenitors and is subsequently downregulated during differentiation. Moreover, *HoxA9* is known to impact mouse thymocyte differentiation and its knock-down the most dramatic hematopoietic phenotype from all *HOX* genes (Izon et al., 1998; Lawrence et al., 1997). $Hoxa9^{-/-}$ mice show an increase in the percentage of DP $TCR\alpha\beta^{hi}$ cells, total number of TCRαβ positive cells and accelerated progression of DP cells to mature SP $CD4^+$ and $CD8^+$ when compared with wild-type cells (Izon et al., 1998). This suggests that lifting the repression exerted by HOXA9 on *TCRA* enhancer accelerated the maturation of thymocytes. Inversely, our HOXA9 overexpression experiments showed the bias of the lineage choice

towards TCRγδ expressing cells in *in vitro* and *in vivo* experiments. This suggests the important role of physiological repression of Eα activity by HOXA proteins on the thymocyte development and lineage choice.

Blockage of Eα activity and subsequent lack of *TCRA* rearrangement during thymic differentiation can have severe implication. The cortical thymic maturation arrest observed in T-ALL that overexpress TLX1 and TLX3 proteins is caused by binding of TLX1/TLX3 to Eα via their interactions with ETS1 leading to reduced enhancer activity and consequently inhibition of *TCRA* gene segment recombination and maturation arrest of DP thymocytes and leukemia development (Dadi et al., 2012). This maturation blockage can be overcome by TLX1/3 abrogation or by downstream TCRαβ expression, which leads to *TCRA* rearrangement and apoptosis. The deregulation of HOX genes has also been reported in acute leukemias (Alharbi et al., 2013), especially when MLL translocations are involved. MLL fusion proteins constitutively upregulate *HOXA9* expression (Ferrando et al., 2003). In addition, *HOX* genes can be involved in T-ALL through translocation *PICALM-MLLT10*, *SET-NUP214* or following chimeric fusion protein of *HOXA* locus with T cell receptor (HOXA-TCR) (Dik et al., 2005b; Soulier et al., 2005; Speleman et al., 2005; Van Vlierberghe et al., 2008). These T-ALLs represent mainly immature or γδ lineage phenotype, suggesting the potential impact of *HOXA* genes overexpression and their inhibiting function on Eα activity in the development of these leukemias.

Taken together, endogenous HOXA5-9 transcription factors restrain the enhancing activity of a fully assembled Eα enhanceosome in DN thymocytes; prevent premature *TCRA* recombination, thereby providing the tight epigenetic control of Eα activation and αβ vs γδ developmental cell fate.

ACKNOWLEDGMENTS

We thank Cell Imaging Platform of Institut des Maladies Génétiques Imagine and the Transcriptomics and Genomics Marseille-Luminy (TGML) sequencing platform. We thank the Pediatric Cardiac Surgery of Hôpital-Necker-Enfants-Malades for the human thymic samples. We thank Cell Therapy Unit of Hôpital Saint-Louis for umbilical cord blood samples. The authors thank all participants in the GRAALL-2003/05 study groups for collecting and providing data and samples of T-ALL patients.

We thank Christoph Plass from German Cancer Research Center in Heidelberg for reading the article and constructive criticism.

A.C. was supported by grants from the Association pour la Recherche sur le Cancer (ARC), the Fondation pour la Recherche Médicale (FRM) and BLEUPRINT projet. M.T was supported by grants from Fondation de France (FdF) and ARC.

Work in the laboratory of S.S. was supported by recurrent funding from INSERM and Aix-Marseille University and by specific grants from the European Union's FP7 Programme (agreement N° 282510-BLUEPRINT), the Foundation for Cancer Research ARC (ARC PJA 20151203149) and A*MIDEX (ANR-11-IDEX-0001-02), Plan Cancer 2015 (C15076AS), PlBio-INCA, and Equipe Labellisée Ligue contre le Cancer.

AUTHOR CONTRIBUTIONS

A.C., M.T., G.C. S.S. and V.A. wrote the manuscript. A.C., M.T., and E-L.M. performed research and data analysis. G.C and M.B. and S.S. performed bioinformatics analysis. J-H.M, E.J-M., M.G., I.G., P.F, and H-G.S are members of BLEUPRINT consortium. All authors validated the manuscript. V.A. and S.S. oversaw conceptual development of the project.

Aurore + Guillaume A

DECLARATION OF INTEREST

The authors declare no competing interests.

FIGURE TITLES AND LEGENDS

Figure 1: Reference epigenomes of human hematopoietic and T cell precursors

(A) MCA of human hematopoietic samples based on chromatin states. Ovals highlight the major subtypes of hematopoietic lineages as described in **Table to do**

(B) Top enriched MSigDB pathways associated with active enhancers found in each hematopoietic population. Vertical lines delineate the main hematopoietic lineages.

(C) Chromatin state profiles of representative genes of the different hematopoietic lineages. The color code is indicated at the right of the panel.

(D) MCA of HSC and human T cell populations based on chromatin states.

(E) Sankey diagram depicting the main epigenomic transitions between the HSC and the human T cell populations. The color code represents the merged chromatin states as indicated at the right of the panel.

(F) Chromatin state profiles of representative genes specifically expressed at different stages of T cell differentiation. The color code is as in C.

See also **Figure S1**.

Figure 2: DNA demethylation is a hallmark of distal regulatory elements irrespective of their activation status in T cells

(A) Heatmap displaying DNA methylation scores of distal hypomethylated regions at the indicated thymocyte populations based on WGBS. KM classification defined eight hypomethylated clusters. The scale of DNA methylation is indicated at the bottom of the panel.

(B) H3K27ac profiles at the indicated thymocyte populations centered on the hypomethylated clusters defined in A.

(C) Median of RNA-seq expression at the indicated thymocyte populations of genes associated with hypomethylated clusters defined in A.

(D) Top GO biological processes enriched at genes associated with hypomethylated clusters defined in A. Relevant hematopoietic and T cell related terms are highlighted.

(E) KM clustering of DNA hypomethylated regions with dynamics H3K27ac levels among thymocytes subpopulation. The heatmap of H3K27ac levels at the indicated thymocytes subpopulation is indicated.

(F) H3K27ac, H3K4me1 and WGBS profiles at the indicated thymocyte populations centered on the clusters defined in E. Same color code as in B.

(G) Z-score of median RNA-seq expression and H3K27ac signal compared at the indicated thymocyte populations of genes associated with clusters defined in E.
(H) Chromatin states, H3K27ac and WGBS profiles of representative loci displaying constant hypomethylation, but dynamic H3K27ac enrichments at the indicated thymocyte populations (red rectangles). The regulation of the associated gene is indicated at the top of each panel specifically expressed at different stages of T cell differentiation. The color code is as in C.
See also **Figure S2**.

Figure 3: Chromatin opening precedes enhancer activation
(A) Overlap between ATAC peaks identified in tCD34 and EC thymocytes.
(B) Clustered heatmap of ATAC signal in tCD34 and EC thymocytes (left panels)
(C) Top GO biological processes enriched in genes associated with distal ATAC peaks specific of tCD34 or EC thymocytes or common to both.
(D) Conserved TFBS enriched at distal ATAC peaks specifics of tCD34 or EC thymocytes or common to both.
(E) Average profiles of WGBS signals corresponding to the ATAC peaks defined in (A).
(F) Heatmap of H3K27ac signals based on common ATAC peaks ordered by the H3K27ac ratio between tCD34 and EC.
(G) Average H3K27ac profiles centered on ATAC peaks with relative high H3K27ac signal in tCD34 or EC thymocytes or common to both, as defined in E.
(H) Violin and box plots of RNA-seq expression at the indicated thymocyte populations of genes associated with distal ATAC peaks defined in E. Left panels represents data from (Casero et al., 2015).
(I) Percentage of significantly regulated genes associated with distal ATAC peaks defined in E,
(J) Top significantly induced genes associated with ATAC peaks and H3K27ac gain. The expression fold change between tCD34 (Thy3) and EC/LC (Thy4) thymocytes is indicated.
See also **Figure S3**.

Figure 4: Epigenetic and transcription factor analysis of the *TCRA* locus in the human thymic subpopulations
(A) ChromH, ATAC-seq, ChIP-seq profiles for histone modifications: H3K27ac, H3K4me1, H3K4me3, H3K36me3, H3K27me3 and BS-seq at the *TCRA/D* locus in human thymic

subpopulation tCD34, EC and LC. The zoom on the *TCRA* enhancer (Eα) is presented on the right.

(B) ATAC sequencing profiles for *TCRA* locus (*TRAJ* gene segments and enhancer alpha in Human thymic subpopulations $CD34^+1a^-$, $CD34^+1a^+$, EC, LC

(C) ChIP-qPCR analysis of ETS1 binding to Eα in human thymic subpopulation tCD34, EC and LC. ChIP enrichment level is presented as % of input and corresponds to the ratios between the Eα and IgG signal.

(D) ChIP-qPCR analysis of RUNX1 binding to Eα in human thymic subpopulation tCD34, EC and LC. ChIP enrichment level is presented as % of input and corresponds to the ratios between the Eα and Actin signal. IgG isotype control was performed to asses absence of non-specific Eα ChIP enrichment. ChIP experiments are representative of at least two independent experiments.

(E) RUNX1 ChIP-seq profiles in thymic CD34+ population (tCD34).

ATAC: Assay for Transposase-Accessible Chromatin, BS-seq: bisulfite sequencing; tCD34: thymic double negative $CD34^+$, EC: early cortical, LC: late cortical.

See also **Figure S4**.

Figure 5: HOXA5-9 proteins are progressively repressed during human thymopoiesis and bind to Eα and repress its activity.

(A) Unsupervised RNA-seq gene expression classification identifying 19 clusters with different profiles of gene expression modulations.

(B) Cluster 13 identifies 85 transcription factors which expression decreases around β-selection (ISP>EC>LC). HOXA5, HOXA7, HOXA9 and HOXA10 transcription factors are highlighted.

(C) TLDA analysis of HOXL expression profiles in thymic subpopulations. The expression level is classified from blue to grey. Only genes that show expression at any of the thymic subpopulation are presented.

(D) RQ-PCR analysis of *HOXA9* gene expression in human thymic subpopulations. Results are represented relative to the *ABL1* housekeeping gene.

(E) RQ-PCR analysis of *Cα* gene expression in human thymic subpopulations. Results are represented relative to the *ABL1* housekeeping gene.

(F) RQ-PCR analysis of *TEA-Cα* gene expression in human thymic subpopulations. Results are represented relative to the *ABL1* housekeeping gene.

(G) RQ-PCR analysis of *TCRA* germline from normal human thymocyte subpopulations Normalization was performed with the Albumin housekeeping gene.

(H) Fluorescent PCR Genescan analysis of Vα-Jα rearrangements detected in sorted thymic populations.

(I) Eα-CAT fold repression in HeLa cells following transfection with encoding vectors for ETS1, RUNX1, LEF1 and individual HOXA5-9 (full length and homeodomain-deleted forms). TLX1 is used as a control positive of repression. Eα-CAT signals were normalized to those of control cells transfected with an empty GFP vector. Data are presented as means of at least triplicate measurements with error bars to represent SEM.

(J) Anti-Flag ChIP-QPCR analysis of HOXA5-9 (full length and ΔHD forms) binding to Eα in Loucy cell line expressing Flagged HOXA5-9. ChIP enrichment level is presented as percentage of input and corresponds to the ratios between the Eα and Actin signal. ChIP with empty NTAP vector was performed as a control. Results represent means and SEM of triplicate reactions.

(K) Anti-Flag ChIP-QPCR analysis of HOXA5-9 (forms full length and ΔHD) binding to Eα in HeLa cells were transfected with ETS1, RUNX1, LEF1, TCRA-CAT and Flagged HOXA5-9 (full length and ΔHD forms) expressing vectors. Enrichment at Eα at substrate TCRA-CAT vector is showed relative to input substrate vector DNA and normalized to empty vector control.

(L) Upper panel: Co-immunoprecipitation (Co-IP) using HOXA9 antibody followed with immunoblotting with anti-ETS1 antibody in human thymocytes. The input corresponds to the 1% of cell extract used in the CoIP experiment. FT: flow-through; LW: last wash. A representative of two independent experiments is shown. Lower panel: immunoprecipitation (IP) using anti-ETS1 antibody followed by immunoblotting with anti-HOXA9 antibody in Loucy cell line. The input corresponds to the 1% of cell extract used in the CoIP.

(M) Duolink Proximity Ligation Assays (PLA) and confocal microscopy analysis of cells labeled with anti-ETS1 (Alexa 488, purple) and anti-HOXA9 (Alexa 555, green) antibodies in Loucy cell line and $tCD34^+/CD3^-$ and $tCD34^-/CD3^+$ thymocytes. Bars, 10µm.

tCD34: thymic double negative $CD34^+$, ISP: immature simple positive: EC: early cortical; LC: late cortical; SP4: simple positive CD4; SP8: simple positive CD8; TLDA: TaqMan Low Density Array.

See also **Figure S5**.

Figure 6: HOXA9 overexpression biases the T cell development in humans and mice.

(A) Representative FACS plots representing TCRγδ expressing cells during T cell differentiation of human CD34+ UCB cells transduced with HOXA9-GFP expressing vector or control GFP vector. Cells represent CD45+CD7+CD1a+ populations.

(B) Quantification of TCRγδ positive cells (as in A). Data are presented as means of three independent experiments with error bars as SEM.

(C) Fluorescent PCR Genescan analysis of *TCRA* rearrangements detected in differentiated CD34+ UCB cells.

(D) Schematic representation of the *in vitro* experimental design in mice.

(E) Representative FACS plots and quantification of TCRαβ and TCRγδ expressing cells in DN thymocytes transduced with a HOXA9-GFP overexpressing vector or GFP control construct cultivated *in vitro* for 6 days on OP9-DL1 stromal cells.

(F) Quantification of TCRαβ and TCRγδ expressing cells in DN thymocytes (as in E). Quantification represents data of three independent experiments.

(G) Schematic representation of the *in vivo* experimental design in mice. Representative FACS plots.

(H) Representative FACS plots of TCRαβ and TCRγδ expressing cells in donor derived thymocytes isolated from $Rag2^{-/-}\gamma c^{-/-}$ mice after 3 weeks post-transplantation.

(I) Quantification of TCRαβ and TCRγδ expressing cells in donor derived thymocytes (as in H). Quantification represents data of three independent experiments.

(J) RQ-PCR analysis of the HOXA9 gene expression in the TCRαβ (n=45) and TCRγδ (n=96) T-ALLs. Results are represented relative to the *ABL1* housekeeping gene. **P=0.003 (two-tailed Mann-Whitney test).

(K) RQ-PCR analysis of the *pre-Tα* gene expression in the HD-deregulated (n=30) and HD-underegulated (n=8) TCRγδ T-ALL. Results are represented relative to the *ABL1* housekeeping gene. *P=0.02 (two-tailed Mann-Whitney test).

(L) RQ-PCR analysis of the *SOX13* gene expression in the HD-deregulated (n=34) and HD-underegulated (n=12) TCRγδ T-ALL. Results are represented relative to the *ABL1* housekeeping gene. **P=0.003 (two-tailed Mann-Whitney test).

See also **Table 1**.

TABLES WITH TITLES AND LEGENDS

Table 1: *TCRB* rearrangement analysis in the TCRγδ T-ALLs according to their deregulation of homeodomain genes.

TCRγδ T-ALL patients (n=95)	No. (%)	*TCRB* rearrangements		
		VDJ	DJ	GL
HD-deregulated	71 (75)	45	23	3
HOXA	39 (41)	16	20	3
TLX1/3	32 (34)	29	3	0
HD-underegulated	24 (25)	3	10	11

Abbreviations: HD, homeodomain; T-ALL, T cell acute lymphoblastic leukemia; VDJ, complete *TCRB* rearrangements; DJ uncomplete *TCRB* rearrangements; GL, germline (*TCRB* not rearranged).

See also **Figure 6**.

STAR METHODS

KEY RESOURCE TABLE

(table is in a separated file in dropbox)

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact Vahid Asnafi (vahid.asnafi@nck.aphp.fr).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines

HeLa cells were maintained in DMEM-Dulbecco's Modified Eagle's Medium, high glucose (Thermo Fisher, Cat#41965062) supplemented with 10% fetal bovine serum (FBS) (Thermo Fisher, Cat#10270106) and 1% penicillin/streptomycin.

Loucy cell line (ACC-394) cell line was cultured in RPMI-1640 medium (Thermo Fisher, Cat#21875091) supplemented with 20% FBS; (Thermo Fisher, Cat#10270106), 1% penicillin/streptomycin and 1x sodium pyruvate.

Platinum-E (Plat-E, the retroviral packaging cell line) was maintained in DMEM supplemented with 10% FBS (Thermo Fisher, Cat#10270106), 1% penicillin/streptomycin, 10 µg/ml blasticidin and 1 µg/ml puromycin.

OP-DL1 stromal cell lines were maintained in MEM α, GlutaMAX supplement medium (ThermoFisher, Cat#32561094) supplemented with 20% FBS Hyclone SH30070.03HI (Fisher Scientific, Cat#10772634) and 1% penicillin/streptomycin.

Primary cells preparation

Human CD34+ umbilical cord blood cells (UCB) were isolated from cord blood using lymphocyte separation medium (Eurobio, Cat#CMSMSL01-01). Next, immature CD34+ UCB cells were positively selected with direct CD34+ progenitor cell isolation kit (Miltenyi Biotec, Cat#130-046-702) using the LS separation columns (Miltenyi, Cat#130-042-401). The positive, CD34+ fraction of cells were further purified by sorting with FACS ARIA III to obtain 98-99% purity of CD34+ UCB cells.

Human thymic subpopulations were obtained from children undergoing heart surgery. The informed consent was obtained from the parents. Thymi were dissociated until the single cell suspension. Specific subpopulations were purified by sorting using FACS ARIA III (BD Biosciences). For the immature subpopulation, thymocytes were pre-purified by depletion of CD3 and CD8 positive fraction by magnetic activated cell sorting using CD3 MicroBeads (Miltenyi, Cat#130-050-101) and CD8 MicroBeads (Miltenyi, Cat#130-045-201) respectively. Antibody used for cell labelling of CD3-/CD8- thymic fraction: CD1a FITC, clone NA1/34 (Dako, Cat#F7141), CD34 APC (BD Biosciences, 345804), CD8 PC7 (Beckman, 737661), CD3 Alexa-700 (BD Biosciences, 557943), CD4 V450, clone RPA-T4 (BD Biosciences, 560345), CD45 V500, clone HI30 (BD Biosciences, 560777). Final purity after sorting was over 95%.

Mouse thymic cells from 6- to 8-week old C57Bl/6Rj mice were extracted and mechanistically disrupted on a 75-µm nylon cell strainer. Cells were washed twice with cold PBS supplemented with 2% of FBS. DN thymocytes were isolated by depletion of CD4 and CD8 positive fraction using magnetic beads (Miltenyi, Cat#130-117-043 and Cat#130-117-044, respectively). For maximum purity, sorting of DN cells was performed using a FACS ARIA III.

Mouse ΔRag thymocytes were obtained from the $Rag2^{-/-}$ mice and consists mainly DN3 cells. ΔRagCD3 were obtained from the $Rag2^{-/-}$ mice injected with CD3ε antibody to promote the transition from DN to DP stage.

Rodents

C57Bl/6 (CD45.2) mice were purchased from Janvier Labs. $Rag2^{-/-}\gamma c^{-/-}$ (CD45.1) mice used in the *in vivo* experiments were purchased from xxx. All experiments were performed according to procedures approved by the Committee xxx.

Patient sample collection

150 adult T-ALL patients were included in the multicenter randomized GRAALL (Group for Research in Adult Acute Lymphoblastic Leukemia) -2003 and -2005 studies. Both studies were conducted in accordance with the Declaration of Helsinki and approved by local and multicenter research ethical committees. Informed consent was obtained from all patients at trial entry. Both trials were registered at http://www.clinicaltrials.gov as NCT00222027 and NCT00327678. Leukemic blasts were isolated from peripheral blood of T-ALL patients using Ficoll gradient separation. The criteria for inclusion in the current study were a diagnosis of

TCR-expressing T-ALLs and the availability of diagnostic material for gene expression evaluation.

METHOD DETAILS

Reference epigenomes of human thymopiesis

Five sorted human thymic subpopulations (tCD34, EC, LC, SP4, SP8) were used for ChIP-seq (H3K4me3, H4K4me1, H3K27ac, H3K27me3, H3K36me3 and H3K9me3) and Whole-Genome Bisulfite Sequencing (WGBS) following the BLEUPRINT protocol (http://dcc.blueprint-epigenome.eu/#/md/methods). Briefly, for ChIP-seq of histone marks, sorted thymis samples were incubated with 1% formaldehyde (Merck, Cat#F8775) in PBS for 10 minutes at room temperature. Next, 1/10 volume of quenching solution of 1.25M glycine (Merck, Cat#GE17-1323-01) was added to the medium and was shaken for 5 minutes at room temperature. Cells were collected by centrifugation and washed one with PBS-BSA 0.5%.For WGBS-seq DNA was extracted from sorted thymic subpopulations using Nucleon BACC2 extraction kit (Merck, Cat#GERPN8502) according to the instructions of the supplier.

Processed data were retrieved from the BLEUPRINT project (Details in Table GC1). To complement the reference epigenomes, additional ChIP samples were sequenced in house in single-end 75nt mode using the NextSeq® 500/550 (Illumina) according to manufacturer's instructions, and processed following the BLEUPRINT protocol. Aligned reads of ChIP-Seq from multiple donors were merged for each thymic subpopulation (Details in Table GC2 to do).

RNA-seq of human thymopiesis

Six sorted human thymic subpopulations (tCD34, ISP, EC, LC, SP4, SP8) were used for Poly(A)-enriched RNA-Seq. Fragments were sequenced in stranded paired-end (2 x 50 bp) using the SOLiD HQ5500XL platform (Life Technologies) and default parameters (Details in Table GC3).

Assay for Transposase Accessible Chromatin (ATAC-seq) of human thymopoiesis

Four sorted human thymic subpopulations (CD34+1a-7+, CD34+1a+, EC, LC) from two donors were used for ATAC-Seq. 5.10^4 thymocytes were washed with cold PBS and lysed in 50µl cold lysis buffer (10 mM Tris HCL pH 7.4; 10 mM NaCl; 3m M MgCl2; 0.1% Igepal CA-630). Transposition reaction was performed in 50 µl of 1x TD reaction buffer supplemented with 2.5 µl of TDE1 Tn5 Transposase from Nextera DNA Library Prep Kit

(Illumina, Cat#FC-121-1030) at 37°C for 30 minutes. Directly after transposition, purification using MinElute PCR Purification kit (Qiagen, Cat#28004) was performed followed by sequencing. Samples were sequenced in house in single-end 75nt mode using the NextSeq® 500/550 (Illumina) according to manufacturer's instructions. Reads were trimmed with sickle and aligned with bowtie2 using default settings. Aligned reads from $tCD34^{+}1a^{-}7^{+}$ and $tCD34^{+}1a^{+}$ populations from the two donors were merged to generate the tCD34 ATAC sample.

Chromatin segmentation and space generation

Merged ChIP-Seq for each thymic subpopulations were segmented with ChromHMM (http://dx.doi.org/10.1038/nprot.2017.124) using the 11-state chromatin model from (Carrillo-de-Santa-Pau et al., 2017) to complement their published chromatin states from other hematopoietic cells. Regions with variable chromatin states across hematopoietic cell types were extracted using Chromdet and 5-state collapse model from (Carrillo-de-Santa-Pau et al., 2017). Multiple Correspondence Analysis was done using FactoMineR R package and cell-type delimitation ellipses were manually added afterward. Sankey plot was generated using riverplot R package.

Genome browser views were produced with IGV then compacted with Inkscape.

Functional enrichment analysis

Functional enrichment analyses were produced using a custom pipeline to automate multi-sample queries to GREAT web-service. In summary, features were converted to hg19 assembly using crossmap and queried with rGREAT R package against GO Biological Process and MSigDB Pathways. 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. Additionally, MSigDB Pathways terms were further filtered for heatmap display to keep only the best 5 terms for each sample according to binomial p-value. For GO Biological Process, GOSemSim R package was used to compute Wang similarity distance between all terms. Terms with a similarity distance higher than 0.1 were grouped together. Only the best term according to Binomial Fold Enrichment for each group and sample was kept for heatmap display.

Definition of distal regulatory elements and association to genes

Hypomethylated CpG regions were defined from WGBS following the BLEUPRINT protocol (http://dcc.blueprint-epigenome.eu/#/md/bs_seq_grch38). Distal regulatory regions were defined as hypomethylated regions in at least one thymic subpopulation falling between 2 and 2000 kb from the closest gene, based on the subset of Ensembl release 93 annotation track containing only protein coding and TR-IG genes. Regions overlapping repeats from repeatMasker, mitochondrial and sexual chromosomes, or exhibiting standard deviation higher than 0.05 between donors samples from the same thymic subpopulation were filtered-out. Constitutive hypomethylated distal regulatory elements were defined as those with CpG mean methylation signal below 0.25 for all thymic subpopulation samples. Distal cis-regulatory regions from WGBS and ATAC peaks were associated to both their nearest upstream and downstream genes as long as they were within 2000 kb range.

Clustering

Clustering was done by k-means using Lloyd algorithm and Pearson distance on mean CpG signal in hypomethylated regions (Figure 2A), on qualitative H3K27ac peak presence (Figure 2E) and on qualitative ATAC peak presence (Figure 3).

Normalization and peak-calling

ChIP-Seq coverage tracks were computed using DANPOS and quantile-normalized against SP8 samples as an arbitrary reference. ATAC-Seq coverage tracks were computed using deepTools bamCoverage with RPKM normalization. Peaks for both ChIP-Seq and ATAC-Seq were called using Macs2. Profiles and heatmaps were generated using deepTools plotProfile and plotHeatmap, respectively.

Gene expression analysis

Solid RNA-Seq data were processed using LifeScope™ (Life Technologies) and TopHat using default parameters. Thy3 and Thy4 RNA-Seq samples from Casero et al. 2015 (Casero et al., 2015) were retrieved and aligned using STAR and Ensembl GRCh38 release 93 annotation track. Quantification in genes for both experiments was done with Subread FeaturesCounts. RPK TMM count normalization and differential expression analysis were done using EdgeR.

Motif enrichment

Conserved transcription factor binding sites were retrieved from UCSC genome browser. Binding motifs for the same transcription factor were merged. Clustered distal ATAC peaks

were queried for enrichment against the merged set of motifs using Pygtftk ologram (http://dx.doi.org/10.1093/bioinformatics/btz116).

All sequencing-based analyses produced in this paper are gathered as a Snakemake workflow available at https://github.com/guillaumecharbonnier/mw-cieslak2019

Gene expression analysis

RQ-PCR. RNA was extracted using the RNeasy Mini Kit (Qiagen, Cat#74104) and converted into cDNA using SuperScript III Reverse Transcriptase (Thermo Fisher, Cat#18080093). Real-time PCR reactions were performed using Taqman Universal PCR Master Mix (Thermo Fisher, Cat#4304437) or SYBR Green PCR Master Mix (Thermo Fisher, Cat#4309155). RQ-PCR was performed on an Applied Biosystems 7900HT Fast Real-Time PCR system. Primer sequences are in the **Table S2** (List of the Nucleotides Used in the Study).

Multiplex *TCRA* RT-PCR analysis

TCRA cDNA was amplified from Cα (FAM fluorophore-labelled) and various Vα in 5 multiplex RT-PCRs (Villarese et al., 2018). Briefly, 20ng of cDNA was amplified in each tube of PCR with 1U HotStarTaq DNA Polymerase (Qiagen, Cat#203203), 2mM of MgCl2, 20% Q solution, 10mM dNTP final concentration and 10pmol of each primer. The Taq polymerase was activated 15 min at 95 °C, after 37 cycles at 94 °C during 30 s, 63 °C 45 s, 72 °C 1 min 30 s. The final elongation was 72 °C during 10 min. Primer sequences are in the **Table S2**. The analysis of rearrangements by multiplex fluorescent PCR was performed by separation of single stand (denaturated) PCR products in a capillary sequencing polymer and detected via automated laser scanning (3130 Genetic Analyzer, Applied Biosystems).

CAT-reporter assay

HeLa cell lines were transfected using Lipofectamine 2000 Transfection Reagent (Thermo Fisher, Cat#11668019) in the Opti-MEM I Reduced Serum Medium (Thermo Fisher, Cat#51985026) according to the manufacturer's instructions. Cells at 70% confluency were transfected with the Eα-CAT reporter construct together with expression vectors for ETS1-HA-His, LEF1-HA, and RUNX1-T7. To test the repression of CAT expression the cells were additionally transfected with expressing vectors for full length forms of HOXA5-9 or homeodomain deleted forms: HOXA5-ΔHD (aa 1-187), HOXA6-ΔHD (aa 1-147), HOXA7-ΔHD (aa 1-126), HOXA9-ΔHD (aa 1-193). All HOXA proteins were cloned into the pEGFP-C1 vector (Clontech, Cat#6084-1) in which they were fused with GFP in their N-terminus.

After 24h of incubation expression of CAT activity was determined using CAT ELISA kit (Merck, Cat#11363727001) according to the manufacture's recommendations.

Chromatin immunoprecipitations (ChIP)

tCD34+, EC and LC thymocytes were crosslinked for 10 minutes with 1% formaldehyde at 20°C and sonicated using Branson 450 Sonifier Cell Disruptor in order to obtain an average length for DNA fragments of approximately 500 bp. Chromatin immunoprecipitation was conducted with anti-RUNX1 (Abcam, Cat#ab23980), anti-ETS1 (C20X) (Santa Cruz, Cat#sc-350) and anti-IgG antibody (Abcam, Cat#ab37415) using Dynabeads Protein G for Immunoprecipitation (Thermo Fisher Cat#10003D). Extracted ChIP DNA was purified by QIAquick PCR Purification Kit (Qiagen, Cat#28104). ChIPed DNA was quantified in RQ-PCR assay using primers presented in **Table S2**.

ChIP analysis of HOXA binding was performed in two cell line models: Loucy and HeLa. Loucy cell line was electroporated using Neon Transfection System (1 pulse of 1700V, 20ms,) with expressing vectors for HOXA5-9 (full length and homeodomain deleted forms) tagged with SBP-Flag epitope and empty vector NTAP as control (vectors are listed in the **Key Resources Table**). After 48h incubation, cells were crosslinked with 1% formaldehyde for 10 minutes. Chromatin immunoprecipitation was performed as described above using anti-Flag antibody (Merck, Cat#F1804). Samples were analyzed by RQ-PCR using Ealpha-F/Ealpha-R for enhancer alpha binding and Actin-F/Actin-R for control. The primers are listed in the **Table S2**. The HeLa cell line was transiently transfected using Lipofectamine 2000 Transfection Reagent (Thermo Fisher, Cat#11668019) in the Opti-MEM I Reduced Serum Medium (Thermo Fisher, Cat#51985026) with expressing vectors for ETS1-HA-His, LEF1-HA, RUNX1-T7, and Eα-CAT reporter constructs together with individual expressing vectors for HOXA5-9 (full length or homeodomain deleted forms) tagged with SBP-Flag or the empty vector SBP-Flag (NTAP) as control. After 24h incubation, cells were crosslinked with 1% formaldehyde for 10 minutes. Chromatin immunoprecipitations were performed with anti-Flag antibody (Merck, Cat#F1804). Samples were analyzed by RQ-PCR with the following primers: vect-Ealpha-F/vect-Ealpha-R for enhancer binding vect-ctr-F/vect-ctr-R for control. The primers are listed in the **Table S2**.

Immunoprecipitation (IP)

2.10^8 of Loucy cell line or total thymocytes were lysed for 30 minute with 8 ml of 1x RIPA buffer (Cell Signaling, Cat#9806) complemented with Complete EDTA-free Protease

Inhibitor Cocktail (Merck, Cat#11873580001). The lysates were incubated overnight at 4°C with 10µg of anti-ETS1 (C20X) (Santa Cruz, Cat#sc-350), anti-HOXA9 antibody (13C11, non-commercial production) followed by 2 hours incubation at 4°C with 20µl of Protein G Agarose, Fast Flow (Merck, Cat#16-266). After 4 washes in washing buffer (100 mM NaCl, 15mM Tris HCl pH 7.8) the bound proteins were eluted and detected by Western blot with anti-HOXA9 (13C11) and anti-ETS1 (C20X) antibodies.

Immunofluorescence analyses

Cells were fixed on poly-L-lysine (0.01%) pre-coated slides with PBS-BSA 1%/3.5% formaldehyde solution during 20 minutes at room temperature, then permeabilized in PBS supplemented with 10% FBS and 1% Triton X-100 for 5 minutes (cell lines) or 10 minutes (thymocytes). Cells were incubated at 4°C overnight with anti-HOXA9 (13C11) (dilution 1/20) and anti-ETS1 (C20X) (dilution 1/250) antibodies, and then labeled with secondary antibodies goat anti-mouse IgG Alexa Fluor A555 (Thermo Fisher, Cat#A-21422) and goat anti-rabbit IgG Alexa Fluor 488 (Thermo Fisher, Cat#A-11008) in dilution of 1/200. For Proximity Ligation Assay (PLA) cells were fixed, permeabilized, and incubated with primary antibodies as described above. The following Duolink Probes were used for protein detection: Duolink In Situ PLA Probe Anti-Mouse MINUS (Merck, Cat#DUO92004) for HOXA9 and Duolink In Situ PLA Probe Anti-Rabbit PLUS (Merck, Cat#DUO92002) for ETS1. The interactions between proteins were detected with Duolink In Situ Detection Reagents Red (Merck, Cat#DUO92008) following the manufacturer's instructions. For EGFP fluorescence analysis HeLa cells were transfected with expressing vectors for full length and homeodomain deleted forms of HOXA5-9 genes fused to eGFP in their N-terminus (p-EGFP-C1 vector, same as for CAT-ELISA experiments). After 24 hours, cells were fixed as described above. Images were collected on a confocal microscope (Carl Zeiss LSM 700) with Zen 2011 software using 63x objectives at room temperature. Images were processed using ImageJ software.

Human T cell differentiation

CD34+ UCB cells stimulated for 16h before transduction in the CellGenix GMP SCGM Stem Cell Growth Medium (CellGenix, Cat#20802-500) supplemented with cytokines: 20 ng/ml hTPO, 100 ng/ml hSCF and 100 ng/ml hFLT3-L (Miltenyi). Stimulated CD34+ UCB cells were transduced with HOXA9-GFP or control GFP VSV-G lentiviral vectors. 48h post transduction, GFP expressing cells were sorted using FACS ARIA III and cultured on OP9-

DL1 stromal cells in homemade Minimal Essential Medium MEMα (Thermo Fisher, Cat#12000063) supplemented with 20% FBS Hyclone SH30070.03HI (Fisher Scientific, Cat#10772634) and cytokines: 5 ng/ml rFLT3-L, 10 ng/ml hSCF, 2 ng/ml hIL7 (Miltenyi) (Six et al., 2011). Stromal OP9DL1cells were changed every week. At different time points of culture, cells were collected and analyzed by FACS for expression of surface TCRγδ and TCRαβ and *TCRA* rearrangement by multiplex RT-PCR analysis.

Production of retrovirus particles and transduction of DN thymocytes

The retroviral vectors pMSCV-GFP and pMSCV-HOXA9-GFP were purchased from Addgene. For production of retroviral vector supernatants, Platinum-E (Plat-E) packaging cells ($3x10^5$ cells/well) cultured in 6 well plates were transiently transfected with pMSCV-GFP or pMSCV-HOXA9-GFP vectors (3,3µg/well) using FuGENE HD Transfection Reagent (Fisher Scientific, Cat#PRE2311) and Opti-MEM Reduced Serum Media. Medium was changed 24h post-transfection. Vector supernatants were collected 48h after transfection and were freshly used for thymocyte transduction. Sorted DN thymocytes were resuspended in vector supernatants with 10µg/ml polybrene and spinoculated for 3 hours at 3500 rpm at room temperature. Next, the sorted cells were cultured in vitro on OP9-DL1 or injected into $Rag2^{-/-}$ $\gamma c^{-/-}$ Ly5.1 mice.

Mouse T cell differentiation

DN thymocytes were FACS sorted from C57/Bl6 mice Ly5.2. During *in vitro* studies, transduced DN cells were cultured on an OP9-DL1 stroma layer in homemade Minimal Essential Medium MEMα (Thermo Fisher, Cat#12000063) supplemented with 20% FBS Hyclone SH30070.03HI (Fisher Scientific, Cat#10772634) and recombinant murine cytokines: 5 ng/ml rFLT3-L (RnD Systems, Cat#427-FL), and 5 ng/ml rIL7 (Cat#407-ML). During *in vivo* studies, transduced DN cells were retro-orbitally injected into $Rag2^{-/-}\gamma c^{-/-}$ Ly5.1 mice that had been previously irradiated (2Gy). After 3 weeks, post-transplant, recipient mice were euthanized and the donor derived cells analyzed.

SUPPLEMENTAL INFORMATION TITLES AND LEGENDS

Figure S1: Major human T cell differentiation stages, Related to Figure 1

(A) Schematic representation of major stages of human thymopoiesis (orange). The thymic sub-populations used to prepare BLEUPRINT reference epigenomes are presented in blue colour.

(B) Plots show the gating strategy used for sorting human thymic sub-populations. Purity after sort was checked to be 95-99%.

tCD34: immature double negative $CD34^+$ ($CD34^+/CD3^-/CD4^-/CD8^-$); EC: early cortical ($TCR\alpha\beta^-/CD3^-/CD4^+/CD8^+$), LC: late cortical ($TCR\alpha\beta^+/CD3^{+low}/CD4^+/CD8^+$), SP4: single positive CD4+ ($TCR\alpha\beta^+/CD3^+/CD4^+/CD8^-$), SP8: single positive CD8+ ($TCR\alpha\beta^+/CD3^+/CD4^-/CD8^+$).

(C) State emissions for the used chromatin segmentation model and biological description.

(D) GREAT gene enrichment analysis for the genomic regions highly correlating with first and second dimensions of MCA from Figure 1D.

Figure S2: Integration of DNA methylation and histone modifications, Related to Figure 2

(A) t-SNE on WGBS from all BLEUPRINT healthy hematopoietic samples.

(B) Histone modifications and methylation profiles around the 8 clustered classes based on hypomethylation dynamics during T cell differentiation (Figure 2A).

(C) Histone modifications and methylation profiles around the 8 clustered classes based on H3K27-acetylation dynamics during T-Cell differentiation at constitutively hypomethylated regions.

(D) GREAT gene enrichment analysis for the 8 clustered classes based on H3K27-acetylation dynamics during T-Cell differentiation at constitutively hypomethylated regions.

Figure S3: Epigenomics profile of an induced gene, Related to Figure 3

Figure S4: ChIP-seq and short-RNA-seq profiles of the *TCRA* locus in the mouse thymocytes, Related to Figure 4

(A) DNaseI-seq and ChIP-seq analysis for ETS1, RUNX1, E47, GATA3, and Ikaros binding to *TCRA* enhancer (Eα) in the ΔRag and double positive (DP) mouse thymocytes.

(B) Short-RNA-seq analysis for the enhancer RNA (eRNA) in the ΔRag and double positive (DP) murine thymocytes.

Figure S5: Expression of *HOXA5-9* genes in human thymic subpopulations and their nuclear localization, Related to Figure 5

(A) Chromatin state profiles for human *HOXA* locus. The color code as in Figure 1C.

(B) Heatmap representation of cluster C13

(C) Unsupervised RNA-seq gene expression of genes of HOXL family.

(D) Column plot representation of *HOXA5-9* gene expression obtained by TLDA. Results are normalized to expression of the *GAPDH* housekeeping gene.

(E) RQ-PCR for *HOXA5-9* gene expression in human thymic subpopulations. Results are represented relative to the *ABL1* housekeeping gene.

(F) Examples of fluorescence microscopic analysis of HOXA5-9 expression in HeLa cells transfected with expression vectors for full length (FL) and homeodomain deleted forms (ΔHD) of HOXA5-9. White scale bar is 10 mm.

Table S1: Design of HOXL assay in TLDA.

Table S2: List of the Nucleotides Used in the Study

REFERENCES

Abarrategui, I., and Krangel, M.S. (2009). Germline transcription: a key regulator of accessibility and recombination. Adv Exp Med Biol *650*, 93–102.

Alharbi, R.A., Pettengell, R., Pandha, H.S., and Morgan, R. (2013). The role of HOX genes in normal hematopoiesis and acute leukemia. Leukemia *27*, 1000–1008.

Asnafi, V., Beldjord, K., Boulanger, E., Comba, B., Le Tutour, P., Estienne, M.H., Davi, F., Landman-Parker, J., Quartier, P., Buzyn, A., et al. (2003). Analysis of TCR, pT alpha, and RAG-1 in T-acute lymphoblastic leukemias improves understanding of early human T-lymphoid lineage commitment. Blood *101*, 2693–2703.

Asnafi, V., Beldjord, K., Libura, M., Villarese, P., Millien, C., Ballerini, P., Kuhlein, E., Lafage-Pochitaloff, M., Delabesse, E., Bernard, O., et al. (2004). Age-related phenotypic and oncogenic differences in T-cell acute lymphoblastic leukemias may reflect thymic atrophy. Blood *104*, 4173–4180.

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*, 823–837.

Bassing, C.H., Tillman, R.E., Woodman, B.B., Canty, D., Monroe, R.J., Sleckman, B.P., and Alt, F.W. (2003). T cell receptor (TCR) alpha/delta locus enhancer identity and position are critical for the assembly of TCR delta and alpha variable region genes. Proc Natl Acad Sci U A *100*, 2598–2603.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., et al. (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell *133*, 1266–1276.

Carey, M. (1998). The enhanceosome and transcriptional synergy. Cell *92*, 5–8.

Carrillo-de-Santa-Pau, E., Juan, D., Pancaldi, V., Were, F., Martin-Subero, I., Rico, D., and Valencia, A. (2017). Automatic identification of informative regions with epigenomic changes associated to hematopoiesis. Nucleic Acids Res. *45*, 9244–9259.

Carroll, S.B. (1995). Homeotic genes and the evolution of arthropods and chordates. Nature *376*, 479–485.

Casero, D., Sandoval, S., Seet, C.S., Scholes, J., Zhu, Y., Ha, V.L., Luong, A., Parekh, C., and Crooks, G.M. (2015). Long non-coding RNA profiling of human lymphoid progenitor cells reveals transcriptional divergence of B cell and T cell lineages. Nat. Immunol. *16*, 1282–1291.

Cauchy, P., Maqbool, M.A., Zacarias-Cabeza, J., Vanhille, L., Koch, F., Fenouil, R., Gut, M., Gut, I., Santana, M.A., Griffon, A., et al. (2016). Dynamic recruitment of Ets1 to both nucleosome-occupied and -depleted enhancer regions mediates a transcriptional program switch during early T-cell differentiation. Nucleic Acids Res. *44*, 3567–3585.

Choo, S.W., White, R., and Russell, S. (2011). Genome-wide analysis of the binding of the Hox protein Ultrabithorax and the Hox cofactor Homothorax in Drosophila. PLoS One *6*, e14778.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U A *107*, 21931–21936.

Dadi, S., Le Noir, S., Payet-Bornet, D., Lhermitte, L., Zacarias-Cabeza, J., Bergeron, J., Villarese, P., Vachez, E., Dik, W.A., Millien, C., et al. (2012). TLX Homeodomain Oncogenes Mediate T Cell Maturation Arrest in T-ALL via Interaction with ETS1 and Suppression of TCRalpha Gene Expression. Cancer Cell *21*, 563–576.

Dik, W.A., Pike-Overzet, K., Weerkamp, F., de Ridder, D., de Haas, E.F., Baert, M.R., van der Spek, P., Koster, E.E., Reinders, M.J., van Dongen, J.J., et al. (2005a). New insights on human T cell development by quantitative T cell receptor gene rearrangement studies and gene expression profiling. J Exp Med *201*, 1715–1723.

Dik, W.A., Brahim, W., Braun, C., Asnafi, V., Dastugue, N., Bernard, O.A., van Dongen, J.J., Langerak, A.W., Macintyre, E.A., and Delabesse, E. (2005b). CALM-AF10+ T-ALL expression profiles are characterized by overexpression of HOXA and BMI1 oncogenes. Leukemia. *19*, 1948-57.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Ferrando, A.A., Neuberg, D.S., Staunton, J., Loh, M.L., Huard, C., Raimondi, S.C., Behm, F.G., Pui, C.H., Downing, J.R., Gilliland, D.G., et al. (2002). Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. Cancer Cell *1*, 75–87.

Ferrando, A.A., Armstrong, S.A., Neuberg, D.S., Sallan, S.E., Silverman, L.B., Korsmeyer, S.J., and Look, A.T. (2003). Gene expression signatures in MLL-rearranged T-lineage and B-precursor acute leukemias: dominance of HOX dysregulation. Blood. *102*, 262-8. Epub 2003 Mar 13.

Giese, K., Cox, J., and Grosschedl, R. (1992). The HMG domain of lymphoid enhancer factor 1 bends DNA and facilitates assembly of functional nucleoprotein structures. Cell *69*, 185–195.

Giese, K., Kingsley, C., Kirshner, J.R., and Grosschedl, R. (1995). Assembly and function of a TCR alpha enhancer complex is dependent on LEF-1-induced DNA bending and multiple protein-protein interactions. Genes Dev *9*, 995–1008.

Goodman, F.R. (2002). Limb malformations and the human HOX genes. Am J Med Genet *112*, 256–265.

Hawwari, A., and Krangel, M.S. (2005). Regulation of TCR delta and alpha repertoires by local and long-distance control of variable gene segment chromatin structure. J Exp Med *202*, 467–472.

Hernandez-Munain, C., Sleckman, B.P., and Krangel, M.S. (1999). A developmental switch from TCR delta enhancer to TCR alpha enhancer function during thymocyte maturation. Immunity *10*, 723–733.

Ho, I.C., Yang, L.H., Morle, G., and Leiden, J.M. (1989). A T-cell-specific transcriptional enhancer element 3' of C alpha in the human T-cell receptor alpha locus. Proc Natl Acad Sci U A *86*, 6714–6718.

Ho, I.C., Bhat, N.K., Gottschalk, L.R., Lindsten, T., Thompson, C.B., Papas, T.S., and Leiden, J.M. (1990). Sequence-specific binding of human Ets-1 to the T cell receptor alpha gene enhancer. Science *250*, 814–818.

Hollenhorst, P.C., Chandler, K.J., Poulsen, R.L., Johnson, W.E., Speck, N.A., and Graves, B.J. (2009). DNA specificity determinants associate with distinct transcription factor functions. PLoS Genet *5*, e1000778.

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.e8.

Huang, C.Y., Golub, R., Wu, G.E., and Kanagawa, O. (2002). Superantigen-induced TCR alpha locus secondary rearrangement: role in tolerance induction. J Immunol *168*, 3259–3265.

Huang, Y., Sitwala, K., Bronstein, J., Sanders, D., Dandekar, M., Collins, C., Robertson, G., MacDonald, J., Cezard, T., Bilenky, M., et al. (2012). Identification and characterization of Hoxa9 binding sites in hematopoietic cells. Blood *119*, 388–398.

Izon, D.J., Rozenfeld, S., Fong, S.T., Komuves, L., Largman, C., and Lawrence, H.J. (1998). Loss of function of the homeobox gene Hoxa-9 perturbs early T-cell development and induces apoptosis in primitive thymocytes. Blood *92*, 383–393.

Krangel, M.S., Carabana, J., Abbarategui, I., Schlimgen, R., and Hawwari, A. (2004). Enforcing order within a complex locus: current perspectives on the control of V(D)J recombination at the murine T-cell receptor alpha/delta locus. Immunol Rev *200*, 224–232.

Lawrence, H.J., Sauvageau, G., Humphries, R.K., and Largman, C. (1996). The role of HOX homeobox genes in normal and leukemic hematopoiesis. Stem Cells *14*, 281–291.

Lawrence, H.J., Helgason, C.D., Sauvageau, G., Fong, S., Izon, D.J., Humphries, R.K., and Largman, C. (1997). Mice bearing a targeted interruption of the homeobox gene HOXA9 have defects in myeloid, erythroid, and lymphoid hematopoiesis. Blood *89*, 1922–1930.

Lewis, E.B. (1978). A gene complex controlling segmentation in Drosophila. Nature *276*, 565–570.

Magli, M.C., Largman, C., and Lawrence, H.J. (1997). Effects of HOX homeobox genes in blood cell differentiation. J Cell Physiol *173*, 168–177.

Mann, R.S., Lelli, K.M., and Joshi, R. (2009). Hox specificity unique roles for cofactors and collaborators. Curr Top Dev Biol *88*, 63–101.

McMurry, M.T., and Krangel, M.S. (2000). A role for histone acetylation in the developmental regulation of VDJ recombination. Science *287*, 495–498.

Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H., and Wolfe, S.A. (2008). Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell *133*, 1277–1289.

Pekowska, A., Benoukraf, T., Zacarias-Cabeza, J., Belhocine, M., Koch, F., Holota, H., Imbert, J., Andrau, J.C., Ferrier, P., and Spicuglia, S. (2011). H3K4 tri-methylation provides an epigenetic signature of active enhancers. Embo J *30*, 4198–4210.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Roberts, J.L., Lauzurica, P., and Krangel, M.S. (1997). Developmental regulation of VDJ recombination by the core fragment of the T cell receptor alpha enhancer. J Exp Med *185*, 131–140.

Six, E.M., Benjelloun, F., Garrigue, A., Bonhomme, D., Morillon, E., Rouiller, J., Cacavelli, L., Blondeau, J., Beldjord, K., Hacein-Bey-Abina, S., et al. (2011). Cytokines and culture medium have a major impact on human in vitro T-cell differentiation. Blood Cells Mol Dis *47*, 72–78.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H.J., et al. (2011). Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. Cell *147*, 1270–1282.

Sleckman, B.P., Bardon, C.G., Ferrini, R., Davidson, L., and Alt, F.W. (1997). Function of the TCR alpha enhancer in alphabeta and gammadelta T cells. Immunity *7*, 505–515.

Soulier, J., Clappier, E., Cayuela, J.M., Regnault, A., Garcia-Peydro, M., Dombret, H., Baruchel, A., Toribio, M.L., and Sigaux, F. (2005). HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). Blood. *106*, 274-86. Epub 2005 Mar 17.

Speleman, F., Cauwelier, B., Dastugue, N., Cools, J., Verhasselt, B., Poppe, B., Van Roy, N., Vandesompele, J., Graux, C., Uyttebroeck, A., et al. (2005). A new recurrent inversion, inv(7)(p15q34), leads to transcriptional activation of HOXA10 and HOXA11 in a subset of T-cell acute lymphoblastic leukemias. Leukemia. *19*, 358-66.

Spicuglia, S., Payet, D., Tripathi, R.K., Rameil, P., Verthuy, C., Imbert, J., Ferrier, P., and Hempel, W.M. (2000). TCRalpha enhancer activation occurs via a conformational change of a pre-assembled nucleo-protein complex. Embo J *19*, 2034–2045.

Spits, H. (2002). Development of alphabeta T cells in the human thymus. Nat Rev Immunol *2*, 760–772.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

Stunnenberg, H.G., and Hirst, M. (2016). The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell *167*, 1897.

Taghon, T., Thys, K., De Smedt, M., Weerkamp, F., Staal, F.J., Plum, J., and Leclercq, G. (2003). Homeobox gene expression profile in human hematopoietic multipotent stem cells and T-cell progenitors: implications for human T-cell development. Leukemia *17*, 1157–1163.

Trinquand, A., Dos Santos, N.R., Tran Quang, C., Rocchetti, F., Zaniboni, B., Belhocine, M., Da Costa de Jesus, C., Lhermitte, L., Tesio, M., Dussiot, M., et al. (2016). Triggering the TCR Developmental Checkpoint Activates a Therapeutically Targetable Tumor Suppressive Pathway in T-cell Leukemia. Cancer Discov. *6*, 972–985.

Van Vlierberghe, P., van Grotel, M., Tchinda, J., Lee, C., Beverloo, H.B., van der Spek, P.J., Stubbs, A., Cools, J., Nagata, K., Fornerod, M., et al. (2008). The recurrent SET-NUP214 fusion as a new HOXA activation mechanism in pediatric T-cell acute lymphoblastic leukemia. Blood *111*, 4668–4680.

Villarese, P., Lours, C., Trinquand, A., Le Noir, S., Belhocine, M., Lhermitte, L., Cieslak, A., Tesio, M., Petit, A., LeLorch, M., et al. (2018). TCRalpha rearrangements identify a subgroup of NKL-deregulated adult T-ALLs associated with favorable outcome. Leukemia *32*, 61–71.

Villey, I., Caillol, D., Selz, F., Ferrier, P., and de Villartay, J.P. (1996). Defect in rearrangement of the most 5' TCR-J alpha following targeted deletion of T early alpha (TEA): implications for TCR alpha locus accessibility. Immunity *5*, 331–342.

Waddington, C.H. (1957). The strategy of the genes; a discussion of some aspects of theoretical biology.

Winoto, A., and Baltimore, D. (1989). A novel, inducible and T cell-specific enhancer located at the 3' end of the T cell receptor alpha locus. Embo J *8*, 729–733.

Zentner, G.E., Tesar, P.J., and Scacheri, P.C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Res *21*, 1273–1283.

Zhang, J., Ding, L., Holmfeldt, L., Wu, G., Heatley, S.L., Payne-Turner, D., Easton, J., Chen, X., Wang, J., Rusch, M., et al. (2012). The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature *481*, 157–163.

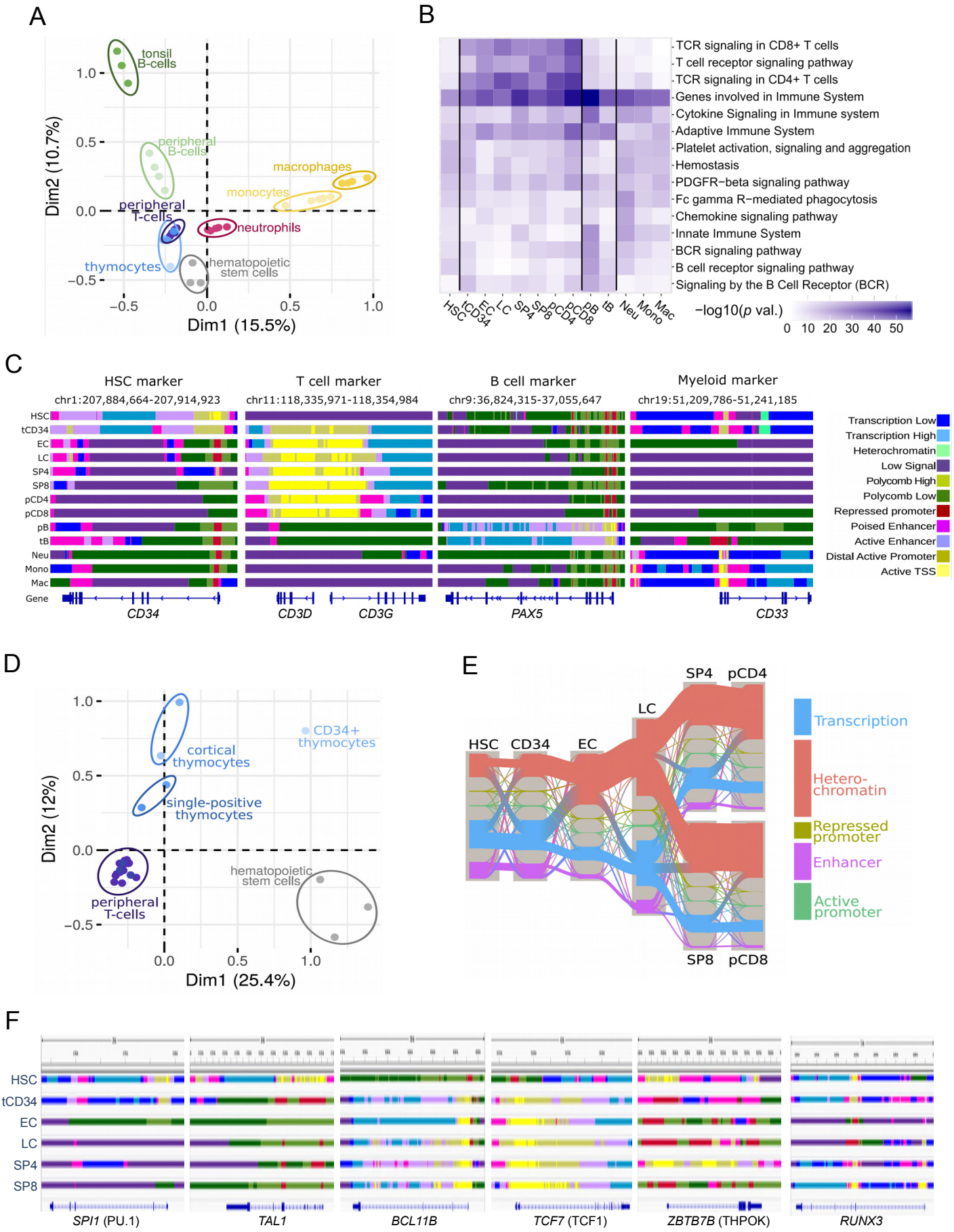

A
tonsil B-cells
peripheral B-cells
macrophages
monocytes
peripheral T-cells
neutrophils
thymocytes
hematopoietic stem cells
Dim2 (10.7%)
Dim1 (15.5%)
B
TCR signaling in CD8+ T cells
T cell receptor signaling pathway
TCR signaling in CD4+ T cells
Genes involved in Immune System
Cytokine Signaling in Immune system
Adaptive Immune System
Platelet activation, signaling and aggregation
Hemostasis
PDGFR−beta signaling pathway
Fc gamma R−mediated phagocytosis
Chemokine signaling pathway
Innate Immune System
BCR signaling pathway
B cell receptor signaling pathway
Signaling by the B Cell Receptor (BCR)
HSC tCD34 EC LC SP4 SP8 pCD4 pCD8 pB tB Neu Mono Mac
−log10(p val.)
0 10 20 30 40 50
C
HSC marker
chr1:207,884,664-207,914,923
T cell marker
chr11:118,335,971-118,354,984
B cell marker
chr9:36,824,315-37,055,647
Myeloid marker
chr19:51,209,786-51,241,185
HSC tCD34 EC LC SP4 SP8 pCD4 pCD8 pB tB Neu Mono Mac Gene
CD34 CD3D CD3G PAX5 CD33
Transcription Low
Transcription High
Heterochromatin
Low Signal
Polycomb High
Polycomb Low
Repressed promoter
Poised Enhancer
Active Enhancer
Distal Active Promoter
Active TSS
D
CD34+ thymocytes
cortical thymocytes
single-positive thymocytes
hematopoietic stem cells
peripheral T-cells
Dim2 (12%)
Dim1 (25.4%)
E
SP4 pCD4
LC
HSC CD34 EC
SP8 pCD8
Transcription
Hetero-chromatin
Repressed promoter
Enhancer
Active promoter
F
HSC
tCD34
EC
LC
SP4
SP8
SPI1 (PU.1) TAL1 BCL11B TCF7 (TCF1) ZBTB7B (THPOK) RUNX3

Figure 1

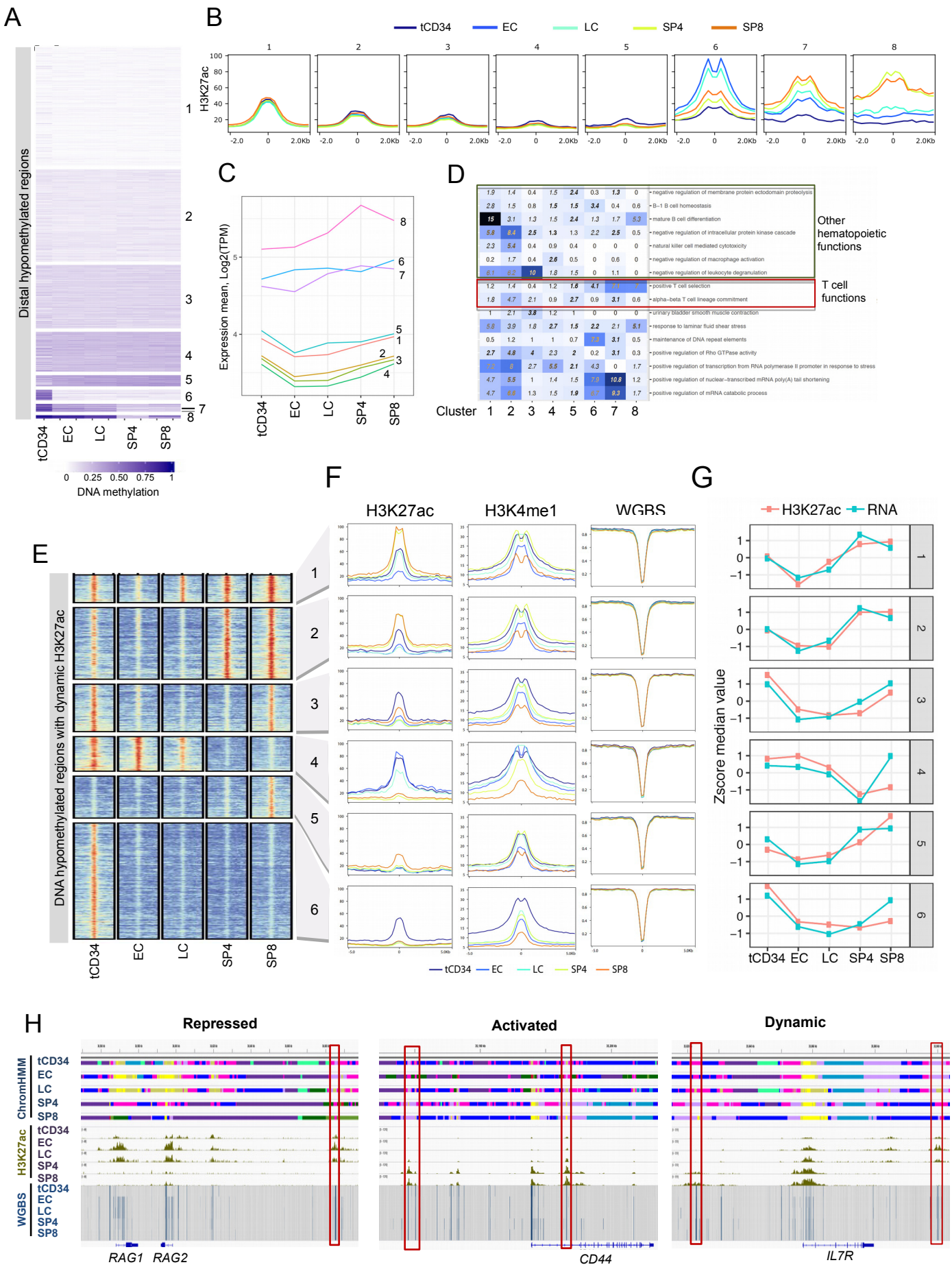

A
B
tCD34
EC
LC
SP4
SP8
H3K27ac
Distal hypomethylated regions
DNA methylation
C
Expression mean, Log2(TPM)
D
Other hematopoietic functions
T cell functions
Cluster
E
DNA hypomethylated regions with dynamic H3K27ac
F
H3K27ac
H3K4me1
WGBS
G
H3K27ac
RNA
Zscore median value
H
Repressed
Activated
Dynamic
ChromHMM
H3K27ac
WGBS
RAG1
RAG2
CD44
IL7R

Figure 2

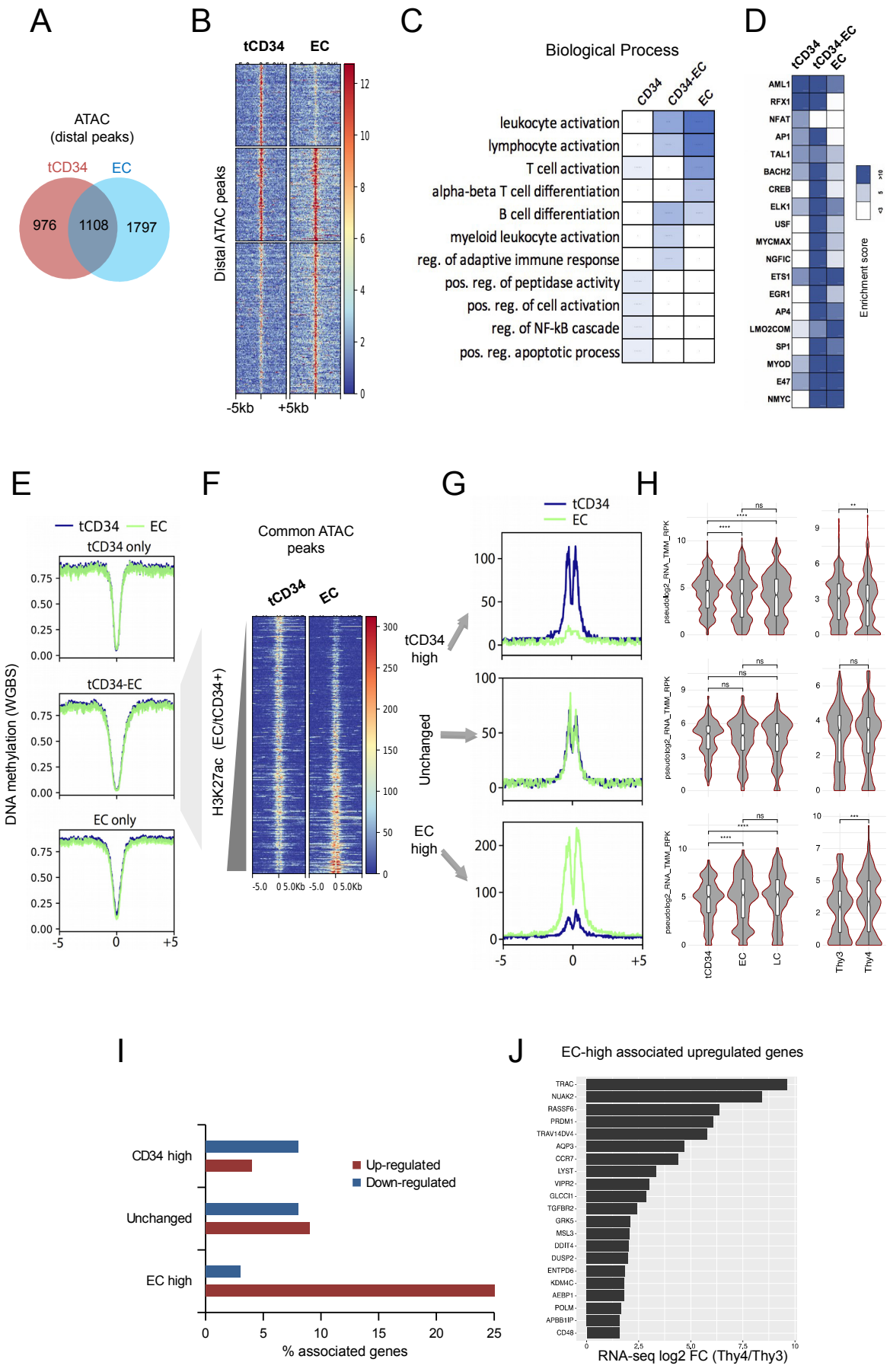

A
ATAC
(distal peaks)
tCD34
EC
976
1108
1797
B
tCD34
EC
Distal ATAC peaks
-5kb
+5kb
C
Biological Process
CD34
CD34-EC
EC
leukocyte activation
lymphocyte activation
T cell activation
alpha-beta T cell differentiation
B cell differentiation
myeloid leukocyte activation
reg. of adaptive immune response
pos. reg. of peptidase activity
pos. reg. of cell activation
reg. of NF-kB cascade
pos. reg. apoptotic process
D
tCD34
tCD34-EC
EC
AML1
RFX1
NFAT
AP1
TAL1
BACH2
CREB
ELK1
USF
MYCMAX
NGFIC
ETS1
EGR1
AP4
LMO2COM
SP1
MYOD
E47
NMYC
Enrichment score
E
tCD34
EC
tCD34 only
tCD34-EC
EC only
DNA methylation (WGBS)
F
Common ATAC
peaks
tCD34
EC
H3K27ac (EC/tCD34+)
G
tCD34
EC
tCD34
high
Unchanged
EC
high
H
pseudolog2_RNA_TMM_RPK
tCD34
EC
LC
Thy3
Thy4
I
CD34 high
Unchanged
EC high
Up-regulated
Down-regulated
% associated genes
J
EC-high associated upregulated genes
TRAC
NUAK2
RASSF6
PRDM1
TRAV14DV4
AQP3
CCR7
LYST
VIPR2
GLCCI1
TGFBR2
GRK5
MSL3
DDIT4
DUSP2
ENTPD6
KDM4C
AEBP1
POLM
APBB1IP
CD48
RNA-seq log2 FC (Thy4/Thy3)

Figure 3

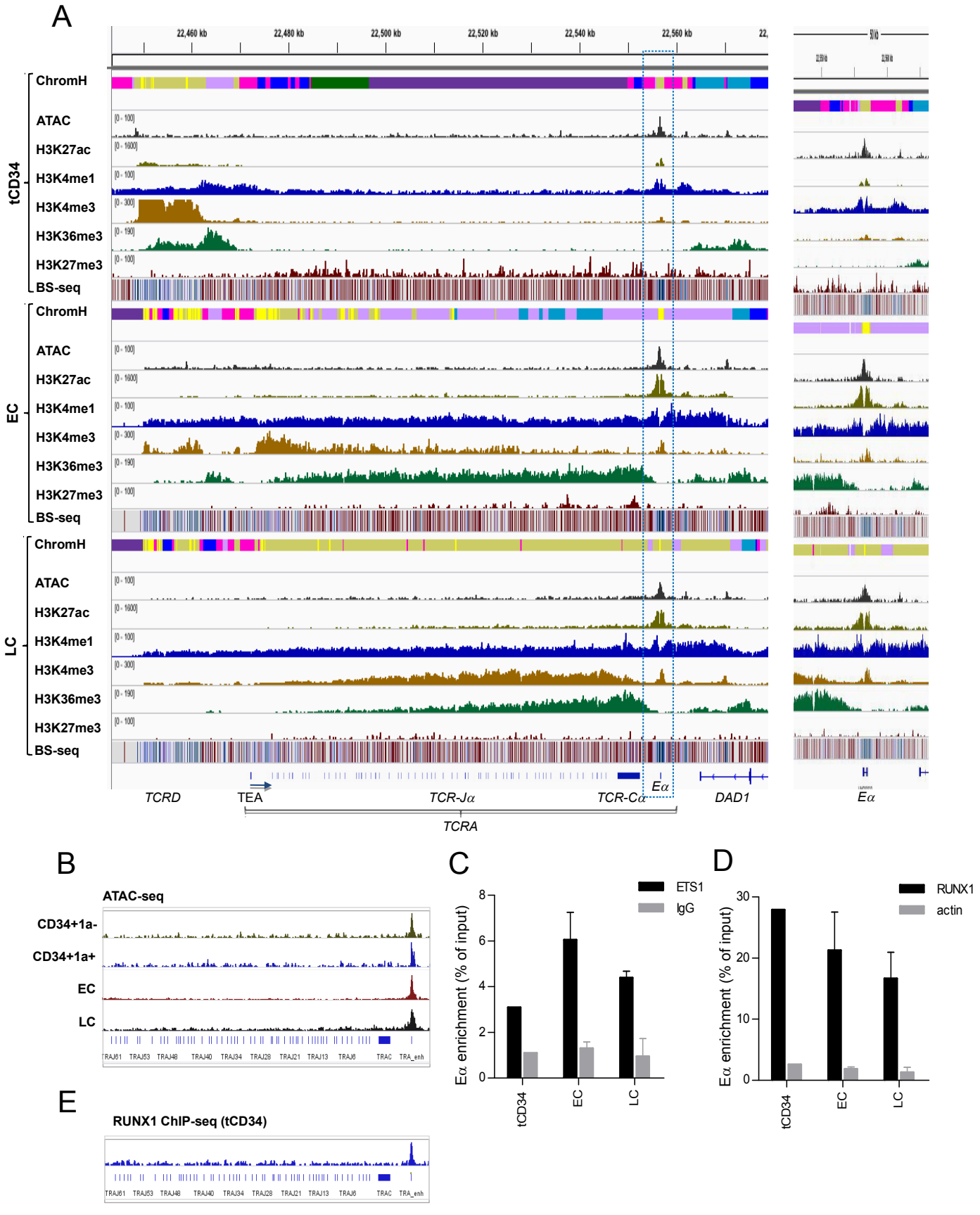

A
22,460 kb
22,480 kb
22,500 kb
22,520 kb
22,540 kb
22,560 kb
tCD34
EC
LC
ChromH
ATAC
H3K27ac
H3K4me1
H3K4me3
H3K36me3
H3K27me3
BS-seq
TCRD
TEA
TCR-Jα
TCR-Cα
Eα
DAD1
TCRA
B
ATAC-seq
CD34+1a-
CD34+1a+
EC
LC
C
ETS1
IgG
Eα enrichment (% of input)
tCD34
EC
LC
D
RUNX1
actin
Eα enrichment (% of input)
E
RUNX1 ChIP-seq (tCD34)

Figure 4

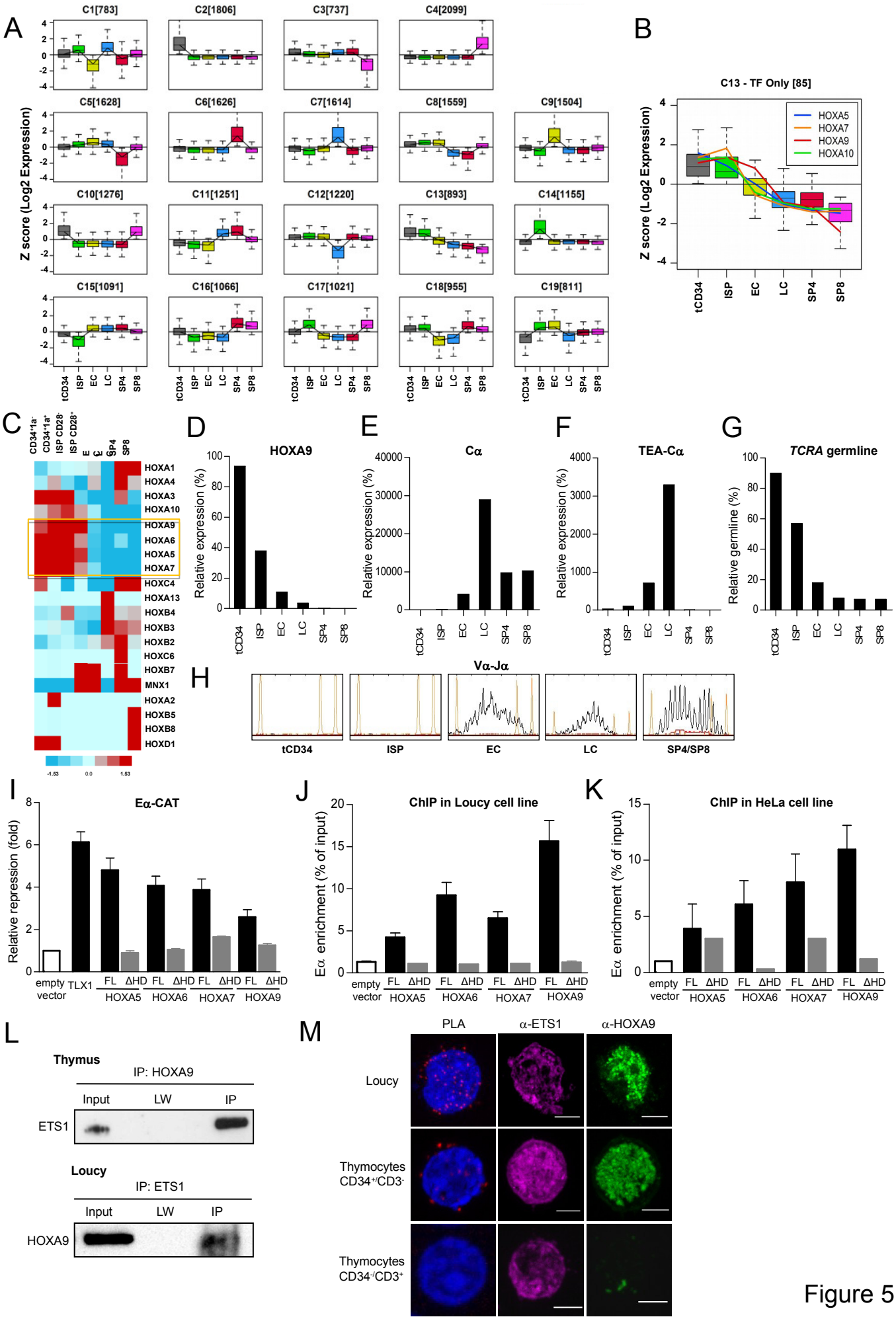

A
C1[783]
C2[1806]
C3[737]
C4[2099]
C5[1628]
C6[1626]
C7[1614]
C8[1559]
C9[1504]
C10[1276]
C11[1251]
C12[1220]
C13[893]
C14[1155]
C15[1091]
C16[1066]
C17[1021]
C18[955]
C19[811]
Z score (Log2 Expression)
tCD34
ISP
EC
LC
SP4
SP8
B
C13 - TF Only [85]
HOXA5
HOXA7
HOXA9
HOXA10
C
CD34+1a-
CD34+1a+
ISP CD28-
ISP CD28+
HOXA1
HOXA4
HOXA3
HOXA10
HOXA9
HOXA6
HOXA5
HOXA7
HOXC4
HOXA13
HOXB4
HOXB3
HOXB2
HOXC6
HOXB7
MNX1
HOXA2
HOXB5
HOXB8
HOXD1
D
HOXA9
Relative expression (%)
E
Cα
F
TEA-Cα
G
TCRA germline
Relative germline (%)
H
Vα-Jα
I
Eα-CAT
Relative repression (fold)
empty vector
TLX1
FL
ΔHD
HOXA5
HOXA6
HOXA7
HOXA9
J
ChIP in Loucy cell line
Eα enrichment (% of input)
K
ChIP in HeLa cell line
L
Thymus
IP: HOXA9
Input
LW
IP
ETS1
Loucy
IP: ETS1
HOXA9
M
PLA
α-ETS1
α-HOXA9
Loucy
Thymocytes CD34+/CD3-
Thymocytes CD34-/CD3+

Figure 5

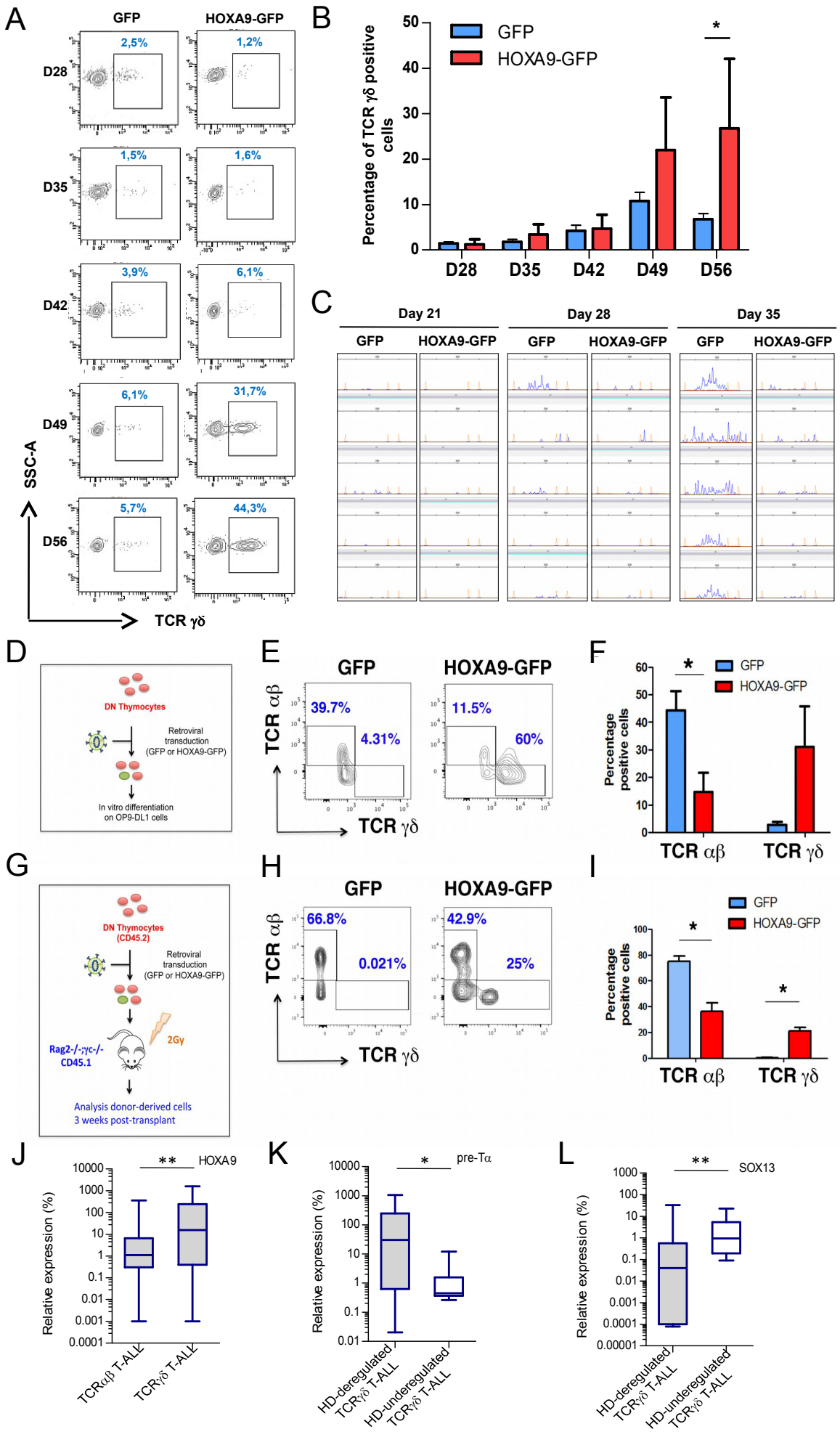
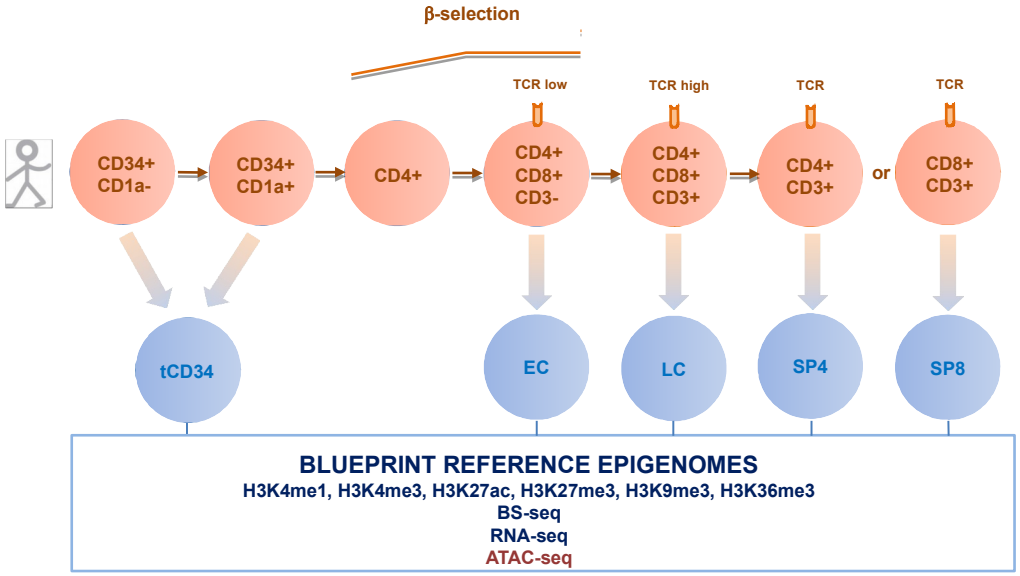
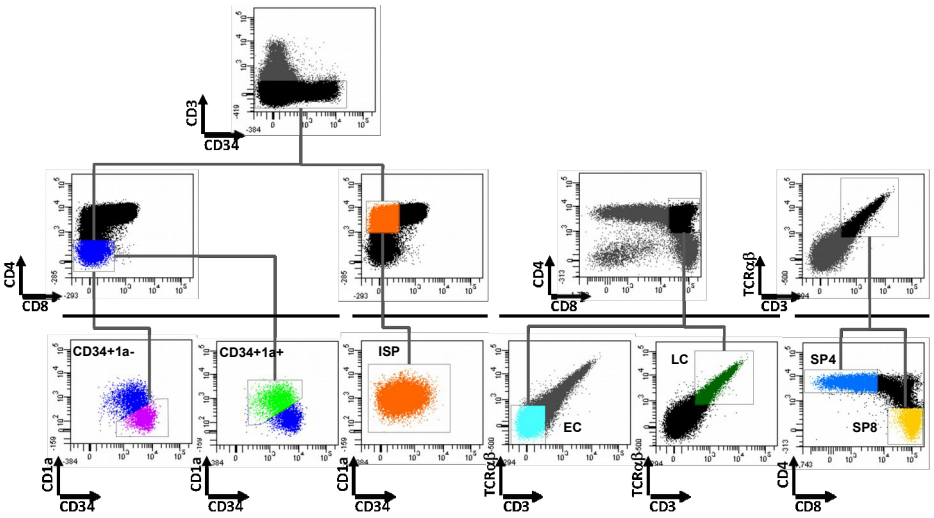

A
GFP
HOXA9-GFP
D28
D35
D42
D49
D56
2,5%
1,2%
1,5%
1,6%
3,9%
6,1%
6,1%
31,7%
5,7%
44,3%
SSC-A
TCR γδ
B
Percentage of TCR γδ positive cells
GFP
HOXA9-GFP
D28
D35
D42
D49
D56
C
Day 21
Day 28
Day 35
GFP
HOXA9-GFP
D
DN Thymocytes
Retroviral transduction (GFP or HOXA9-GFP)
In vitro differentiation on OP9-DL1 cells
E
GFP
HOXA9-GFP
TCR αβ
TCR γδ
39.7%
4.31%
11.5%
60%
F
Percentage positive cells
GFP
HOXA9-GFP
TCR αβ
TCR γδ
G
DN Thymocytes (CD45.2)
Retroviral transduction (GFP or HOXA9-GFP)
Rag2-/-γc-/- CD45.1
2Gy
Analysis donor-derived cells 3 weeks post-transplant
H
GFP
HOXA9-GFP
TCR αβ
TCR γδ
66.8%
0.021%
42.9%
25%
I
Percentage positive cells
GFP
HOXA9-GFP
TCR αβ
TCR γδ
J
HOXA9
Relative expression (%)
TCRαβ T-ALL
TCRγδ T-ALL
K
pre-Tα
Relative expression (%)
HD-deregulated TCRγδ T-ALL
HD-underregulated TCRγδ T-ALL
L
SOX13
Relative expression (%)
HD-deregulated TCRγδ T-ALL
HD-underregulated TCRγδ T-ALL

Figure 6

A

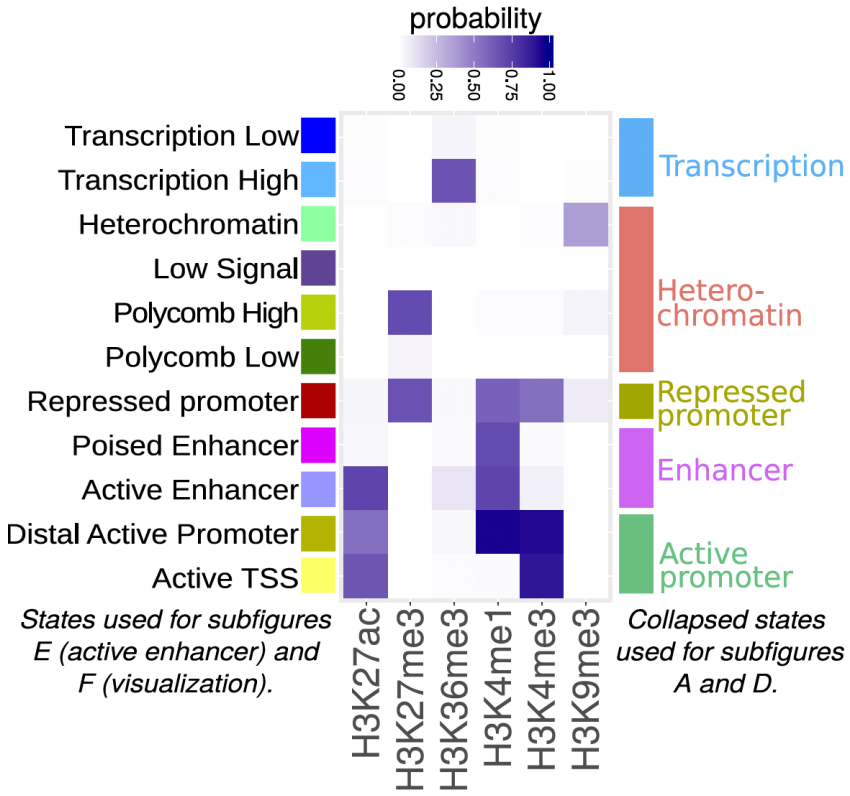


B



Related to Figure 1

Figure S1



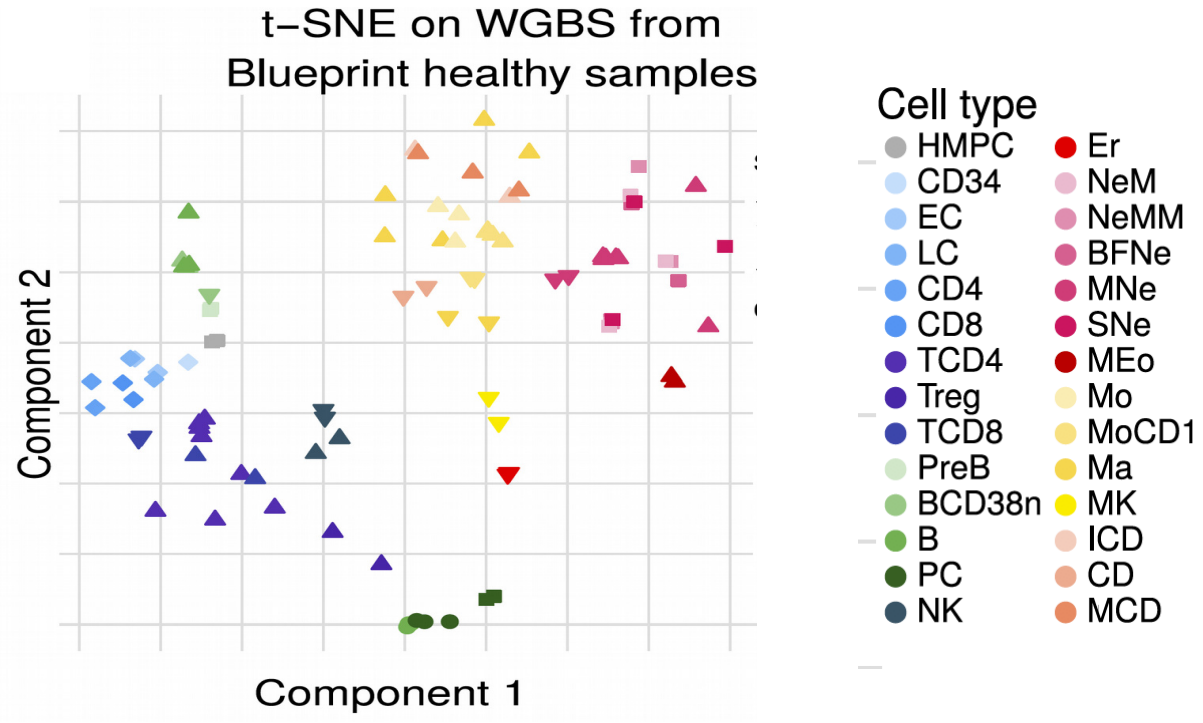
Related to Figure 1

Figure S1C

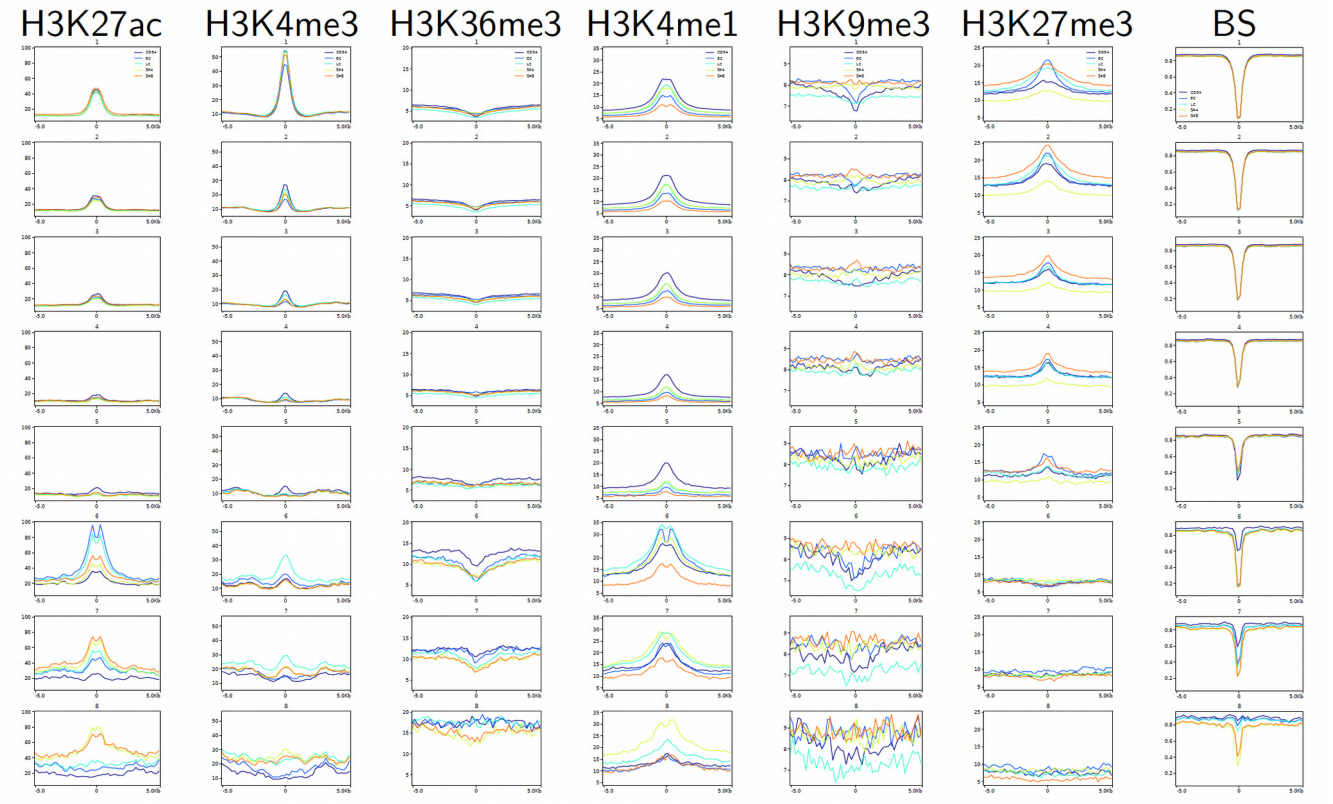
Related to Figure 1D

Figure S1D

A



B

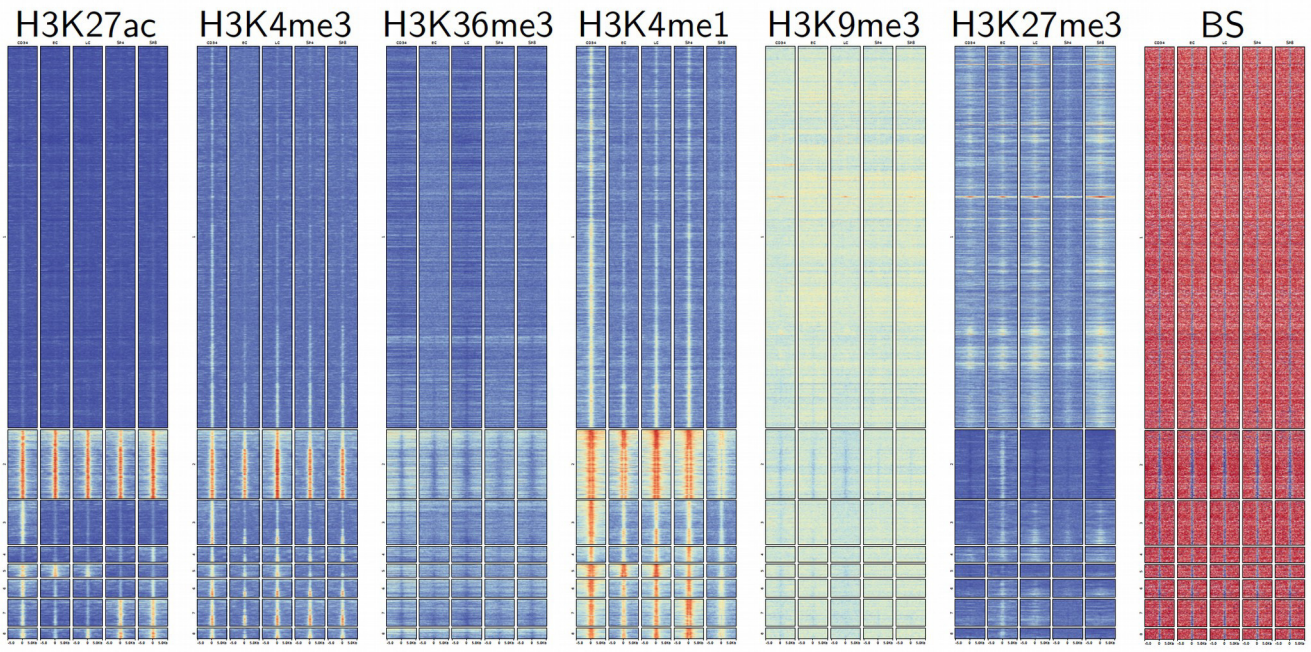


Related to Figure 2

Figure S2

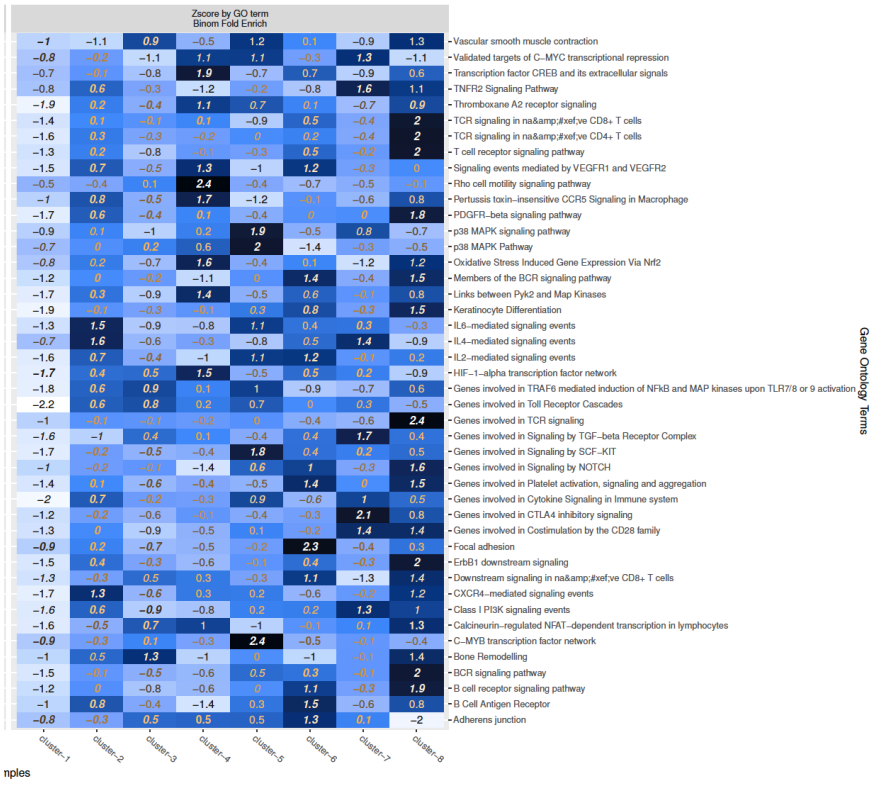
Histone modifications at constitutively hypomethylated regions

C



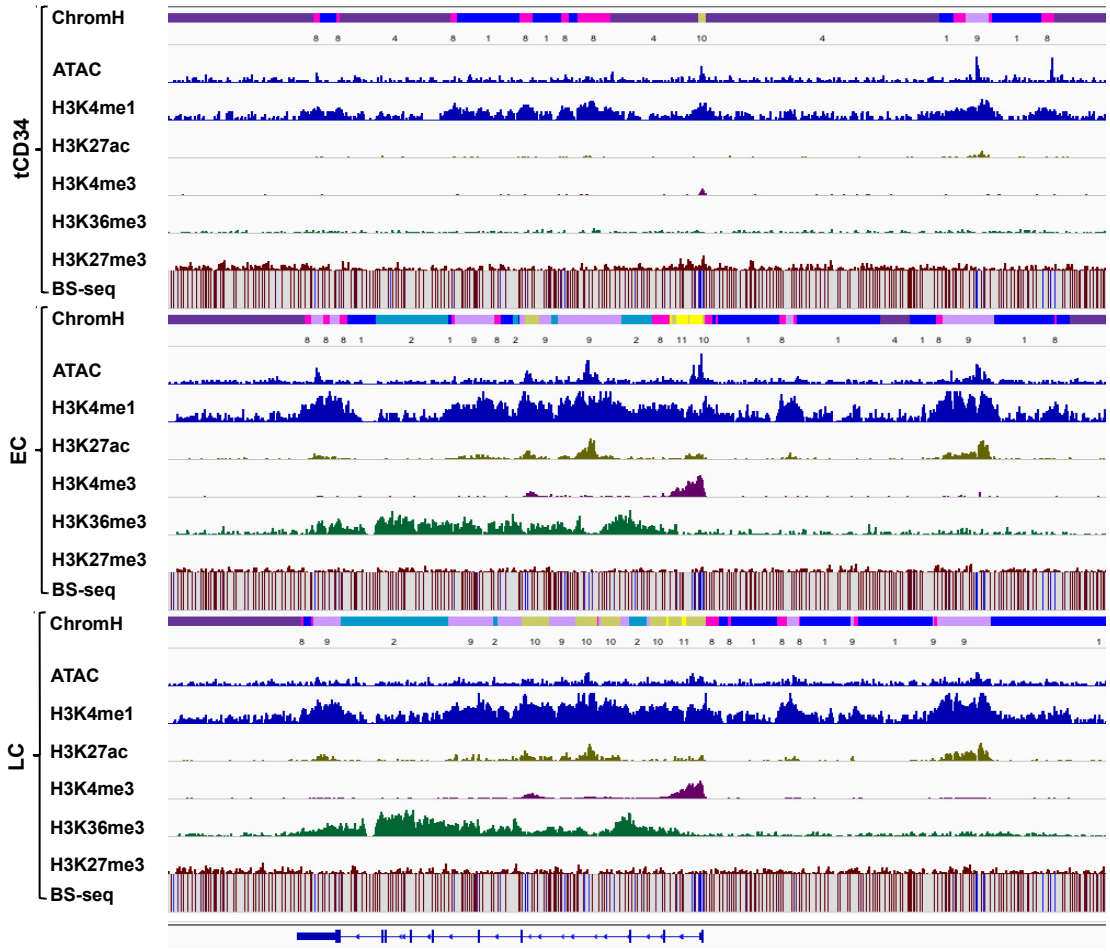
D

GO terms on H3K27ac clusters (DNA hypomethylated regions)



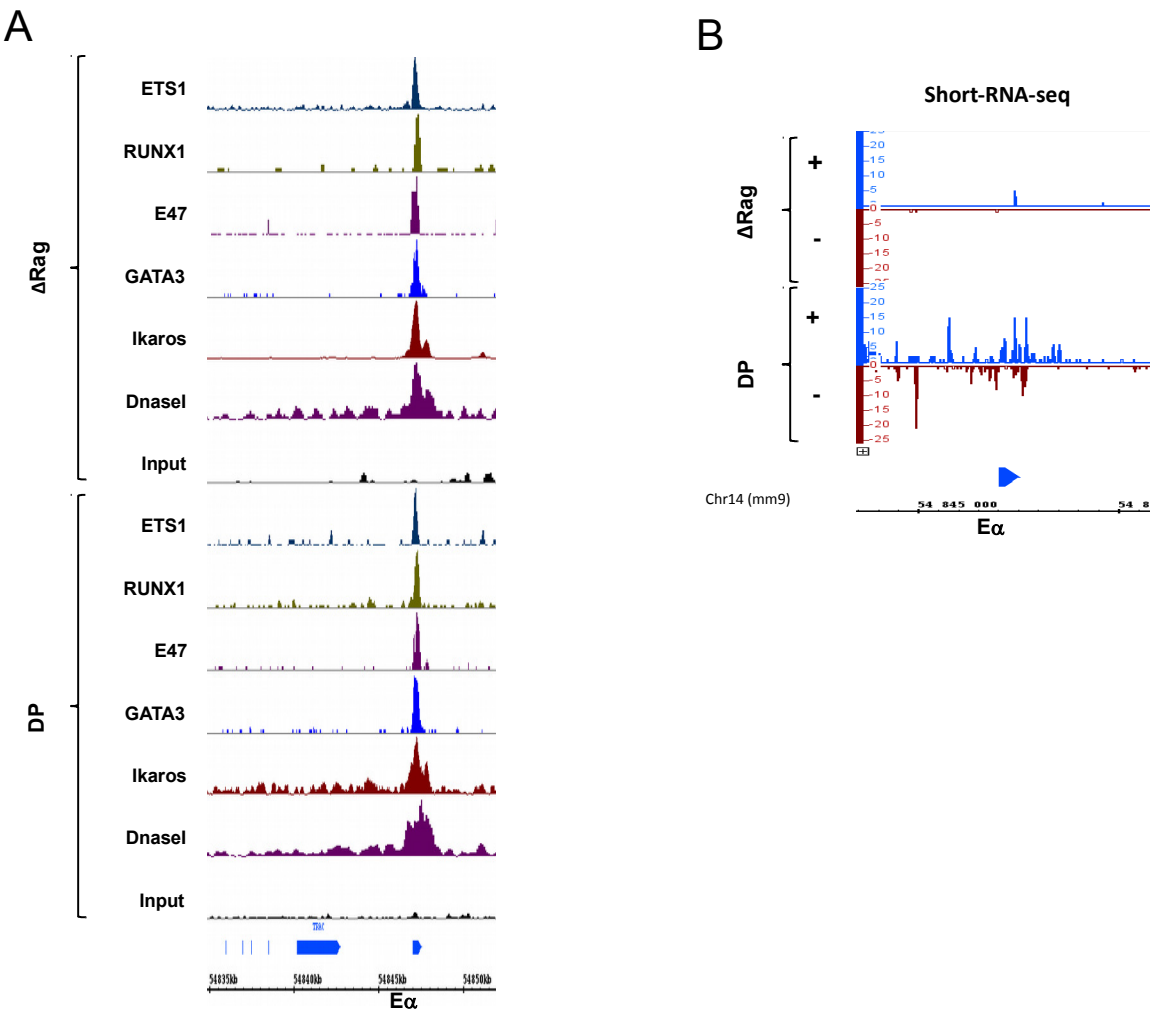
Related to Figure 2E

Figure S2



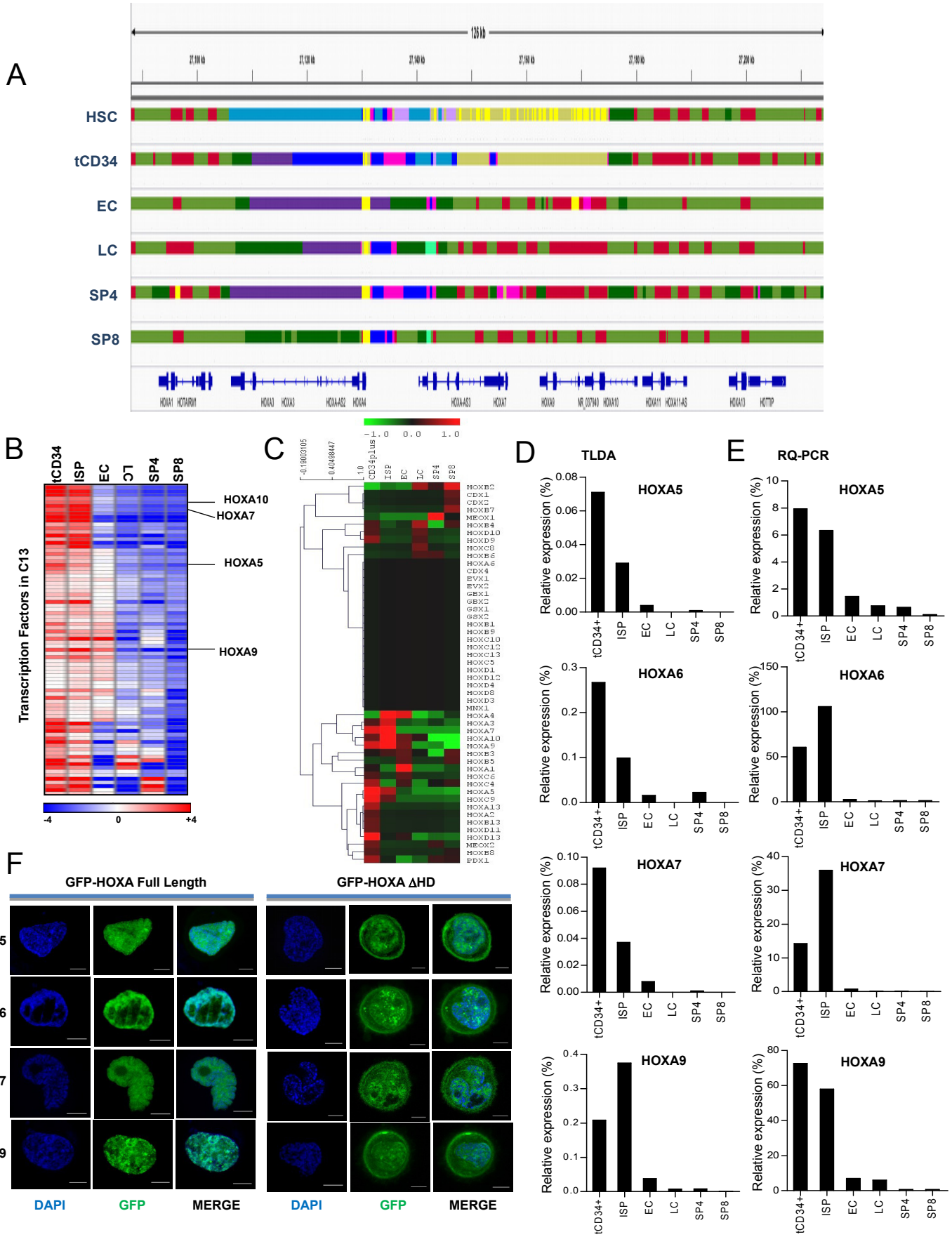
Related to Figure 3

Figure S3



Related to Figure 4

Figure S4



Related to Figure 5

Figure S5

2.2 Régulation épigénétique et remaniements chromatiniens au cours de la spermatogénèse

2.2.1 Nouveau langage d'histones basé sur l'acétylation et la butyrylation des lysines 5 et 8 de l'histone 4

Alors que les acétylations et méthylations de multiples lysines d'histones sont largement caractérisées dans la littérature, les études portant sur d'autres modifications étaient rares il y a encore quelques années. L'équipe du docteur Saadi Khochbin a rapporté la présence de nombre d'entre elles dans les cellules de la lignée geminale mâle murine (TAN, LUO, LEE et al. 2011). Après avoir mis en évidence l'implication de la crotonylation et de la 2-hydroxyisobutyrylation (DAI, PENG, MONTELLIER et al. 2014) des lysines 5 et 8 de l'histone 4 dans le contrôle des programmes d'expression des cellules de la lignée germinale mâle, l'intérêt de l'équipe s'est porté sur les caractéristiques de la butyrylation de ces sites. Le choix de ces modications parmi toutes celles non caractérisées a été guidé par la mise en évidence de Brdt, une protéine à double bromodomaine de la famille des *Bromo- and Extra-Terminal domain* (BET), capable de fixer uniquement l'acétylation simultanée de H4K5 et H4K8 (MORINIÈRE, ROUSSEAUX, STEUERWALD et al. 2009) et agissant comme un régulateur maître du programme d'expression post méiotique (GAUCHER, BOUSSOUAR, MONTELLIER et al. 2012). Étant donné le rôle critique des acétylations de H4K5 et H4K8 dans les activités de Brdt, l'hypothèse initiale était que les régions marquées par d'autres modifications mutuellement exclusives auraient des propriétés fonctionnelles différentes.

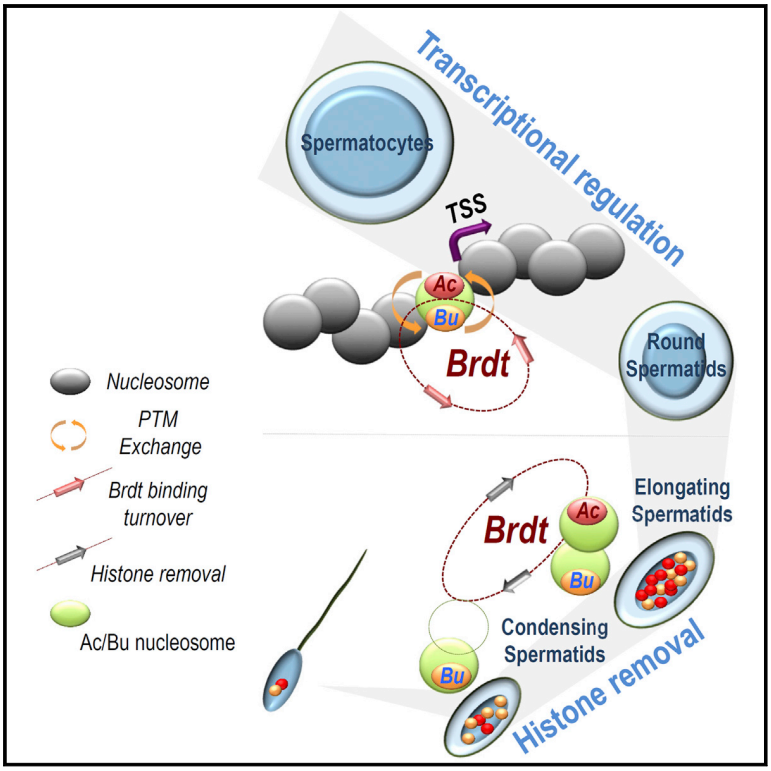
Différentes approches expérimentales ont été appliquées pour vérifier cette hypothèse et caractériser la butyrylation. Les résultats obtenus ont fait l'objet d'un article (ARTICLE 2 ; GOUDARZI, ZHANG, HUANG et al. 2016). En tant que co-auteur, mon rôle a été d'analyser les ChIP-seq de H4K5ac, H4K5bu, H4K8ac et H4K8bu réalisés pour des spermatocytes au stade pachytène et des spermatides rondes. Ces différentes marques se sont révélées être particulièrement colocalisées entre elles à chaque stade, en particulier au niveau des promoteurs de gènes (FIGURE 3A de l'ARTICLE 2). Le croisement avec des données de biopuces disponibles dans la littérature (MONTELLIER, BOUSSOUAR, ROUSSEAUX et al. 2013) a permis de montrer que la présence dynamique de ces marques sur un promoteur est associée à niveau élevé d'activité transcriptomique de son gène relatif (FIGURES 3B et 4 de l'ARTICLE 2). Les approches expérimentales réalisées par les autres auteurs ont permis de préciser comment la présence compétitive de l'acétylation et de la butyrylation est permise et les conséquences sur les méchanismes de régulation transcriptionnelle et de remplacement des histones médiés par Brdt.

Article

Molecular Cell

Dynamic Competing Histone H4 K5K8 Acetylation and Butyrylation Are Hallmarks of Highly Active Gene Promoters

Graphical Abstract



Highlights

- Active gene TSSs are marked by competing H4 K5K8 acetylation and butyrylation
- Histone butyrylation directly stimulates transcription
- H4K5 butyrylation prevents binding of the testis specific gene expression-driver Brdt
- H4K5K8 butyrylation is associated with delayed histone removal in spermatogenic cells

Authors

Afsaneh Goudarzi, Di Zhang, He Huang, ..., Robert G. Roeder, Yingming Zhao, Saadi Khochbin

Correspondence

yzhao2@bsd.uchicago.edu (Y.Z.), saadi.khochbin@univ-grenoble-alpes.fr (S.K.)

In Brief

Histone butyrylation stimulates gene transcription while competing with acetylation at H4K5 to control Brdt bromodomain binding. Differential chromatin labeling with interchangeable H4 acylations is an important epigenetic regulatory mechanism controlling gene expression and chromatin reorganization.

Accession Numbers

GSE77277
GSE39910

Goudarzi et al., 2016, Molecular Cell 62, 169–180
April 21, 2016 ©2016 The Authors
http://dx.doi.org/10.1016/j.molcel.2016.03.014

Molecular Cell
Article

Dynamic Competing Histone H4 K5K8 Acetylation and Butyrylation Are Hallmarks of Highly Active Gene Promoters

Afsaneh Goudarzi,[1,8] Di Zhang,[2,8] He Huang,[2] Sophie Barral,[1] Oh Kwang Kwon,[2] Shankang Qi,[2] Zhanyun Tang,[3] Thierry Buchou,[1] Anne-Laure Vitte,[1] Tieming He,[4] Zhongyi Cheng,[4] Emilie Montellier,[1] Jonathan Gaucher,[1,5] Sandrine Curtet,[1] Alexandra Debernardi,[1] Guillaume Charbonnier,[6] Denis Puthier,[6] Carlo Petosa,[7] Daniel Panne,[5] Sophie Rousseaux,[1] Robert G. Roeder,[3] Yingming Zhao,[2,9,*] and Saadi Khochbin[1,9,*]
[1]CNRS UMR 5309, INSERM, U1209, Université Grenoble Alpes, Institut Albert Bonniot, 38700 Grenoble, France
[2]Ben May Department of Cancer Research, The University of Chicago, Chicago, IL 60637, USA
[3]Laboratory of Biochemistry and Molecular Biology, The Rockefeller University, New York, NY 10065, USA
[4]Jingjie PTM Biolab (Hangzhou) Co., Ltd., Hangzhou 310018, China
[5]EMBL Grenoble, BP 181, 71 Avenue des Martyrs, 38042 Grenoble Cedex 9, France
[6]TAGC, UMR, S 1090 INSERM Aix-Marseille Université, U928 Parc Scientifique de Luminy case 928 163, Avenue de Luminy, 13288 Marseille Cedex 9, France
[7]Université Grenoble Alpes/CEA/CNRS, Institut de Biologie Structurale, 38027 Grenoble, France
[8]Co-first author
[9]Co-senior author
*Correspondence: yzhao2@bsd.uchicago.edu (Y.Z.), saadi.khochbin@univ-grenoble-alpes.fr (S.K.)
http://dx.doi.org/10.1016/j.molcel.2016.03.014

SUMMARY

Recently discovered histone lysine acylation marks increase the functional diversity of nucleosomes well beyond acetylation. Here, we focus on histone butyrylation in the context of sperm cell differentiation. Specifically, we investigate the butyrylation of histone H4 lysine 5 and 8 at gene promoters where acetylation guides the binding of Brdt, a bromodomain-containing protein, thereby mediating stage-specific gene expression programs and post-meiotic chromatin reorganization. Genome-wide mapping data show that highly active Brdt-bound gene promoters systematically harbor competing histone acetylation and butyrylation marks at H4 K5 and H4 K8. Despite acting as a direct stimulator of transcription, histone butyrylation competes with acetylation, especially at H4 K5, to prevent Brdt binding. Additionally, H4 K5K8 butyrylation also marks retarded histone removal during late spermatogenesis. Hence, alternating H4 acetylation and butyrylation, while sustaining direct gene activation and dynamic bromodomain binding, could impact the final male epigenome features.

INTRODUCTION

Besides lysine acetylation, we recently identified a variety of short-chain lysine acylations in core histones, including lysine propionylation, butyrylation, 2-hydroxyisobutyrylation, crotonylation, malonylation, succinylation, and glutarylation (Chen et al., 2007; Dai et al., 2014; Tan et al., 2011, 2014; Xie et al., 2012). Emerging data suggest that these new histone lysine acylations may have unique functions that depend not only on cell metabolism, but also on their ability to be deposited or removed by specific enzymes (Dai et al., 2014; Montellier et al., 2012; Rousseaux and Khochbin, 2015; Sabari et al., 2015; Sin et al., 2012; Tan et al., 2011). Nevertheless, the functional impact of differential histone acylation on chromatin recognition by specific factors has remained unexplored.

This study aims to understand the functional consequences of differential histone acylation. In particular, we decided to investigate histone butyrylation, because, in contrast to the acetyl (2-carbon) and propionyl (3-carbon) groups, the butyryl (4-carbon) group restricts the binding of bromodomains (Flynn et al., 2015). More specifically, we focused our attention on histone H4 at K5 and K8, whose acetylation is required to bind the first bromodomain of Brdt, a testis-specific member of the BET protein family (Morinière et al., 2009). Our previous work showed that Brdt stimulates the transcription of certain spermatogenic-specific genes by recruiting the P-TEFb complex and by directly binding to their transcriptional start sites (TSSs). Additionally, during late spermatogenesis, Brdt's first bromodomain is necessary for the replacement of histones by non-histone sperm-specific transition proteins (TPs) and protamines (Prms) (Gaucher et al., 2012). Given the critical role of H4K5 and H4K8 acetylation in Brdt-driven activities, we hypothesized that other mutually exclusive histone marks at these two residues might have key regulatory roles in sperm cell genome programming.

Here, we identify major histone lysine butyrylation sites in cells from different species, including mouse spermatogenic cells. Using spermatogenesis as an integrated biological model system, in addition to in vitro experiments and targeted proteomic approaches, we demonstrate new characteristics of active gene TSSs. Our data indicate that interchangeable acetylation

Molecular Cell 62, 169–180, April 21, 2016 ©2016 The Authors 169

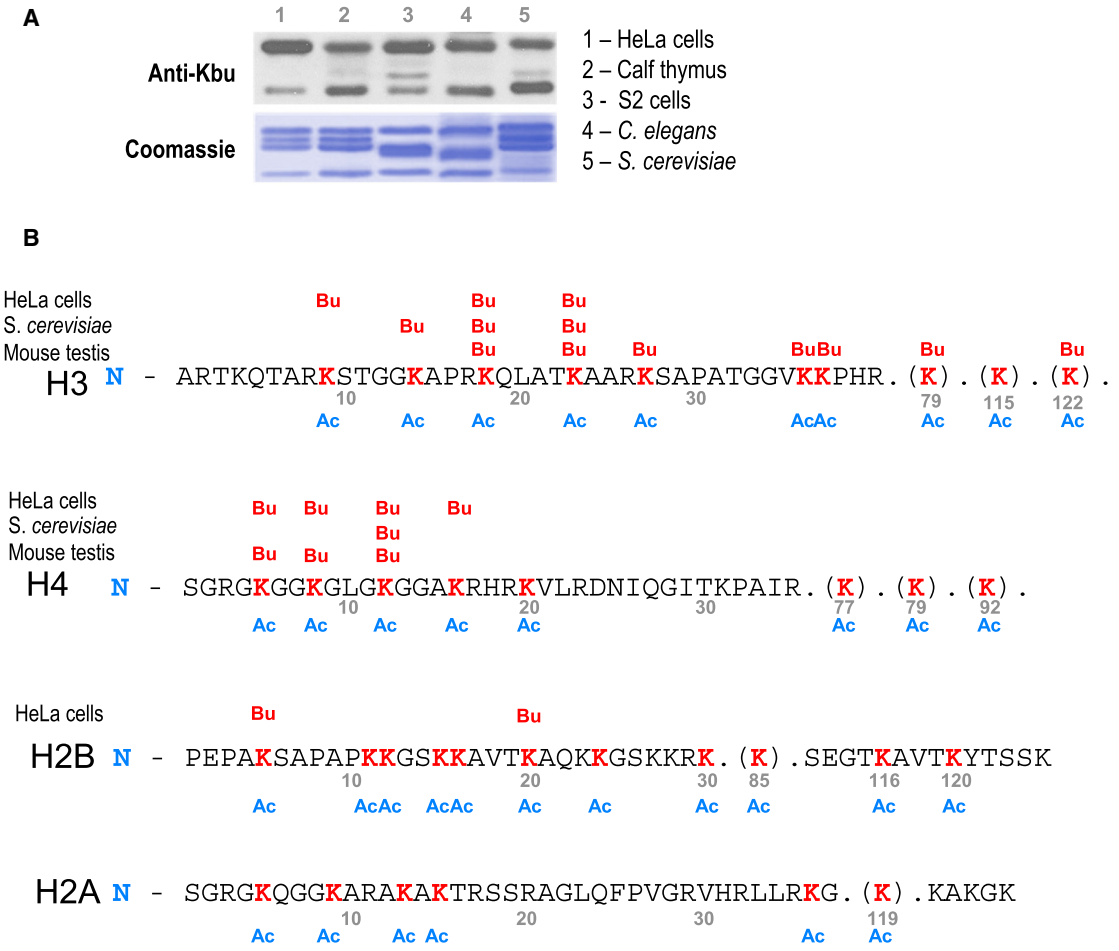


Figure 1. Histone Lysine Butyrylation Is an Evolutionarily Conserved PTM in Eukaryotic Cells

(A) The Kbu residues in core histones from the indicated species were detected by western blotting using a pan anti-Kbu antibody (upper). The corresponding Coomassie blue stained gel is shown (lower).

(B) Illustrations of histone Kac and Kbu sites in core histones identified by tandem mass spectrometry (MS/MS) (acetyl, Ac and butyryl, Bu). The annotated MS/MS spectra for histone Kbu peptides and the specific co-occurrence of K5bu and K8bu in H4 from spermatogenic cells are shown in Data S1.

and butyrylation at H4K5 and H4K8 not only stimulates transcription, but could also underlie a highly dynamic interaction of histone post-translational modification (PTM)-binding factors such as Brdt. Additional data further show that stable differential use of acetylation and butyrylation could also durably affect genome organization in the maturing sperm. Altogether, these findings indicate how competition between histone acylation states could be an important epigenetic regulatory mechanism.

RESULTS

Histone Lysine Butyrylation Is an Evolutionarily Conserved PTM

To identify histone butyryllysine (Kbu) sites and study their function, we first confirmed the presence of histone Kbu by western blotting. Our data suggest that histone Kbu is an evolutionarily conserved PTM in eukaryotic cells (Figure 1A). We then used mass spectrometry to identify possible Kbu sites in core histones from three species (Chen et al., 2007; Kim et al., 2006). Kbu sites were detected in the N-terminal tails of H3 (K9, K14, K18, K23, K27, K36, K37, K79, and K122), H4 (K5, K8, K12, and K16), and H2B (K5 and K20) (Figure 1B; Data S1).

Functional Significance of Histone Butyrylation

To examine the function of Kbu in chromatin biology, we used mammalian spermatogenesis as a model system. Spermatogenic cells can be roughly classified into three major types: proliferative progenitor cells (spermatogonia), meiotic cells (spermatocytes), and post-meiotic haploid cells (spermatids). During differentiation, highly specific gene expression programs are activated in both meiotic and early post-meiotic cells. Large-scale genome reorganization also takes place in spermatids, where a genome-wide replacement of histones by TPs and Prms occurs in post-meiotic cells known as elongating and condensing spermatids (Gaucher et al., 2010; Goudarzi et al., 2014; Govin et al., 2004).

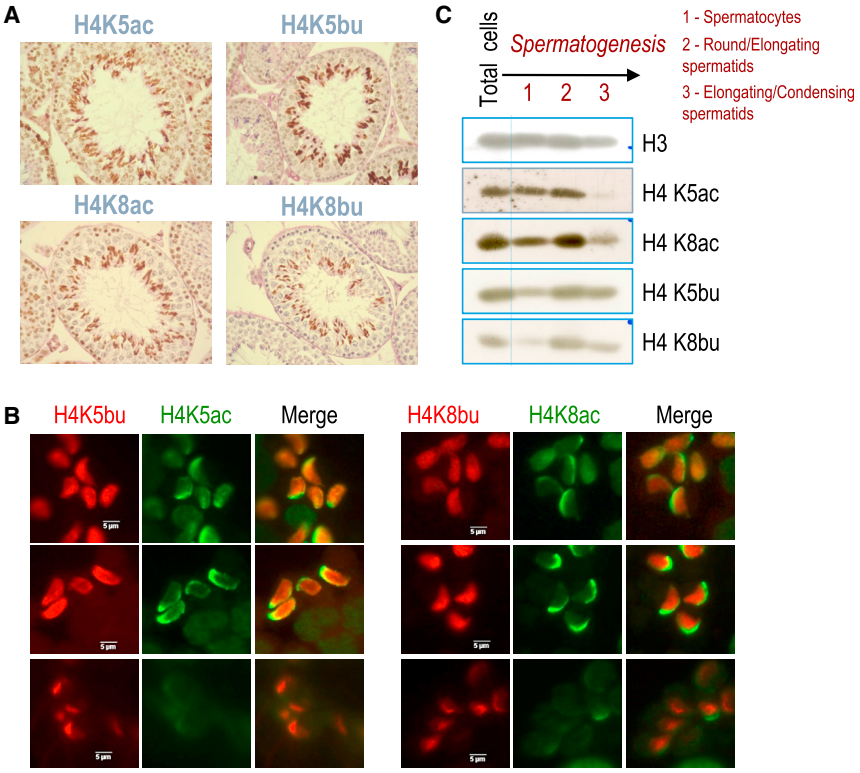


Figure 2. Occurrence of H4K5ac, K5bu, K8ac, and K8bu Marks during Mouse Spermatogenesis

(A and B) Detection by immunohistochemistry on paraffin sections of staged testis tubules (A) and by immunofluorescence in elongating/condensing spermatids (B) using the indicated antibodies are shown. The tubule sections shown in (A) represent stage IX or X of spermatogenesis. The use of mouse monoclonal anti-H4K5ac or H4K8ac antibodies together with rabbit polyclonal anti-H4K5bu or H4K8bu antibodies allowed for the co-detection of the indicated marks. (B) Spermatids are presented as a function of their differentiation along spermiogenesis (upper to lower) as judged by their morphology and degree of genome compaction. The scale bars represent 5 μm. The images were acquired and treated under the same conditions.

(C) Total extracts from fractionated spermatogenic cells corresponding to cells enriched at the indicated stages along with extracts from the total (non-fractionated) spermatogenic cells were used to detect the indicated marks by western blotting. Compared to acetylation, longer exposure times were required for the butyrylation signal to be detected in all samples.

To confirm the existence of histone butyrylation in spermatogenic cells, we identified histone Kbu sites by mass spectrometry analysis of histones from mouse testis. We detected ten butyrylation sites including H4K5bu and H4K8bu (occurring separately or in combination), supporting the presence of these two histone marks in spermatogenic cells (Figure 1B; Data S1; see also Figure 7B). We then used highly specific anti-H4K5bu and anti-H4K8bu antibodies, along with the anti-H4K5ac and anti-H4K8ac antibodies, to investigate the stage-specific presence of these marks in spermatogenic cells. Immunohistochemistry analysis showed that H4K5 and K8 butyrylation is enhanced in elongating spermatids (Figure 2A), similar to earlier observations for histone H4K5 and H4K8 acetylation (Hazzouri et al., 2000).

To study the dynamic changes of butyrylation versus acetylation at both H4K5 and H4K8 sites in spermatogenic cells, we examined the co-existence of H4K5ac and H4K5bu as well as that of H4K8ac and H4K8bu. In elongating spermatids, H4K5ac and H4K8ac are widely distributed, but their localization becomes biased toward the sub-acrosomal regions in later stages. In contrast, H4K5bu- and H4K8bu-containing nucleosomes are homogenously distributed in the same cells (Figure 2B). In late elongating spermatids, while acetylated histones are removed and degraded (Gaucher et al., 2012; Qian et al., 2013), butyrylated H4 species escape this wave of acetylation-dependent histone removal, to finally disappear in condensing spermatids (Figure 2B). Interestingly, immunoblotting experiments, using fractionated spermatogenic cells, confirmed that butyrylation persists longer than acetylation on H4K5 and K8 sites (Figure 2C). These results highlight a bimodal histone removal process, whereby the removal of H4 K5/K8 butyrylated nucleosome occurs after that of H4 K5/K8 acetylated nucleosomes.

Genome-wide Distribution of H4K5 and H4K8 Acetylation and Butyrylation

Our immunofluorescence analysis showed distinctive patterns of acetylation and butyrylation at H4K5 and H4K8, with both marks showing enhanced labeling in elongating spermatids, when cellular transcription dramatically decreases (Gaucher et al., 2010). This observation motivated us to study their genomic distributions and potential functions at earlier stages of spermatogenesis, when specific spermatogenic gene expression programs are activated (Dai et al., 2014; Gaucher et al., 2012; Tan et al., 2011). Toward this goal, mouse spermatogenic cells were fractionated into spermatocytes and post-meiotic round spermatids (Dai et al., 2014; Gaucher et al., 2012; Tan et al., 2011). The two pools of cells were subjected to chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) using four anti-PTM-specific antibodies (anti-H4K5ac, -H4K8ac, -H4K5bu, and -H4K8bu). Analysis of the genomic distribution of these marks revealed that regions surrounding gene TSSs (located upstream of TSSs and at the 5′UTR) exhibit the highest coverage by the four PTMs in both spermatocytes and spermatids (Figure 3A). Further analyses showed that the four H4 PTMs are enriched at TSSs in a manner dependent on the transcriptional activity of the corresponding genes (Figure 3B).

A critical concern in the above experiments is the possibility that histone sites subject to acetylation may be butyrylated at

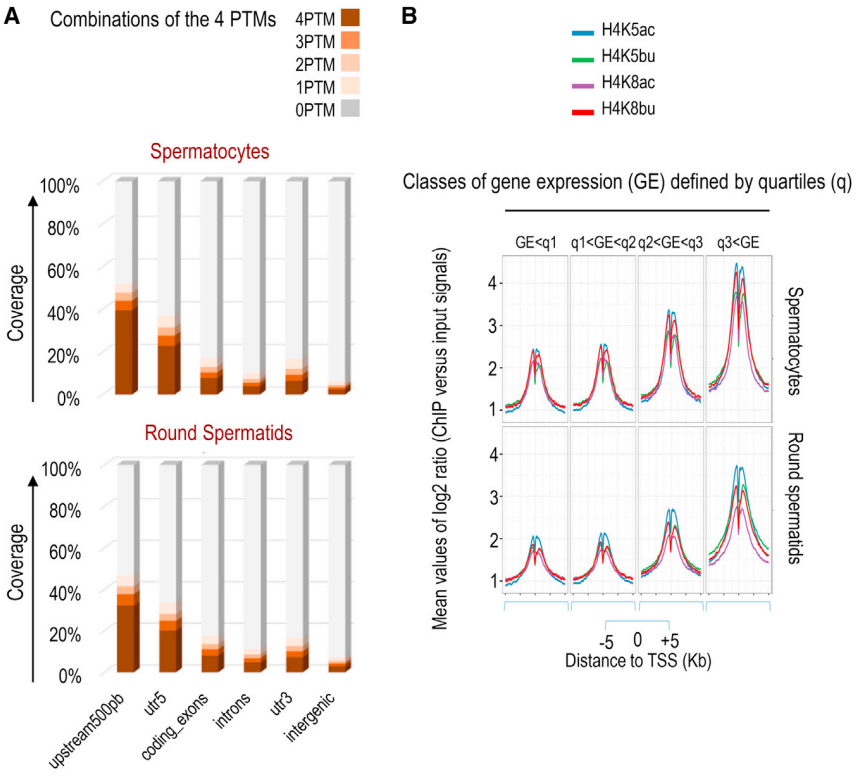


Figure 3. Genome-wide Mapping of H4K5ac, K5bu, K8ac, and K8bu Marks in Meiotic Spermatocytes and Post-meiotic Round Spermatids

(A) Bar diagrams showing the coverage of each of the genomic elements (as indicated) by different combinations of 4, 3, 2, 1, and 0 marks. 100% represents the total annotated genomic element in base pairs.

(B) Log2 ratio of ChIP-seq versus input signal corresponding to the four histone marks (with the indicated color code) is visualized over the gene TSS regions (± 5,000 bp). There are four gene categories that are ordered from left to right according to the level of their expression in the indicated cell types, spermatocytes, or round spermatids. To this aim, we used our transcriptomic analyses of these cells (Montellier et al., 2013). Antibody characterizations and ChIP-qPCR on selected regions are shown in Figure S1 and Table S1.

only low background levels, which would be misleadingly overestimated by the ChIP-seq analysis due to vastly different affinities of the antibodies used. To address this issue, we used surface plasmon resonance (SPR) to measure the affinities of antibodies for their respective targets. These measurements showed that all four antibodies have similar ranges of affinity (Figure S1A). Additionally, a ChIP-qPCR approach demonstrated that the four histone marks are significantly detected at selected genomic regions (Figures S1B and S1C). Together, these experiments confirmed that H4K5K8 butyrylation occurs at levels that largely exceed background noise.

Following these control experiments, we investigated in more detail the relationship between gene expression and the co-occurrence of the four TSS-associated histone marks. Remarkably, the most active genes were found associated with all four marks at their TSS regions (Figures 3B and 4A). In contrast, genes lacking any one of these marks at their TSS showed significantly reduced expression (Figure 4A). Other genomic elements did not show such a direct relationship between the co-existence of the four histone marks and gene expression (Figure S2).

To further analyze the relationship between the four acylation histone marks and gene activity, we took advantage of the differences in gene expression programs between spermatocytes (meiotic cells) and the transcriptionally active haploid round spermatids (generated after meiosis). Using our ChIP-seq data from these two cell types, we divided genes into four categories according to the intensity of TSS labeling by the four H4 acylation marks, namely: (1) genes bearing none of the four marks (labeled "no") and (2–4) genes bearing all four marks, either with (2) comparable ("="), (3) higher (">"), or (4) lower ("<") peak intensities in spermatocytes compared to round spermatids (Figure 4B). We observed that the four groups of genes belong to different gene expression programs. Genes in category (1) are largely unexpressed, while those in category (2) exhibited no change in expression level between meiotic and post-meiotic cells. In striking contrast, genes in categories (3) and (4) exhibited differential expression in the two cell types, which was positively associated with the change in intensity of TSS labeling by the four histone marks (Figure 4B).

Additional support for a positive correlation between the presence of the four PTMs in the TSSs and gene expression is the observation that all four marks are depleted on the sex chromosomes compared to autosomes, consistent with the chromosome-wide meiotic transcriptional inactivation known to characterize sex chromosomes (Figure 4C). To further investigate this observation, we specifically considered the fraction of sex chromosome-linked genes that escapes meiotic sex chromosome inactivation (Namekawa et al., 2006). While all four PTMs were identified on a majority (73%) of TSSs associated with active autosomal genes, co-occurrence of the four PTMs was observed in only a minority (23%) of TSSs associated with sex-linked genes that are reactivated in post-meiotic cells. Hence, over three-quarters of this latter category of TSSs bear between zero and three of the four histone marks. Interestingly, most of the TSSs bearing 1–3 of the PTMs were depleted of H4K8ac, but not of H4K8bu (Figure 4C). This is consistent with other studies showing depletion of acetylation on the TSSs of sex-linked post-meiotic genes and their labeling with other acyl groups such as crotonyl (Sin et al., 2012; Tan et al., 2011).

Histone Butyrylation Directly Stimulates Transcription

Our ChIP-seq data showed that, like acetylation, histone butyrylation is associated with high levels of gene expression,

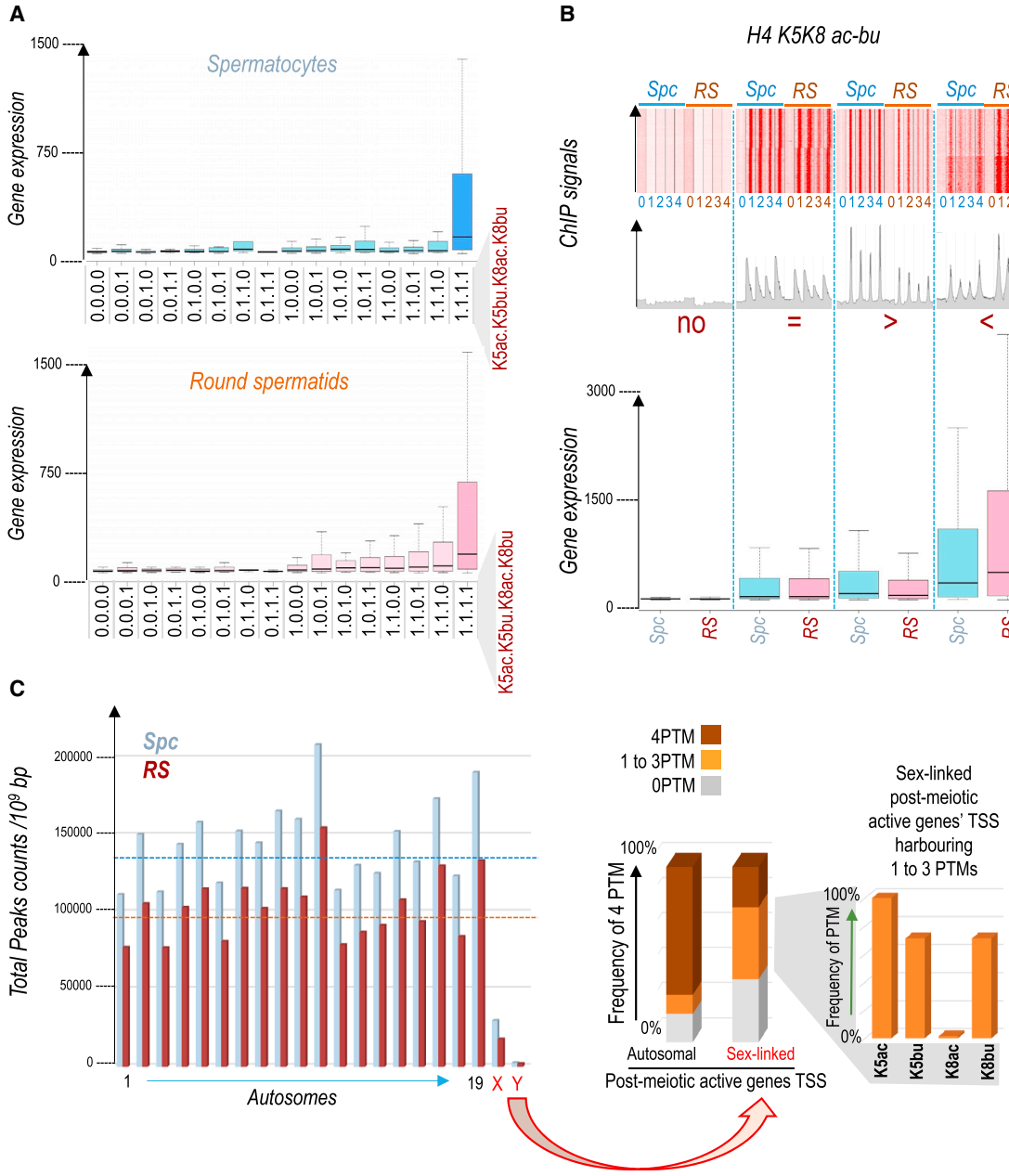


Figure 4. Tight Relationship between TSSs Labeled with H4K5ac, K5bu, K8ac, and K8bu Marks and Gene Transcriptional Activity

(A) The transcriptional activities of genes whose TSS regions bear the indicated histone H4 modifications are shown as box plots in spermatocytes (upper) and round spermatids (lower). The absence or presence of a histone mark is represented by the numbers "0" and "1", respectively, and the four histone marks are arranged from bottom to top as follows: H4K5ac, H4K5bu, H4K8ac, and H4K8bu. The gene transcriptional activities as a function of various combinations of the four H4 PTMs on different genomic elements are shown in Figure S2.

(B) The heatmaps and profiles (upper) show the peak intensities of the ChIP signal at the TSS regions of genes associated with each of the four histone marks ("0" is input; "1" is K5ac; "2" is K5bu; "3" is K8ac; and "4" is K8bu), in spermatocytes ("Spc"), and in round spermatids ("RS"). The genes were assigned to four groups corresponding to: (1) genes with no ChIP peaks ("no"), (2) genes with moderate intensity of TSS labeling by the four marks and no significant change between spermatocytes and round spermatids (=), (3) genes with variation in TSS labeling intensities by the four marks with high meiotic peak intensity and a decreased intensity in post-meiotic cells (>), and (4) genes with an increased peak intensity in post-meiotic cells compared to spermatocytes (<). The box plots (lower) show the expression of the corresponding genes in meiotic and in round spermatids. These data were obtained using transcriptomic data of control samples from our previous work (Montellier et al., 2013; GSE46136).

(C) Total peak counts (acetyl and butyryl H4s) per million base pair for each chromosome in spermatocytes (blue bars) and round spermatids (red bars) are shown. The expected random distributions of the ChIP-seq peaks are indicated as dashed lines (upper). The right panel shows the proportion of gene TSSs harboring 0,

(legend continued on next page)

suggesting the possibility that histone butyrylation directly stimulates gene expression. To test this hypothesis, we exploited a reconstituted activator-dependent in vitro transcription system. Our early studies had shown that in vitro, p300 and CREB binding protein (CBP) can catalyze lysine butyrylation by transferring the butyryl group from [^{14}C] butyryl-CoA to core histone proteins (Chen et al., 2007). This activity was also observed in ex vivo transfection experiments (Chen et al., 2007). Moreover, in vitro butyrylation activity was confirmed both on a reconstituted chromatin template and on histone octamers by mass spectrometry analysis. Further analysis of this in vitro activity revealed that p300 efficiently butyrylates the sites of interest, H4K5 and H4K8, in histone octamers as well as in chromatin (Figures 5A and 5B; Data S1).

After confirming that p300 is a histone butyryltransferase, we used a p300- and p53-dependent in vitro transcription system (Figure 5C) to test if histone butyrylation could stimulate transcription. We observed that p300-catalyzed histone butyrylation indeed directly stimulates transcription (Figure 5D). The mutation of lysine residues to arginine either on H3 or H4 tails eliminated acyl-CoA-stimulated transcription, indicating that acetyl/butyryl-CoA activates transcription through p300-catalyzed histone lysine acylation (Figure 5E). This experiment clearly demonstrates that, like acetylation, histone butyrylation can also directly stimulate gene transcriptional activity.

Brdt Binds to Gene TSSs Harboring H4 K5/K8 Acetylation and Butyrylation

Association of the four histone acylation marks with the TSSs of most of the highly active genes in spermatogenic cells raises the possibility that the high transcriptional activity of these genes is mediated by the binding of Brdt to the acylation marks on their TSSs. To test this hypothesis, we compared previously identified Brdt-bound TSSs (Gaucher et al., 2012) with TSSs labeled with the four histone marks. This analysis showed that most of the Brdt-bound gene TSSs also bore high levels of acetylation and butyrylation marks at H4K5 and H4K8 (Figures 6A and 6B). To study if Brdt's first bromodomain (BD1) mediates this interaction, we used spermatogenic cells from mice expressing a mutated form of Brdt lacking BD1 (Gaucher et al., 2012; Shang et al., 2007). The ChIP-seq analysis showed that, in both spermatocytes and round spermatids, the deletion of BD1 considerably weakens Brdt binding to TSSs bearing the four acylation marks (Figure 6A, Brdt ΔBD1). This result indicates that BD1 has a major role in targeting gene TSSs bearing H4K5K8 acetylation/butyrylation.

To further characterize histone acylation at gene TSSs bound by Brdt, we considered the occurrence of the four studied H4 PTMs at TSSs as a function of Brdt-binding. Remarkably, the vast majority of Brdt-associated TSSs bore all four acylation marks (Figure 6B). These observations therefore led us to question the ability of Brdt to bind H4 bearing butyrylation at either K5 or K8 or both.

H4 K5 Butyrylation Inhibits Brdt Binding

To test the ability of Brdt's bromodomains to bind a butyrylated H4 tail, we first carried out an in vitro binding assay using biotinylated H4 tail peptides bearing all four possible combinations of the K5ac, K5bu, K8ac, and K8bu acylation marks. We incubated the peptides with extracts from transfected Brdt-expressing Cos7 cells and performed pull-down experiments. As expected, Brdt efficiently interacts with H4 peptides that are either fully acetylated (i.e., on K5, K8, K12, and K16) or diacetylated on K5 and K8 (Figure S3A, upper). Strikingly, whereas Brdt binding was only slightly affected by the replacement of H4K8ac by H4K8bu, it was completely abolished by the substitution of H4K5ac by H4K5bu. Pull-down experiments using Brdt mutants bearing inactive bromodomains BD1 or BD2 demonstrated that Brdt binding to all the tested peptides depends on the integrity of its first bromodomain (Figure S3A, lower). These data clearly imply that the butyrylation of H4K5 inhibits the binding of Brdt to histone H4.

To validate this result, we performed the same experiment with nuclear extracts from mouse testis. Our results confirmed that butyrylation at H4K5 abolishes the binding of Brdt to H4 tails (Figure 6C, upper). As further confirmation, we repeated the pull-down assay on protein extracts from wild-type mouse testis using either fully acetylated or fully butyrylated immobilized H4 tail peptides and analyzed the bound fractions by mass spectrometry. Brdt was easily identified among the proteins affinity-isolated by the H4ac-containing peptide, but not by the H4bu-containing peptide (Figure S3B; Data S1). The use of extracts from mice testes expressing the Brdt ΔBD1 mutant confirmed that the Brdt-H4 tail interactions described above are primarily mediated by the BD1 domain (Figure 6C, lower right). This result was further corroborated by an experiment with JQ1, a BET bromodomain inhibitor, which abolished the binding of testis-derived Brdt to acetylated H4 and H4K5acK8bu peptides (Figure 6C, lower left). Taken together, these findings establish that H4K5bu, but not H4K8bu, abolishes the interaction between Brdt and the histone H4 tail.

Structural Analysis of the Effect of Butyrylation on the Brdt BD1-H4 Tail Interaction

To understand the molecular basis of the inhibitory effect of H4K5 butyrylation on Brdt binding, we carried out a structural modeling analysis. In the crystal structure of Brdt-BD1 bound to H4K5acK8ac, residue K5ac is intimately recognized by BD1, whereas K8ac makes fewer contacts (Morinière et al., 2009). Modeling shows that replacing the K8 acetyl group with a butyryl group allows the additional atoms to be accommodated without compromising any of the interactions between BD1 and the peptide (Figure 6D). In contrast, replacing the K5 acetyl group by a butyryl group results in a steric clash with residues in the domain's ZA loop, implying that some structural adjustments to the model are required to accommodate the butyryl group. Comparison with the published co-crystal structure of Brd4-BD1

1–3, or 4 of the studied H4 PTMs on autosomes and on the X chromosome in round spermatids (color coded). Of the X-linked genes that escape inactivation, the majority has either no H4K5K8 acetyl/butyryl marks or harbors one to three of these PTMs. A detailed consideration of these PTMs indicates that these TSSs are nearly always depleted in H4K8ac. The list of X-linked genes that are activated in post-meiotic cells was established based on our previous detailed post-meiotic transcriptomic analysis (Boussouar et al., 2014).

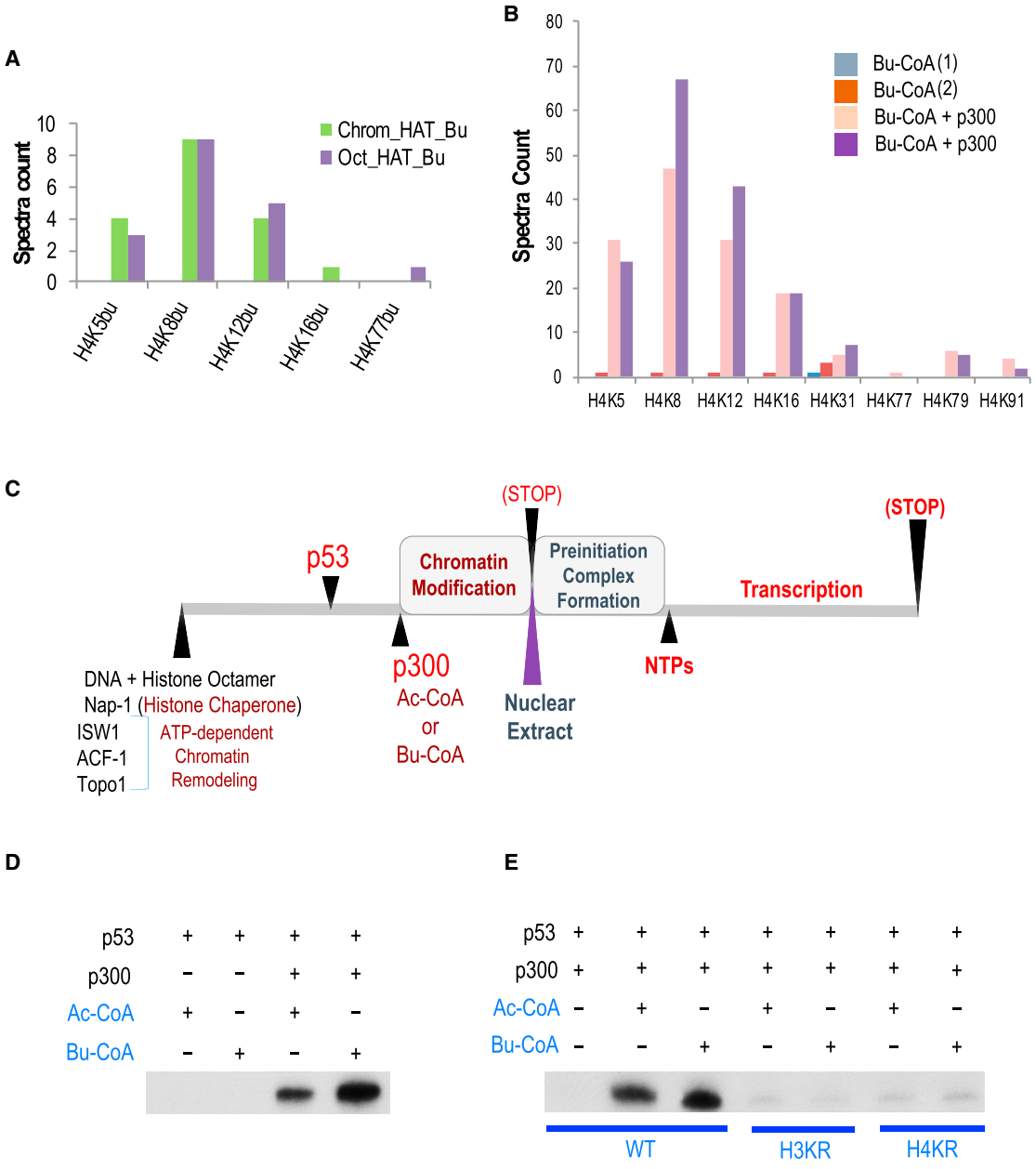


Figure 5. p300 Uses Butyryl-CoA to Butyrylate Histones in Histone Octamers and in Chromatin and to Directly Stimulate Gene Transcription

(A) In vitro reconstituted histone octamers or a chromatin template (used in the assay shown in C) were incubated with butyryl-CoA in the presence of purified p300 and butyrylated histones were detected by MS.

(B) In another assay, histone octamers were incubated with butyryl-CoA in the presence or absence of p300 to measure the degree of non-enzymatic histone butyrylation. The histograms represent the spectra counts of peptides containing the indicated butyrylated lysines in two different experiments (1 and 2: color coded). The annotated MS/MS spectra for histone Kbu peptides are presented in Data S1.

(C) Schematic representation of the p53- and p300-dependent in vitro transcription assay showing the order of the added components (An and Roeder, 2004).

(D) Autoradiography of RNA products obtained by in vitro transcription under the conditions shown.

(E) The same reactions as in (D) were performed with chromatin templates generated by the use of either wild-type histone H3 and H4 or the indicated H3 and H4 K to R mutants.

bound to H3K14bu indicates the type of adjustments required to accommodate H4K5bu (Vollmuth and Geyer, 2010). Aligning the latter structure with that of the Brdt-BD1/H4K5acK8ac complex shows that H3K14bu occupies approximately the same position as K5ac (Figure 6E). However, the bulkier butyryl group results in the displacement of a water molecule within the ligand-binding

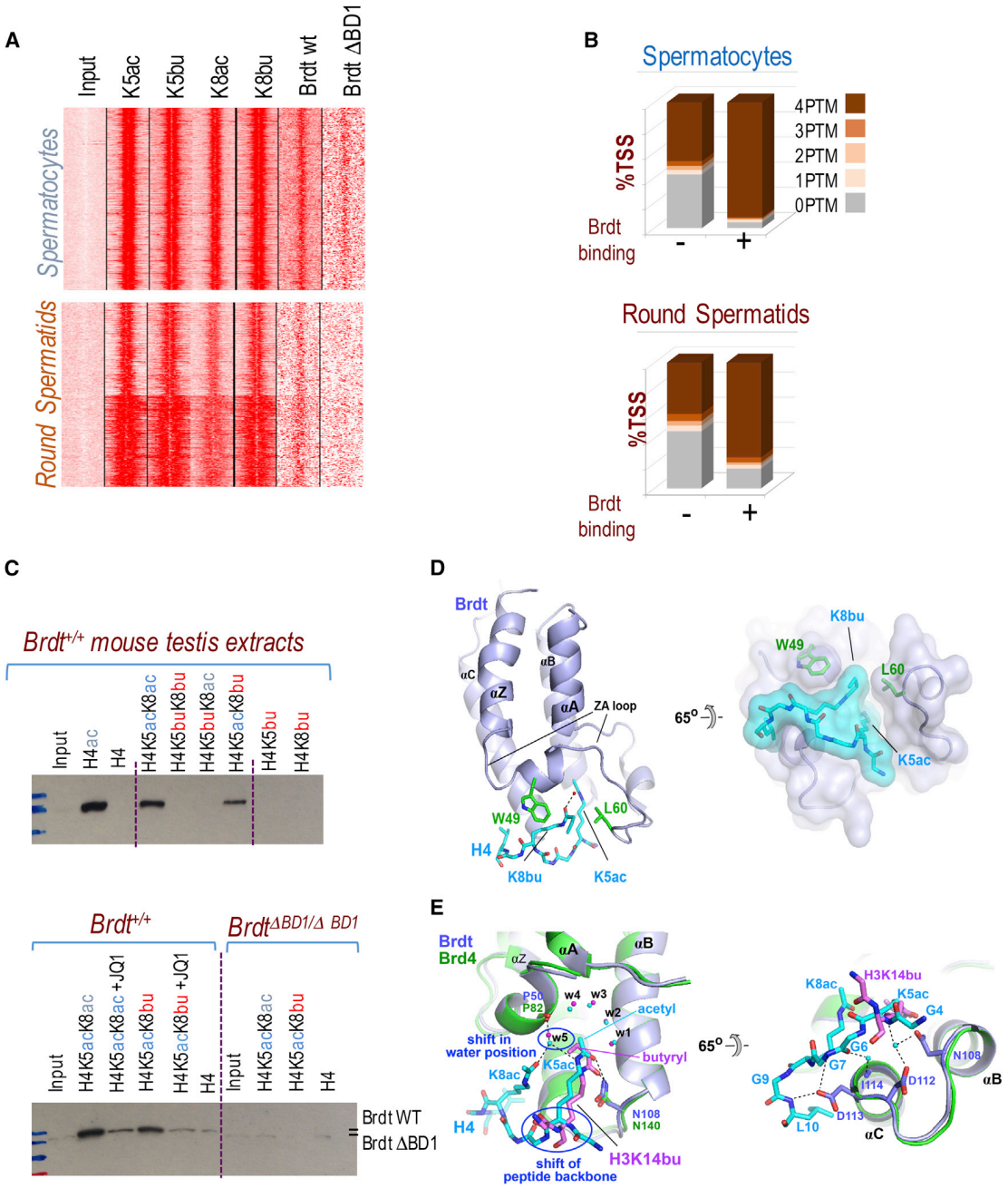


Figure 6. Brdt Is Preferentially Recruited to Gene TSS Regions Enriched in H4K5ac, K5bu, K8ac, and K8bu Marks

(A) ChIP-seq data from chromatin immunoprecipitation of Brdt either from wild-type fractionated spermatogenic cells or the corresponding fractionated cells expressing a truncated Brdt lacking its first bromodomain (Brdt ΔBD1) were obtained and compared with ChIP-seq data from the four indicated histone H4 marks from wild-type spermatogenic cells. Seqminer software illustrates gene TSSs bound by the wild-type Brdt (Brdt wt), Brdt ΔBD1, and the occurrence of the four H4 marks on the same regions. The input corresponds to the sequencing of chromatin fragments before ChIP from wild-type cells. Exactly the same profile was obtained for the input chromatin fragments from Brdt ΔBD1 cells (data not shown).

(B) Gene TSS regions were divided into two categories, either bound or unbound by Brdt, in spermatocytes (upper) or round spermatids (lower). For each category, the proportion of genes whose TSS regions are enriched with none, 1, 2, 3, or 4 of the four histone PTMs are represented with the indicated colors.

(C) Mouse total testis extracts were prepared from wild-type mice (upper and lower left) or mice expressing the truncation mutant Brdt ΔBD1 (lower right) and incubated with the indicated peptides, Brdt was then visualized after pull-down using an anti-Brdt antibody (Gaucher et al., 2012). In some experiments, BET bromodomain inhibitor JQ1 was added to the extract prior to the peptide pull-down step, as indicated. In another experiment Brdt was identified by MS/MS (Figure S3B), and the annotated MS/MS spectra for Brdt peptides are presented in Data S1. The pull-down experiments were also performed on Brdt-expressing transfected cells and the data shown in Figure S3A.

(legend continued on next page)

pocket and causes the K14bu main-chain and side-chain atoms to be shifted relative to those of K5ac. In the context of an H4K5buK8ac peptide, such shifts would be predicted to disrupt several direct and indirect hydrogen bonds between BD1 and the peptide (see figure legend for details). Thus, the modeling approach provides a plausible structural basis for the poor affinity observed for the binding of Brdt to the H4K5buK8ac peptide compared with that to the H4K5acK8ac and H4K5acK8bu peptides.

Dynamic Mixed Labeling of H4K5K8 by Acetylation and Butyrylation at Active Chromatin Sites

Our pull-down and structural modeling data strongly suggest that Brdt is inhibited from binding TSS regions where histone H4 is modified by lysine butyrylation at the K5 position. This conclusion appears contradictory to the ChIP-seq data, where H4K5bu and Brdt were both associated with the same TSS regions. A hypothesis that would reconcile these observations is that acetylation and butyrylation of H4K5 exhibit a rapid turnover. Indeed, acyltransferases associated with highly active gene TSSs might feasibly drain cellular acetyl-CoA as well as butyryl-CoA toward these sites, leading to a mixture of histone H4 acetylation and butyrylation marks. To test this hypothesis, we performed in vitro assays by incubating reconstituted histone octamers with purified p300 and an equimolar mixture of acetyl-CoA and butyryl-CoA. Histones were then analyzed by mass spectrometry to detect acetylated and/or butyrylated peptides. Our results show that p300 can use acyl-CoAs to catalyze acetylation and butyrylation at both H4K5 and H4K8 sites (Figure 7A), as we detected H4 peptides bearing various combinations of acetylated or butyrylated H4K5 and H4K8 (Figure 7A; Data S1). Encouraged by this in vitro result, we then investigated whether histone H4 isolated from spermatogenic cells also contains diverse lysine acylation marks. Mass spectrometry analysis of these samples detected H4 peptides with various combinations of acetylation and butyrylation at H4K5 and H4K8.

In addition, depending on cell type, the stoichiometry of H4K5bu and H4K8bu could be higher than some of the widely studied histone marks such as H3K4me3 (Kulej et al., 2015) (Figures 7B and S4A; Table S2), but lower than those of H4 tail acetylation, such as that of K16, which can be as high as 20% of H4 species (Kulej et al., 2015).

Acetylation is known to have a high turnover rate at gene TSSs (Crump et al., 2011). Therefore, due to continuous acylation by histone acetyltransferase (HATs) such as p300 and the rapid turnover of these PTMs, it is feasible that H4K5 and K8 alternate between acetylated and butyrylated states. This model is supported by the detection of various combinations of H4K5K8 acetylation and butyrylation either in in vitro HAT assays (Figure 7A) or in vivo in spermatogenic cells (Figure 7B). A direct consequence of such alternating histone acetylation/butyrylation would be the dynamic binding of Brdt, which would oscillate between high- and low-affinity states depending on the acylation status of H4K5 (Figure 7C). Such dynamic histone H4 acylation could be facilitated by open nucleosomes on the corresponding gene TSSs by specific histone variants such as H2A.Lap1, a histone H2A variant capable of inducing unstable and open nucleosomes and known to associate with active gene TSSs in spermatogenic cells (Nekrasov et al., 2013; Soboleva et al., 2012). Indeed, analysis of ChIP-seq data for H2A.Lap1 revealed that the four H4 PTMs are particularly enriched on H2A.Lap1-associated TSSs (Figure S4B).

Taken together, the above results support a model whereby alternating competing histone acetylation and butyrylation underlie a dynamic interaction between the histone modifications and the cognate bromodomain.

DISCUSSION

The present study reports our findings regarding the interplay between histone acetylation and butyrylation on the histone H4 tail during sperm cell differentiation. We found that in vitro, p300 uses available acetyl-CoA and butyryl-CoA sources to acylate the H4 tail at all the acceptor lysines in an indiscriminate manner. A proteomic approach also revealed the co-existence of the same combinations of H4 K5K8 acetyl and butyryl marks in different spermatogenic cell types. ChIP-seq analysis on fractionated spermatogenic cells further demonstrated that H4K5/K8 acetyl/butyryl are particularly enriched at the TSSs of the most active genes. Strikingly, however, functional and structural analysis revealed that the acylation state of H4K5 is a critical determinant of Brdt binding affinity, as Brdt binds the acetylated, but not the butyrylated state. In this context, a stable maintenance of differential acetylation and butyrylation could have important functional consequences for the genome reorganization observed during spermatogenesis. We found that, in contrast to earlier stages, in elongating spermatids, H4K5 and H4K8 acetylation and butyrylation become more markedly associated with specific regions of the genome. This could have a direct consequence on the action of Brdt in these cells.

(D and E) Modeling of Brdt-BD1 bound to H4 peptides bearing mixed butyryl and acetyl groups.

(D) Hypothetical model of an H4K5acK8bu peptide bound to Brdt-BD1. The model is based on the crystal structure of the Brdt/H4K5acK8ac complex and was built by substituting the K8 acetyl group with a butyryl group. The substitution allows all contacts between BD1 and the diacetylated peptide to be preserved, while the additional atoms of the butyryl group are accommodated within the space between residues Trp49 and Leu60.

(E) Comparison of the Brdt and Brd4 BD1 bromodomains bound to diacetylated and monobutyrylated peptides, respectively. The two structures were aligned via the pocket-defining BC and ZA loops. Brd4 is shown in green, and its associated water molecules and butyrylated peptide (for which only two residues are present in the crystal structure) are shown in magenta. Brdt is shown in light blue, and its water molecules and peptide ligand are shown in cyan. For clarity, most of the ZA loop is omitted. Brdt-BD1 residues 108 and 112–114 make direct and water-mediated hydrogen bonds to the peptide backbone atoms of H4 residues K5ac, G6, G7, and L10. The shift in peptide backbone position required to accommodate a butyryl group on K5 would disrupt some or all of these interactions. In the Brdt-BD1/H4K5acK8ac complex, hydrogen bonds mediated by water molecule w5 play a key structural role in linking the two acetylated lysine residues to each other and to the backbone carbonyl of BD1 residue Pro50. In the structure of Brd4-BD1/H3K14Bu, w5 is displaced by 1.1 Å toward Pro82, while the amide nitrogen of K14Bu is displaced in the opposite direction by 1.3 Å. Thus, compared to the diacetylated peptide, the presence of the butyryl group disrupts a water-mediated hydrogen bond between BD1 and the peptide.

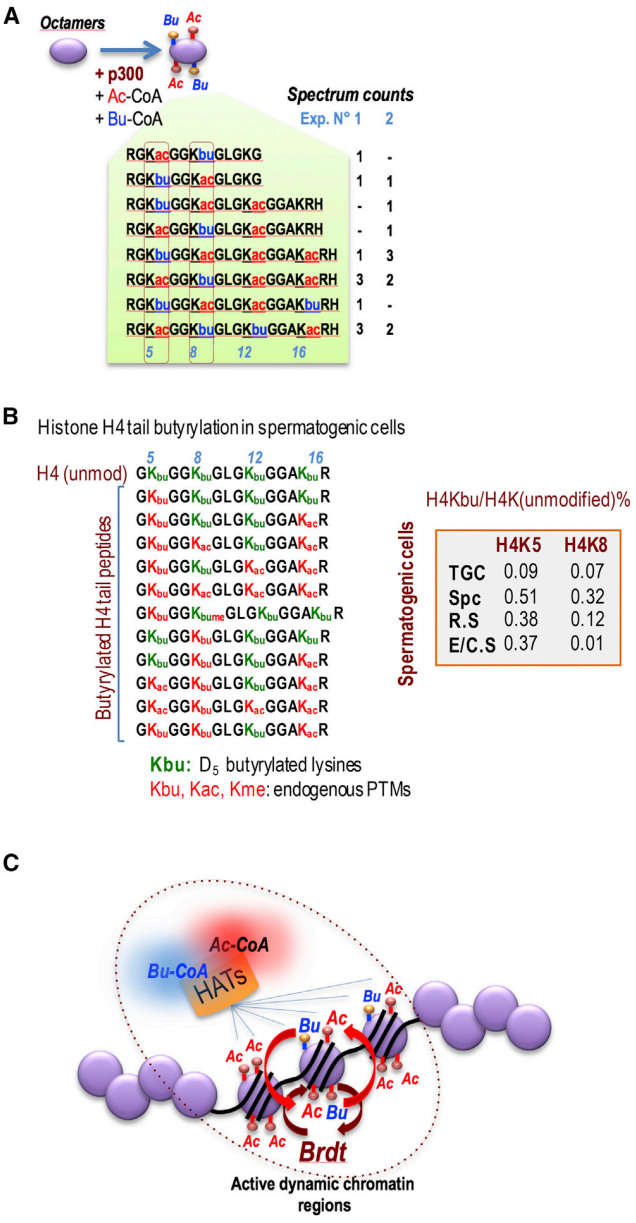


Figure 7. p300-Mediated Acetylation and Butyrylation of H4K5K8 and Functional Implications for Gene Promoter Activity

(A) In vitro reconstituted octamers were incubated with purified p300 in the presence of equal amounts of acetyl-CoA and butyryl-CoA, and the histones were subsequently analyzed by MS. The unique histone H4 peptides bearing multiple lysine modifications, which were specifically identified, are indicated along with their corresponding spectrum counts. The annotated MS/MS spectra for histone Kbu/ac peptides are presented in Data S1.

(B) Stoichiometry of H4K5bu and K8bu in sperm cells. The green letters indicate the in vitro chemical butyrylation, and the red letters indicate the endogenous modifications. The % represents the respective ratios of the peptide bearing H4K5bu and H4K8bu (summed peak areas) over the corresponding unmodified H4 (total germ cells: TGC; spermatocytes: Spc; round spermatids: RS; and elongating/condensing spermatids: E/CS). The values of peak area were used for calculations (Table S2; Figure S4A).

(C) Model of the functional interplay between acylation marks. We propose that at active gene TSSs, p300 (and possibly other HATs) randomly use acetyl- or butyryl-CoA to modify histone tails. HDACs and histone variants, also present at dynamic chromatin regions, accelerate the turnover of these histone marks. Alternate modifications of H4K5 with acetyl and butyryl marks, while maintaining the region competent for transcription, also lead to an oscillation in Brdt binding. The co-localization of H2A.Lap1 and the four H4 PTMs at gene TSSs is shown in Figure S4B. Dynamic Brdt binding could help the pre-initiation complex and RNA PolII to reassemble at TSSs and would favor successive cycles of gene transcription.

Indeed, we previously showed that in elongating spermatids, when histone hyperacetylation and a general transcriptional shut-down occur, BD1 is indispensable for the replacement of acetylated histones by TPs (Gaucher et al., 2012). Additionally, the hyperacetylated histones are known to be directly targeted for degradation by a PA200-containing specialized proteasome (Qian et al., 2013), suggesting that the Brdt-bound histone population enters this pathway. Here, we show that in elongating spermatids, butyrylated histones survive this wave of acetylation-dependent histone removal and degradation. This survival is perfectly consistent with the inability of Brdt to recognize butyrylated H4. These observations suggest that a stable differential labeling of H4 by acetylation and butyrylation may control the timing of histone removal. In this context, it is also possible to speculate that nucleosomes bearing butyrylated H4 could undergo a direct histone-to-Prm replacement due to the inability of Brdt to mediate the exchange of histones by TPs. Indeed, *TP1-TP2* double KO cells can undergo direct histone-to-Prm replacement but, in this case, Prms are unable to tightly compact the genome (for review, see Gaucher et al., 2010)). The prediction is therefore that regions bearing butyrylated histones would evolve to a less compact structure in mature sperms than regions that are marked by acetylated histones before their removal. This regulatory mechanism could be an elegant way to introduce differences in genome compaction by Prms.

Differential histone tail acylation might also play important roles in the control of somatic cell gene expression, differentiation, and genome programming. Indeed, a recent study of p300-mediated histone H3K18 crotonylation revealed that an increase in cellular levels of crotonyl-CoA favors a more efficient de novo gene activation (Sabari et al., 2015). Although the precise mechanism underlying the role of H3K18cr in transcription is unknown, the differential affinity of a bromodomain protein toward crotonyl versus acetyl marks could conceivably account for the observed phenomenon.

Previous reports demonstrated a rapid turnover of histone acetylation on a sub-fraction of nucleosomes (Clayton et al., 2006; Waterborg, 2002), overlapping with active gene TSSs (Crump et al., 2011). However, the functional implications of the short half-life of histone acetylation on active chromatin regions, especially on highly transcribed gene TSSs, have remained elusive. The results described in our present study are consistent with a rapid turnover of both acetyl and butyryl marks on the H4 tail. Given the differential affinity of Brdt BD1 for the acetylated and butyrylated forms of H4K5, a rapid alternation of acylation states at H4K5 would result in a highly dynamic interaction between Brdt and chromatin. Thus, it is tempting to speculate that the generally observed rapid turnover of acetyl marks may be functionally significant because it enables rapid

transitions between alternative states of lysine acylation. Such a mechanism, while maintaining histones permanently modified, would allow for a dynamic association with specific bromodomains, which might be important for sustaining successive cycles of transcription. Thus, either a change in the ratio of cellular acyl-CoAs (for instance caused by metabolic disorders; Pougovkina et al., 2014) or differential activities of histone deacetylase (HDACs) in removing acyl-groups (Rousseaux and Khochbin, 2015) could reprogram gene expression profiles.

In conclusion, we provide here the first demonstration that the interchangeable use of two closely related histone acylation marks at a specific site, H4K5, has important functional consequences by modulating the ability of a transcriptional regulator, Brdt, to recognize chromatin. This finding should improve our understanding of gene transcriptional regulation and its link to cell metabolism.

EXPERIMENTAL PROCEDURES

Antibodies and Other Reagents

Pan anti-Kbu antibody, anti-histone site-specific Kbu, and acetyllysine antibodies used were purchased from PTM Biolabs and anti-histone H3 and anti-histone H4 antibodies were from Abcam. Mouse monoclonal antibodies against H4K5ac and H4K8ac were a generous gift from Dr. H. Kimura. Anti-Brdt is a homemade antibody previously described (Gaucher et al., 2012). Butyryl-CoA and acetyl-CoA were purchased from Sigma-Aldrich. The modified porcine trypsin was purchased from Promega. HPLC-grade acetonitrile, water, and ethanol were purchased from EMD Chemicals. Peptides bearing one or a few acetyl and Kbu residues were custom synthesized and were verified by HPLC and mass spectrometry. JQ1 was synthesized by Charles McKenna and Elena Ferri (University of Southern California) (Emadali et al., 2013) and used as described in this reference.

The antigen recognition capacities of the antibodies used in ChIP-seq and ChIP-qPCR were determined using SPR as described in the Supplemental Information.

Identification of Kbu Sites in Core Histones by Affinity Enrichment and Mass Spectrometry

Acetylation and butyrylation of histones were determined on 200 μg of core histones extracted from the different studied cell types. The detailed procedures are described in the Supplemental Information.

Quantification of H4K5K8 Butyrylation in Spermatogenic Cell Populations

Histone from spermatogenic cells underwent chemical butyrylation with deuterated (D5) butyryl anhydride and processed as described in the Supplemental Information.

Analysis of Mouse Spermatogenic Cells, ChIP, Bioinformatics, and In Vitro Transcription

All the experimental procedures, including immunostaining, cell fractionation, and ChIP-seq were carried out exactly as described previously (Dai et al., 2014; Gaucher et al., 2012; Tan et al., 2011). ChIP-qPCR was performed following our ChIP-seq protocol, but the recovered DNA was amplified using quantitative PCR and specific primers corresponding to regions indicated in Figure S1. Table S1 indicates the sequence of these primers and details of the ChIP experiments are described in Supplemental Information.

Brdt pull-down assays were performed as previously described (Huang et al., 2010; Pivot-Pajot et al., 2003). The in vitro transcription assay was also previously described (An et al., 2002; An and Roeder, 2004; Sabari et al., 2015; Tang et al., 2013). The bioinformatic analyses followed a pipeline similar to the one previously described (Dai et al., 2014; Gaucher et al., 2012; Montellier et al., 2013; Tan et al., 2011) and are detailed in the Supplemental Information.

p300 Expression and Purification for In Vitro Acetyl/Butyryl-Transferase Assays on Histone Octamer

In vitro acetyl/butyryl-transferase assays on histone octamer used recombinant p300 expressed and purified from baculovirus infected in Sf21 insect cells. Protein purification and the details of the HAT assays are described in the Supplemental Information.

ACCESSION NUMBERS

The accession numbers for the H4 K5 and K8 acetylation and butyrylation ChIP-seq data and Brdt ChIP-seq data reported in this paper are GEO: GSE77277 and GSE39910, respectively.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, two tables, and one data file and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2016.03.014.

AUTHOR CONTRIBUTIONS

Y.Z. and S.K. designed and coordinated the whole project. A.G. performed most of the experiments on spermatogenic cells. S.B., J.G., and E.M. were involved in the ChIP setup and S.B. helped with the final experiments. A.-L.V. designed appropriate primers and performed qPCR after ChIP from total germ cells. J.G. and D.P. performed in vitro HAT assays. Z.T. and R.G.R. performed in vitro transcription assays. S.C. helped with pull-down assays. T.B. helped with spermatogenic cell fractionation. C.P. helped with structural analyses. A.D., G.C., and D.P. helped with the bioinformatic analyses under the supervision of S.R. D.P. D.Z. performed the IP/pull-down/MS/MS experiments, PTM quantifications, antibody characterization (SPR), and coordinated the MS analyses. H.H. and O.K. performed the HPLC-MS/MS experiments and data analysis. S.Q. constructed the ChIP-seq library for sequencing. T.H. and Z.C. were involved in antibody production and characterization.

ACKNOWLEDGMENTS

The S.K. laboratory is supported by a grant from the Fondation pour la Recherche Medicale (FRM) “analyse bio-informatique pour la recherche en biologie” program as well as the ANR Episperm3 program (ANR-15-CE12-0005-02 to S. K. and D. P.) and by INCa (2013-082 program to S.K. and C. P.). D. P. laboratory is also supported by the World Wide Cancer Research foundation (16-0280). A.G. has been a recipient of a Marie Curie Initial Training Network funded by the European Commission (FP7-PEOPLE-2011-ITN, PITN-GA-289880) for three years and is now supported by “Fondation ARC.” G.C. is supported by the above-mentioned FRM grant. The Y.Z. laboratory is supported by the NIH (GM105933, DK107868, and GM115961). Y.Z. is a shareholder and a member of the scientific advisory board of PTM BioLabs, Co., Ltd. T.H. and Z.C. are PTM BioLabs employees. Work in the laboratory of R.G.R. was supported by grants from the NIH and the Starr Foundation Tri-Institutional Stem Cell Initiative (2014-034). We gratefully acknowledge the generous gift of mouse monoclonal antibodies against H4K5ac and H4K8ac by Dr. H. Kimura. JQ1 was synthesized and provided to us by Charles McKenna and Elena Ferri. We also acknowledge Miho Shimada for the preparation of histone mutants.

Received: November 4, 2015
Revised: February 5, 2016
Accepted: March 10, 2016
Published: April 21, 2016

REFERENCES

An, W., and Roeder, R.G. (2004). Reconstitution and transcriptional analysis of chromatin in vitro. Methods Enzymol. *377*, 460–474.

An, W., Palhan, V.B., Karymov, M.A., Leuba, S.H., and Roeder, R.G. (2002). Selective requirements for histone H3 and H4 N termini in p300-dependent transcriptional activation from chromatin. Mol. Cell *9*, 811–821.

Boussouar, F., Goudarzi, A., Buchou, T., Shiota, H., Barral, S., Debernardi, A., Guardiola, P., Brindle, P., Martinez, G., Arnoult, C., et al. (2014). A specific CBP/p300-dependent gene expression programme drives the metabolic remodelling in late stages of spermatogenesis. Andrology *2*, 351–359.

Chen, Y., Sprung, R., Tang, Y., Ball, H., Sangras, B., Kim, S.C., Falck, J.R., Peng, J., Gu, W., and Zhao, Y. (2007). Lysine propionylation and butyrylation are novel post-translational modifications in histones. Mol. Cell. Proteomics *6*, 812–819.

Clayton, A.L., Hazzalin, C.A., and Mahadevan, L.C. (2006). Enhanced histone acetylation and transcription: a dynamic perspective. Mol. Cell *23*, 289–296.

Crump, N.T., Hazzalin, C.A., Bowers, E.M., Alani, R.M., Cole, P.A., and Mahadevan, L.C. (2011). Dynamic acetylation of all lysine-4 trimethylated histone H3 is evolutionarily conserved and mediated by p300/CBP. Proc. Natl. Acad. Sci. USA *108*, 7814–7819.

Dai, L., Peng, C., Montellier, E., Lu, Z., Chen, Y., Ishii, H., Debernardi, A., Buchou, T., Rousseaux, S., Jin, F., et al. (2014). Lysine 2-hydroxyisobutyrylation is a widely distributed active histone mark. Nat. Chem. Biol. *10*, 365–370.

Emadali, A., Rousseaux, S., Bruder-Costa, J., Rome, C., Duley, S., Hamaidia, S., Betton, P., Debernardi, A., Leroux, D., Bernay, B., et al. (2013). Identification of a novel BET bromodomain inhibitor-sensitive, gene regulatory circuit that controls Rituximab response and tumour growth in aggressive lymphoid cancers. EMBO Mol. Med. *5*, 1180–1195.

Flynn, E.M., Huang, O.W., Poy, F., Oppikofer, M., Bellon, S.F., Tang, Y., and Cochran, A.G. (2015). A subset of human bromodomains recognizes butyryllysine and crotonyllysine histone peptide modifications. Structure *23*, 1801–1814.

Gaucher, J., Reynoird, N., Montellier, E., Boussouar, F., Rousseaux, S., and Khochbin, S. (2010). From meiosis to postmeiotic events: the secrets of histone disappearance. FEBS J. *277*, 599–604.

Gaucher, J., Boussouar, F., Montellier, E., Curtet, S., Buchou, T., Bertrand, S., Hery, P., Jounier, S., Depaux, A., Vitte, A.L., et al. (2012). Bromodomain-dependent stage-specific male genome programming by Brdt. EMBO J. *31*, 3809–3820.

Goudarzi, A., Shiota, H., Rousseaux, S., and Khochbin, S. (2014). Genome-scale acetylation-dependent histone eviction during spermatogenesis. J. Mol. Biol. *426*, 3342–3349.

Govin, J., Caron, C., Lestrat, C., Rousseaux, S., and Khochbin, S. (2004). The role of histones in chromatin remodelling during mammalian spermiogenesis. Eur. J. Biochem. *271*, 3459–3469.

Hazzouri, M., Pivot-Pajot, C., Faure, A.K., Usson, Y., Pelletier, R., Sèle, B., Khochbin, S., and Rousseaux, S. (2000). Regulated hyperacetylation of core histones during mouse spermatogenesis: involvement of histone deacetylases. Eur. J. Cell Biol. *79*, 950–960.

Huang, R., Holbert, M.A., Tarrant, M.K., Curtet, S., Colquhoun, D.R., Dancy, B.M., Dancy, B.C., Hwang, Y., Tang, Y., Meeth, K., et al. (2010). Site-specific introduction of an acetyl-lysine mimic into peptides and proteins by cysteine alkylation. J. Am. Chem. Soc. *132*, 9986–9987.

Kim, S.C., Sprung, R., Chen, Y., Xu, Y., Ball, H., Pei, J., Cheng, T., Kho, Y., Xiao, H., Xiao, L., et al. (2006). Substrate and functional diversity of lysine acetylation revealed by a proteomics survey. Mol. Cell *23*, 607–618.

Kulej, K., Avgousti, D.C., Weitzman, M.D., and Garcia, B.A. (2015). Characterization of histone post-translational modifications during virus infection using mass spectrometry-based proteomics. Methods *90*, 8–20.

Montellier, E., Rousseaux, S., Zhao, Y., and Khochbin, S. (2012). Histone crotonylation specifically marks the haploid male germ cell gene expression program: post-meiotic male-specific gene expression. BioEssays *34*, 187–193.

Montellier, E., Boussouar, F., Rousseaux, S., Zhang, K., Buchou, T., Fenaille, F., Shiota, H., Debernardi, A., Héry, P., Curtet, S., et al. (2013). Chromatin-to-nucleoprotamine transition is controlled by the histone H2B variant TH2B. Genes Dev. *27*, 1680–1692.

Morinière, J., Rousseaux, S., Steuerwald, U., Soler-López, M., Curtet, S., Vitte, A.L., Govin, J., Gaucher, J., Sadoul, K., Hart, D.J., et al. (2009). Cooperative binding of two acetylation marks on a histone tail by a single bromodomain. Nature *461*, 664–668.

Namekawa, S.H., Park, P.J., Zhang, L.F., Shima, J.E., McCarrey, J.R., Griswold, M.D., and Lee, J.T. (2006). Postmeiotic sex chromatin in the male germline of mice. Curr. Biol. *16*, 660–667.

Nekrasov, M., Soboleva, T.A., Jack, C., and Tremethick, D.J. (2013). Histone variant selectivity at the transcription start site: H2A.Z or H2A.Lap1. Nucleus *4*, 431–438.

Pivot-Pajot, C., Caron, C., Govin, J., Vion, A., Rousseaux, S., and Khochbin, S. (2003). Acetylation-dependent chromatin reorganization by BRDT, a testis-specific bromodomain-containing protein. Mol. Cell. Biol. *23*, 5354–5365.

Pougovkina, O., Te Brinke, H., Wanders, R.J., Houten, S.M., and de Boer, V.C. (2014). Aberrant protein acylation is a common observation in inborn errors of acyl-CoA metabolism. J. Inherit. Metab. Dis. *37*, 709–714.

Qian, M.X., Pang, Y., Liu, C.H., Haratake, K., Du, B.Y., Ji, D.Y., Wang, G.F., Zhu, Q.Q., Song, W., Yu, Y., et al. (2013). Acetylation-mediated proteasomal degradation of core histones during DNA repair and spermatogenesis. Cell *153*, 1012–1024.

Rousseaux, S., and Khochbin, S. (2015). Histone acylation beyond acetylation: Terra incognita in chromatin biology. Cell J. *17*, 1–6.

Sabari, B.R., Tang, Z., Huang, H., Yong-Gonzalez, V., Molina, H., Kong, H.E., Dai, L., Shimada, M., Cross, J.R., Zhao, Y., et al. (2015). Intracellular crotonyl-CoA stimulates transcription through p300-catalyzed histone crotonylation. Mol. Cell *58*, 203–215.

Shang, E., Nickerson, H.D., Wen, D., Wang, X., and Wolgemuth, D.J. (2007). The first bromodomain of Brdt, a testis-specific member of the BET sub-family of double-bromodomain-containing proteins, is essential for male germ cell differentiation. Development *134*, 3507–3515.

Sin, H.S., Barski, A., Zhang, F., Kartashov, A.V., Nussenzweig, A., Chen, J., Andreassen, P.R., and Namekawa, S.H. (2012). RNF8 regulates active epigenetic modifications and escape gene activation from inactive sex chromosomes in post-meiotic spermatids. Genes Dev. *26*, 2737–2748.

Soboleva, T.A., Nekrasov, M., Pahwa, A., Williams, R., Huttley, G.A., and Tremethick, D.J. (2012). A unique H2A histone variant occupies the transcriptional start site of active genes. Nat. Struct. Mol. Biol. *19*, 25–30.

Tan, M., Luo, H., Lee, S., Jin, F., Yang, J.S., Montellier, E., Buchou, T., Cheng, Z., Rousseaux, S., Rajagopal, N., et al. (2011). Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell *146*, 1016–1028.

Tan, M., Peng, C., Anderson, K.A., Chhoy, P., Xie, Z., Dai, L., Park, J., Chen, Y., Huang, H., Zhang, Y., et al. (2014). Lysine glutarylation is a protein posttranslational modification regulated by SIRT5. Cell Metab. *19*, 605–617.

Tang, Z., Chen, W.Y., Shimada, M., Nguyen, U.T., Kim, J., Sun, X.J., Sengoku, T., McGinty, R.K., Fernandez, J.P., Muir, T.W., and Roeder, R.G. (2013). SET1 and p300 act synergistically, through coupled histone modifications, in transcriptional activation by p53. Cell *154*, 297–310.

Vollmuth, F., and Geyer, M. (2010). Interaction of propionylated and butyrylated histone H3 lysine marks with Brd4 bromodomains. Angew. Chem. Int. Ed. Engl. *49*, 6768–6772.

Waterborg, J.H. (2002). Dynamics of histone acetylation in vivo. A function for acetylation turnover? Biochem. Cell Biol. *80*, 363–378.

Xie, Z., Dai, J., Dai, L., Tan, M., Cheng, Z., Wu, Y., Boeke, J.D., and Zhao, Y. (2012). Lysine succinylation and lysine malonylation in histones. Mol. Cell. Proteomics *11*, 100–107.

2.2.2 Rôle du variant d'histone H2AL2 dans le remplacement des histones par les protamines

Des analyses protéomiques menées par l'équipe du docteur Saadi Khochbin ont permis d'identifier par spectrométrie de masse de nouveaux variants d'histones, H2AL1, H2AL2 et H2AL3, provenant de spermatides condensées (GOVIN, ESCOFFIER, ROUSSEAUX et al. 2007). Parmi ces variants, H2AL2 présentait les critères pour être un bon candidat impliqué dans le remplacement des histones. Premièrement, il est exprimé tardivement au cours de la spermiogénèse, dans les spermatides en cours d'élongation et simultanément aux protéines de transition. De plus, il est retrouvé associé à TH2B dans des structures de transition subnucléosomales. Celles-ci ont été mises en évidence dans les spermatides condensées par une digestion poussée à la MNase. Ces petites structures chromatiniennes établiraient un état transitoire de la chromatine en fin de spermiogénèse.

Afin de déterminer le rôle du variant H2AL2 dans le remplacement des histones par les protamines, un modèle de souris invalidant son gène a été généré. Les résultats des approches expérimentales comparant ce modèle à un modèle *wild-type* ont fait l'objet d'une publication (ARTICLE 3, BARRAL, MOROZUMI, TANAKA et al. 2017). Les analyses de données MNase-seq que j'ai mené ont permises d'établir précisément la taille (73 bp) des fragments ADN protégés de la MNase par les petites structures dans les spermatides condensées (FIGURE 5B de l'ARTICLE 3). En l'absence de H2AL2, ces petites structures sont quand même observées mais les fragments possèdent une taille inférieure (60 bp). De même, les nucléosomes restant en spermatides condensées protègent davantage l'ADN (157 bp). Le processus de transformation d'un nucléosome en petite structure a pu être observée précisément en se positionnant au niveau des nucléosomes stables en spermatides rondes (FIGURE 5C de l'ARTICLE 3). Cette approche a révélé que la petite structure avait un positionnement conservé entre cellules pour chaque nucléosome et que ce positionnement est déterminé plus par la séquence ADN que par la position originelle du nucléosome. Le croisement de ces résultats avec ceux d'approches biochimiques et protéomiques a permis de proposer un nouveau modèle moléculaire de remplacement des histones par les protamines incluant H2AL2 avant l'entrée en action des protéines de transition.

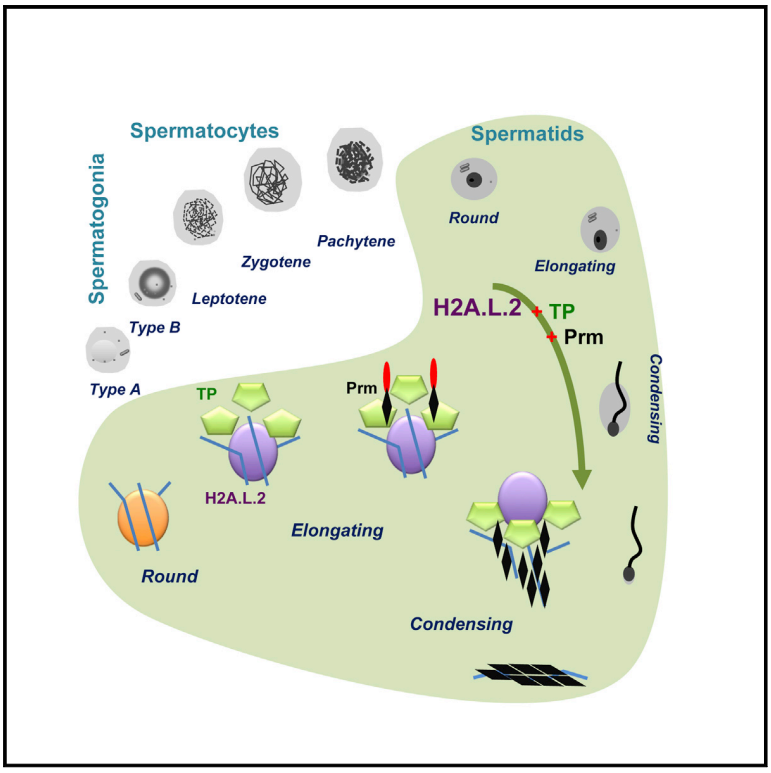
Un projet annexe visait à déterminer l'impact de l'incorporation de H2AL2 sur les régions péricentromériques. Suite à une expérience d'hybridation *in situ* en fluorescence de sondes ARN (RNA FISH) mettant en évidence la transcription des séquences répétées péricentromériques lors de l'incorporation de H2AL2 dans ces régions, l'hypothèse d'un impact de H2AL2 sur cette transcription a été formulée. Afin de la vérifier, des RNA-seq orientés en triplicats à différents stades de la spermatogénèse ont été réalisés et analysés. Les niveaux transcriptomiques observés n'ont révélé aucune différence en présence ou absence du variant d'histone étudié (FIGURE 2.1).

Article

Molecular Cell

Histone Variant H2A.L.2 Guides Transition Protein-Dependent Protamine Assembly in Male Germ Cells

Graphical Abstract



Authors

Sophie Barral, Yuichi Morozumi, Hiroki Tanaka, ..., Sophie Rousseaux, Hitoshi Kurumizaka, Saadi Khochbin

Correspondence

saadi.khochbin@univ-grenoble-alpes.fr

In Brief

Histone-to-protamine transition is essential for procreation, yet its molecular basis has remained obscure. Barral et al. show that a histone variant, H2A.L.2, directs the transformation of nucleosomes by allowing the loading onto chromatin of non-histone transition proteins, which in turn control the final genome organization and compaction by protamines.

Highlights

- H2A.L.2 and transition proteins (TPs) are co-expressed in post-meiotic germ cells
- Lack of H2A.L.2 leads to defective spermatozoa genome compaction and male infertility
- H2A.L.2 assembly opens nucleosomes, releases flexible DNA ends, and allows TP loading
- TP loading drives protamine processing and organized compaction of the sperm genome

Barral et al., 2017, Molecular Cell *66*, 89–101
April 6, 2017
http://dx.doi.org/10.1016/j.molcel.2017.02.025

Molecular Cell
Article

Histone Variant H2A.L.2 Guides Transition Protein-Dependent Protamine Assembly in Male Germ Cells

Sophie Barral,[1,7] Yuichi Morozumi,[1,2,7] Hiroki Tanaka,[2] Emilie Montellier,[1] Jérôme Govin,[3] Maud de Dieuleveult,[4] Guillaume Charbonnier,[5] Yohann Couté,[3] Denis Puthier,[5] Thierry Buchou,[1] Fayçal Boussouar,[1] Takashi Urahama,[2] François Fenaille,[6] Sandrine Curtet,[1] Patrick Héry,[4] Nicolas Fernandez-Nunez,[5] Hitoshi Shiota,[1] Matthieu Gérard,[4] Sophie Rousseaux,[1] Hitoshi Kurumizaka,[2] and Saadi Khochbin[1,8,*]

[1]CNRS UMR 5309, Inserm U1209, Université Grenoble Alpes, Institute for Advanced Biosciences, Grenoble 38700, France
[2]Laboratory of Structural Biology, Graduate School of Advanced Science and Engineering, Research Institute for Science and Engineering, Institute for Medical-oriented Structural Biology, Waseda University, 2-2 Wakamatsu-cho, Shinjuku-ku, Tokyo 162-8480, Japan
[3]Université Grenoble Alpes, Inserm U1038, CEA, BIG-BGE, Grenoble 38000, France
[4]Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, Gif-sur-Yvette 91198, France
[5]TGML, platform IbiSA, Aix Marseille Univ, Inserm U1090, TAGC, Marseille 13288, France
[6]Laboratoire d'Etude du Métabolisme des Médicaments, DSV/iBiTec-S/SPI, CEA Saclay, Gif-sur-Yvette 91191 Cedex, France
[7]Co-first author
[8]Lead Contact
*Correspondence: saadi.khochbin@univ-grenoble-alpes.fr
http://dx.doi.org/10.1016/j.molcel.2017.02.025

SUMMARY

Histone replacement by transition proteins (TPs) and protamines (Prms) constitutes an essential step for the successful production of functional male gametes, yet nothing is known on the underlying functional interplay between histones, TPs, and Prms. Here, by studying spermatogenesis in the absence of a spermatid-specific histone variant, H2A.L.2, we discover a fundamental mechanism involved in the transformation of nucleosomes into nucleoprotamines. H2A.L.2 is synthesized at the same time as TPs and enables their loading onto the nucleosomes. TPs do not displace histones but rather drive the recruitment and processing of Prms, which are themselves responsible for histone eviction. Altogether, the incorporation of H2A.L.2 initiates and orchestrates a series of successive transitional states that ultimately shift to the fully compacted genome of the mature spermatozoa. Hence, the current view of histone-to-nucleoprotamine transition should be revisited and include an additional step with H2A.L.2 assembly prior to the action of TPs and Prms.

INTRODUCTION

Spermatogenesis generates the only cells—spermatozoa—capable of leaving the parent organism and transmitting the male genome to the next generation. Mature spermatozoa are produced through a continuous differentiation of progenitor cells, spermatogonia, that undergo meiotic divisions to generate haploid cells (spermatids). In mammals, the universal mode of DNA packaging in eukaryotes undergoes a profound transformation when spermatid nuclei start the elongation process (elongating spermatids) (De Vries et al., 2012; Gaucher et al., 2010). Genome-wide histone hyperacetylation seems to signal the initiation of the process (Goudarzi et al., 2014), which is characterized by the replacement of histones by non-histone basic proteins, transition proteins (TPs), and protamines (Prms), in subsequent stages in condensing spermatids (Gaucher et al., 2010).

The current vision of these transformations proposes that histones are replaced by TPs (Meistrich et al., 2003), which in turn are removed and replaced by Prms (Lewis et al., 2003). However, TPs and Prms seem to have the same time frame of appearance in elongating/condensing spermatids when histones are still detectable (Meistrich et al., 2003; Zhao et al., 2004b) and their precise order of action still remains obscure.

Seminal works showed that full production of Prms is required for tight genome packaging (Cho et al., 2001, 2003). In addition, the generation of mouse TP1 and TP2 double knockout (KO) (Shirley et al., 2004; Zhao et al., 2004a, 2004b) shows that, although these mice are infertile, Prms alone are capable of histone eviction. These data therefore question the functional utility of TPs in the process of histone removal.

These studies also demonstrate that TPs are somehow necessary for the processing of Prm2, which is synthesized as a preprotein, and for the ultimate male genome compaction (Shirley et al., 2004; Zhao et al., 2004a), but these observations do not explain the molecular basis of the functional relationship between TP assembly, Prm incorporation, and histone replacement.

In the past years, our laboratory has developed several lines of research to tackle the molecular basis of histone eviction and the establishment of the final Prm-dependent genome packaging (Govin et al., 2012; Rousseaux et al., 2011). One of these

strategies, which is aimed at the identification of histones present in chromatin during the process of histone-to-Prm replacement, led to the discovery of a group of novel H2A- and H2B-type histone variants (Govin et al., 2007). One of the identified H2A variants, H2A.L.2, particularly attracted our attention since it is expressed during late spermatogenesis, precisely at the time of histone replacement (Govin et al., 2007). Further investigations showed that, along with the H2B histone variant, TH2B, H2A.L.2 is essentially present in micrococcal nuclease (MNase)-sensitive sub-nucleosomal particles in elongating/condensing spermatids (Govin et al., 2007). TH2B was first identified many years ago in male germ cells, but its function had remained unexplored until recently (Montellier et al., 2013; Shinagawa et al., 2014). Our recent work demonstrates that TH2B is expressed as soon as spermatogonia commit into meiotic divisions and that it gradually replaces H2B at a genome-wide level (Montellier et al., 2013). Interestingly, we also observed that H2A.L.2 preferentially dimerizes with TH2B (Govin et al., 2007). Therefore, at the time of its expression, H2A.L.2 should compete with other H2As for dimerization with TH2B and hence be massively incorporated into the chromatin of elongating spermatids. Hence, H2A.L.2 seemed to be an excellent candidate whose functional study could bring valuable information on the molecular basis of histone-to-Prm transition.

Following this reasoning, we generated mice lacking H2A.L.2. Interestingly, the phenotypes observed in the absence of H2A.L.2 appeared to be very close to those previously reported for TP1/TP2 double-knockout mice (Shirley et al., 2004; Zhao et al., 2004a, 2004b). This observation led us to hypothesize that, in the absence of H2A.L.2, TPs could be non-functional. Focusing our investigations on TPs through a series of biochemical, proteomic, and genomic approaches using wild-type (WT) and H2A.L.2-KO condensing spermatids, we explored, in detail, all the critical steps of the transformation of a nucleosomal structure into nucleoprotamine. As a result, we demonstrate that the current vision of histone-to-Prm transition needs to be completed and updated to include a critical step, which is the incorporation of this H2A histone variant prior to the action of transition proteins.

RESULTS

H2A.L.2 Is Co-expressed with Transition Proteins

Previously, we identified H2A.L.2 as a new histone variant specifically expressed in post-meiotic elongating spermatids, when TPs also start to accumulate (Govin et al., 2007). Here, we designed several experiments to precisely define the timing of H2A.L.2 accumulation compared to TPs. Immunoblots on extracts from developing mouse testes show that H2A.L.2 is first detected at 29 days post-partum (dpp), at the same time as TPs (Figure 1A). Co-detection of H2A.L.2 and TP2 by immunofluorescence on cells from 28 and 29 dpp mouse testes, respectively, before and after H2A.L.2 protein accumulation, shows not only that H2A.L.2 is expressed precisely at the same time as TP2, but also that the two proteins are in fact co-expressed in the same cells (Figure 1A). Detection of H2A.L.2 and TP1 and TP2 on testis sections by immunohistochemistry (IH) confirmed the above findings, highlighting that TPs and H2A.L.2 expression is coordinated and occurs in the same cells at the same time (Figure 1B).

H2A.L.2 Is Required for Final Sperm Genome Compaction and Male Fertility

We then generated H2A.L.2-KO mice following the same strategy as previously used in our TH2B mouse model (Montellier et al., 2013). Briefly, the insertion of a Neomycin (*Neo*) cassette in the 3′ part of the gene (Figure S1) leads to a complete lack of H2A.L.2 protein production, since no trace of the protein was detected using either immunodetection or UHPLC-MS approaches in whole-testis extracts or in mature spermatozoa (Figure 2A). Despite normal testis histology (Figure S2A), normal production of TPs and Prms (Figure 2B; Figure S3A), and normal spermatozoa counts (Figure 2B), H2A.L.2-KO male were completely sterile, unable to father any pups (Figure 2C, upper panel). Furthermore, the analysis of oocytes after in vivo or in vitro fertilization using H2A.L.2-KO male mice spermatozoa showed an absence of fertilized eggs (Figure 2C, middle and lower panels, respectively).

A high-resolution analysis of chromatin texture in mature spermatozoa from WT and H2A.L.2-KO mice by electron microscopy (EM) showed that, whereas the chromatin of WT sperm is relatively uniformly dense with only a few translucent regions, many translucent areas are visible in the chromatin of spermatozoa lacking H2A.L.2, which are indicative of a defective global genome compaction (Figure 2D, upper panels).

Because of the observed spatiotemporal association between H2A.L.2 and TPs expression, we paid specific attention to the published phenotypes of TP1/TP2 double-KO mice, which presented similar defective genome compaction in mature spermatozoa as well as marked defects in the processing of pre-Prm2 (Shirley et al., 2004; Zhao et al., 2004a). To further examine the phenotypical similarities between the absence of H2A.L.2 and TPs, we also monitored the processing of pre-Prm2 in the absence of H2A.L.2 compared to WT spermatozoa. Figure 2D (lower panels) shows that, in the absence of H2A.L.2, a significant proportion of pre-Prm2 remains unprocessed.

Finally, since we previously showed that a fraction of H2A.L.2 remains in mature spermatozoa and that H2A.L.2 preferentially dimerizes with TH2B (Govin et al., 2007), we also visualized TH2B and H2A.L.2 in epididymal spermatozoa from both genotypes. Figure S2B shows that the absence of H2A.L.2 also leads to defective TH2B retention but does not affect the amounts of retained H3 and H4.

Altogether, these investigations led us to hypothesize that H2A.L.2 and TPs are functioning in interdependent molecular mechanisms, whose respective dysfunctions would lead to the same global defective genome compaction, and prompted us to specifically consider the activity of TPs and Prms in the absence of H2A.L.2.

H2A.L.2 Is Required for Transition Proteins Loading onto the Nucleosome and for Efficient Prm Assembly

In the absence of H2A.L.2, TP1 and TP2 are normally expressed in the expected cell types (Figure 3A, upper panels; Figure S3A). However, a high-resolution analysis of TP1 and TP2 distribution in the nucleus of condensing spermatids demonstrates that, while these proteins are associated with chromatin in WT cells,

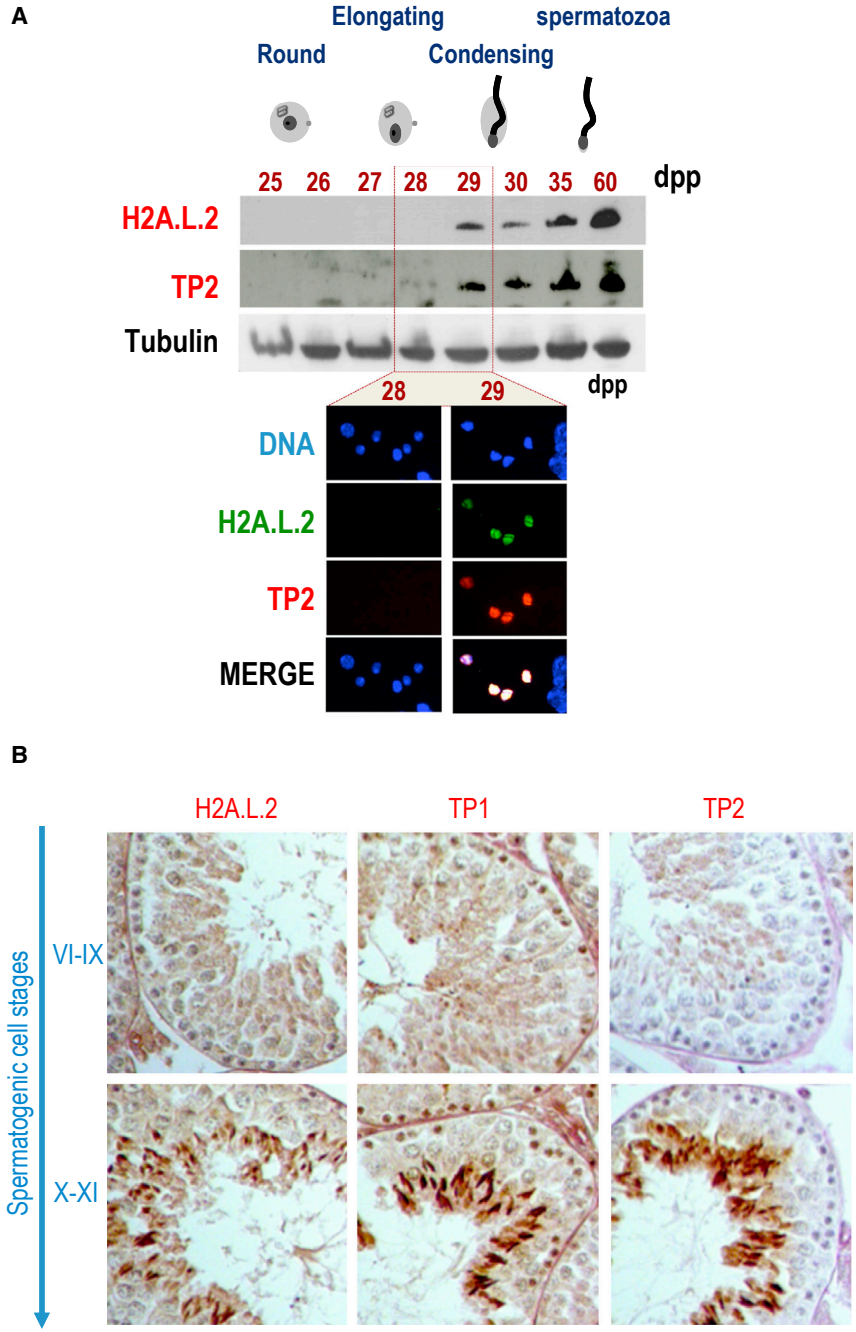


Figure 1. H2A.L.2 and Transition Proteins Are Co-expressed

(A) Total extracts from testes taken at the indicated days after birth (dpp) were successively probed with anti-H2A.L.2, anti-TP2, and anti-tubulin antibodies as indicated (upper panels). The lower panels show the co-detection of H2A.L.2 (rabbit antibody) and TP2 (goat antibody) in testis squash preparations, using testes taken at 28 and 29 dpp as indicated. DNA was stained by DAPI, and the merge panels show the superposition of DNA, TP2, and H2A.L.2 detections.

(B) Sections of paraffin-embedded testes were used to detect cells expressing H2A.L.2, TP1, and TP2. The spermatogenic cell stages for each section are indicated.

they both show defective localization in the absence of H2A.L.2 (Figure 3A, lower panels). Simultaneous detection of TP1 and TP2 and of the acrosome by lectin in confocal microscopy confirmed the aberrant accumulation of TP1 and TP2 in the peri-nuclear regions that do not overlap with the acrosome (Figure S3B).

These results clearly demonstrate the inability of TPs to normally associate with chromatin in the absence of H2A.L.2.

To further characterize TPs and their associated proteins in the presence or absence of H2A.L.2, we prepared a 0.5 M salt extract from purified condensing spermatids, which solubilizes TPs and the associated salt-resistant interactants but does not extract histones from the remaining chromatin (Figure S3C). These extracts were used to immunoprecipitate TP2. Figure 3B shows that TP2 is efficiently and specifically immunoprecipitated in both situations, with no trace of TP2 or its associated proteins, when an irrelevant antibody was used (Figure 3B, Ctl lanes). Co-immunoprecipitation of H3 and TH2B showed that significant amounts of histones become associated with TP2 independently of H2A.L.2. In the absence of H2A.L.2, pre-Prm2 is detected in association with TP2, whereas in the presence of H2A.L.2, only trace amounts of pre-Prm2 are observed (Figure 3B). The presence of pre-Prm2 in soluble extracts from H2A.L.2-KO spermatids and its enrichment in the TP2 immunoprecipitates imply that these cells face a Prm2 deficiency for full genome packaging.

The simplest explanation for the trapping of pre-Prm2 by TP2 would be the ability of the two proteins to interact. To test this hypothesis, we ectopically expressed TP2 in COS cells along with full-length pre-Prm2, its N-terminal half (normally removed during assembly), or its C-terminal half (the assembled part of the protein). 0.5 M salt extracts were prepared, and TP2 was immunoprecipitated. This experiment demonstrated that TP2 specifically interacts with the full-length Prm2 and with its C-terminal half, but not with its N-terminal half, which is normally proteolitically removed (Figure S4).

Therefore, when TPs are not loaded on chromatin, the accumulating TP2 could create a "sink" that traps a fraction of the incoming pre-Prm2, preventing full Prm2 assembly.

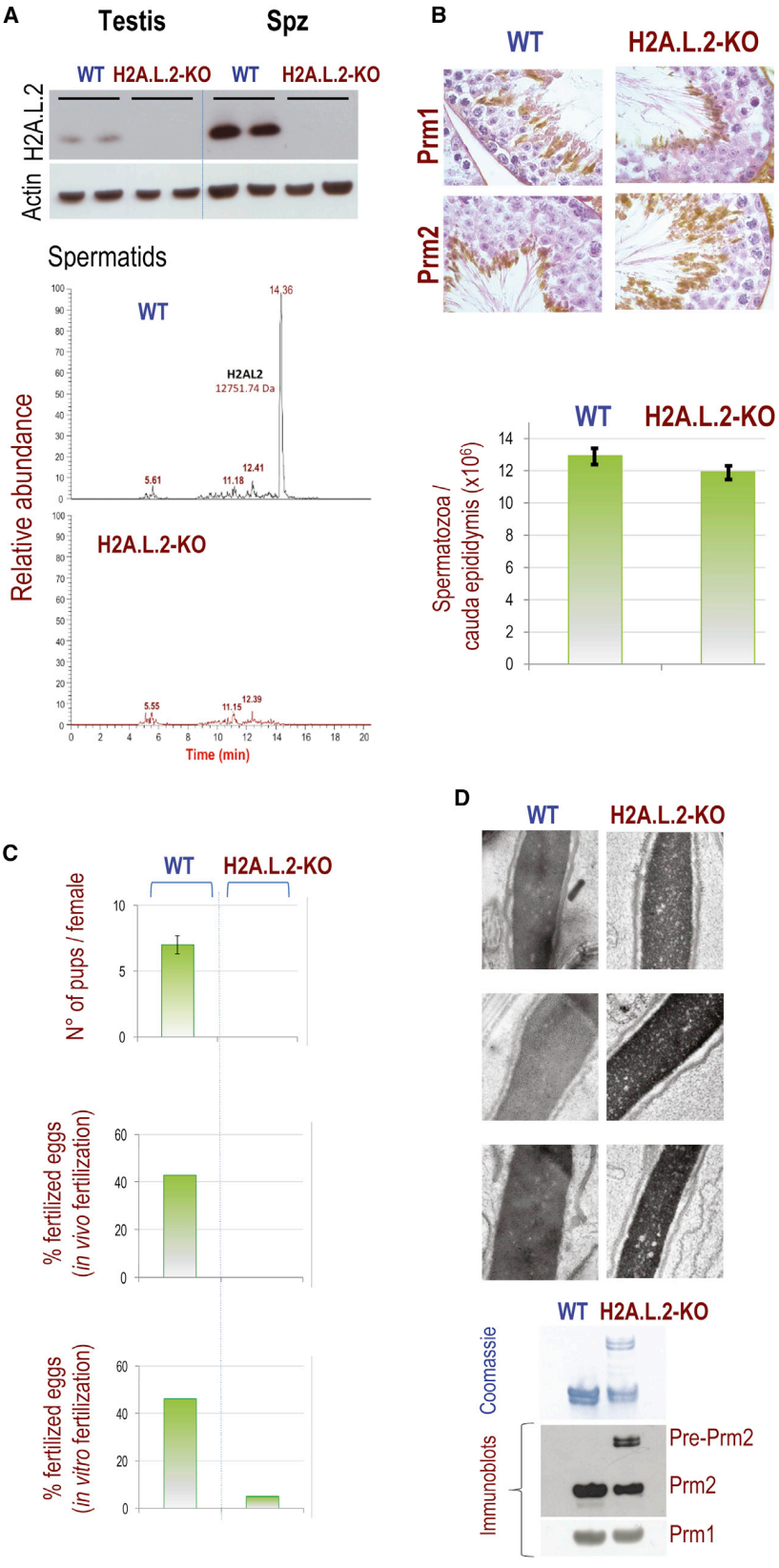


Figure 2. Lack of H2A.L.2 Is Associated with Male Infertility and Results in Deficient Final Genome Compaction and Abnormal Pre-protamine 2 Processing

(A) Whole testis or epididymal spermatozoa extracts from WT and H2A.L.2-KO mice were analyzed by immunoblotting using anti-H2A.L.2 and anti-actin antibodies (upper panels). The lower panels show the UHPLC-MS profiling of histones extracted from the WT or the H2A.L.2-KO condensing spermatids.

(B) Sections of paraffin-embedded testes from WT and H2A.L.2-KO mice were probed with anti-Prm1 and anti-Prm2 antibodies as indicated. The histograms indicate the number of epididymal spermatozoa counted from seven WT and seven H2A.L.2-KO males.

(C) The number of pups per female is indicated after crossing WT female mice with WT (n = 5) or H2A.L.2-KO (n = 5) males as indicated (upper panel). Oocytes from six WT females mated with WT (n = 3) or H2A.L.2-KO (n = 3) males were examined for fertilization, and the number of fertilized oocytes is presented as the percentage of analyzed oocytes (middle panels). Oocytes from the same number of WT females as above were obtained and used for in vitro fertilization (IVF) with spermatozoa from WT (n = 3) or H2A.L.2-KO (n = 3) males. The fertilization success rate is indicated as a percentage of oocytes giving rise to stage 2 embryos (lower panels).

(D) Genome compaction in WT and H2A.L.2-KO spermatozoa was monitored by electron microscopy (upper panels). Acid extracts from WT and H2A.L.2-KO spermatozoa were analyzed after Triton-Acid-Urea (TAU) gel electrophoresis. The gel was stained by Coomassie (indicated) or transferred onto a membrane for immunoblots with the indicated antibodies (lower panels). For (B) and (C), bars show the standard errors.

See Figure S1 for the strategy to generate mice with inactivated *h2a.l.2* gene. The resulting genetically modified mice are called in this manuscript H2A.L.2-KO.

See also Figures S2A and S3A for testis histology and the pattern of TP1 and TP2 expression, respectively.

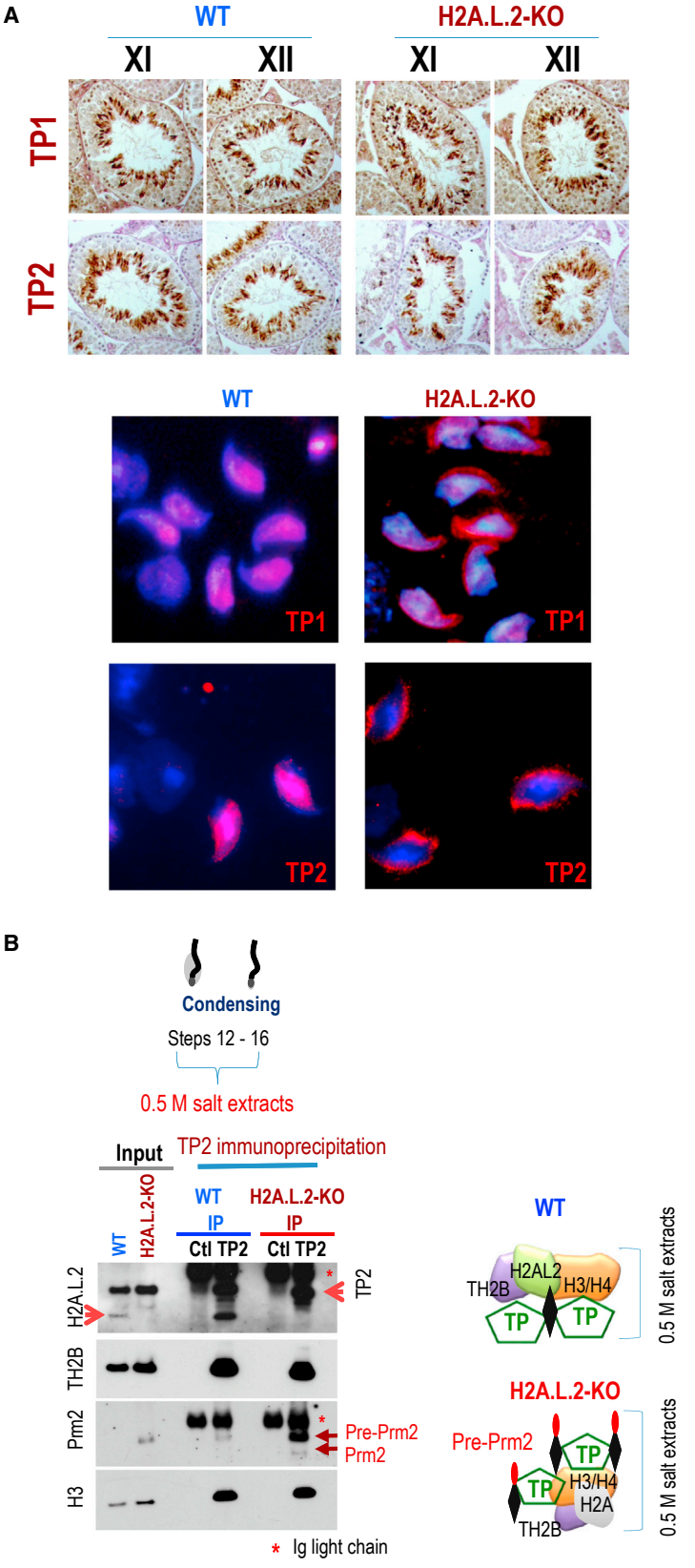


Figure 3. Characterization of Transition Proteins' Activities in the Presence or Absence of H2A.L.2

(A) TP1 and TP2 were immunodetected on testis sections from WT and H2A.L.2-KO spermatogenic cells as indicated. The spermatogenic cell stages are indicated (upper panels). The lower panels show merged images after DNA staining with DAPI (blue) and TP1 and TP2 immunodetection (red) in WT and H2A.L.2-KO condensing spermatids.

(B) Steps 12–16 spermatids were purified from WT and H2A.L.2-KO testes, and the corresponding 0.5 M salt nuclear extracts prepared. The extracts were then used to immunoprecipitate TP2 (IP TP2). An anti-goat Ig was used as a control antibody (IP Ctl). The immunoprecipitated materials were then probed with the indicated antibodies. Please note that the upper panel shows the co-detection of both TP2 and H2A.L.2 as indicated. The arrows show the position of pre-Prm2 and Prm2 and "*" indicates Ig light chains.

The schemes on the right summarize the results of these immunoprecipitations by indicating the proteins of interest found in association with TP2 in the 0.5 M salt extracts. See also Figure S3B for the co-detection of TP1 and TP2 with the acrosome, Figure S3C for the characterization of 0.5 M salt extracts, and Figure S4 for the validation of TP2-Prm2 interaction.

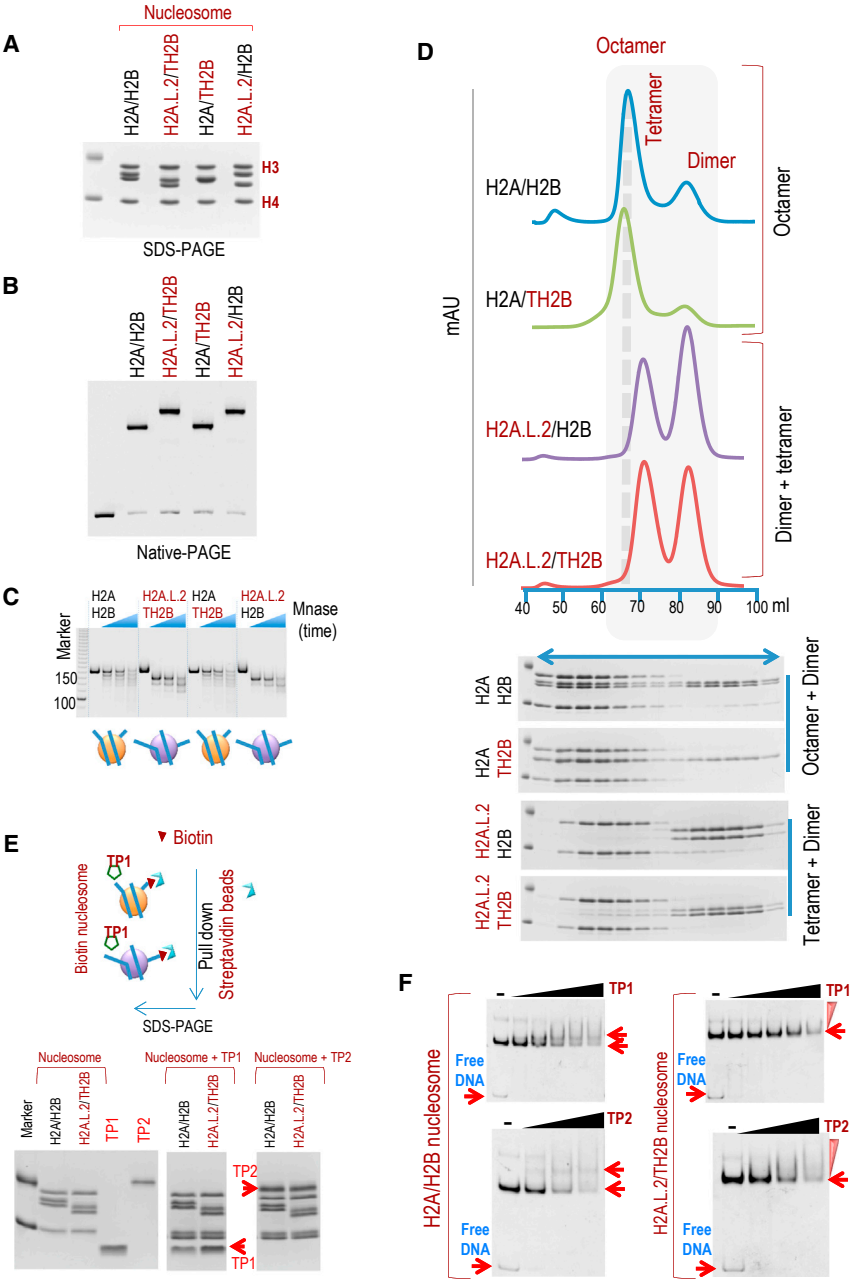


Figure 4. H2A.L.2 Is Required for the Invasion of Nucleosomes by Transition Proteins

(A and B) The indicated purified nucleosomes were analyzed by 18% SDS polyacrylamide gel electrophoresis (PAGE) (A) and 6% native PAGE (B).

(C) The reconstituted nucleosomes shown in (B) were digested with MNase during 0, 5, 10, and 20 min, and the DNA was purified and visualized after migration.

(D) The histone combinations shown in (A) were used to generate histone octamers, which were then analyzed by gel filtration. The gel filtration profiles of the indicated histone combinations used in the octamer formation assays are indicated. Gel filtration fractions (boxed) were analyzed on SDS-PAGE to visualize the presence of histones (lower panels).

(E) H2A-H2B or H2A.L.2-TH2B containing nucleosomes were reconstituted on biotinylated DNA ("Nucleosome") and incubated with equimolar amounts of TP1 and TP2 (indicated). The beads were washed and the bound proteins were recovered in SDS-PAGE loading buffer. The bound proteins ("Nucleosome + TP1" and "Nucleosome + TP2") were analyzed on an 18% SDS-PAGE and visualized by Coomassie brilliant blue staining.

(F) Nucleosomes shown in (B) were incubated with increasing concentrations of purified recombinant TP1 and TP2 and their association with nucleosomes visualized after migration on native PAGE. Figure S5A shows the quantifications of TP1 and TP2 binding corresponding to the experiment shown in (E) from three independent assays.

H2A.L.2 Is Required for the Invasion of Nucleosomes by TPs

To better understand the role of H2A.L.2 in mediating the action of TPs on nucleosomes, we expressed and purified core histones, including H2A, H2B, TH2B, and H2A.L.2, and then used them for in vitro nucleosome reconstitution (Figures 4A and 4B). These nucleosomes were then digested with MNase. Figure 4C shows that the presence of H2A.L.2 is sufficient to generate nucleosomal DNA fragments that are rapidly digested by MNase. These data are perfectly consistent with the data previously published by various groups, where the H2A.B and H2A.L variants, which lack the characteristic H2A acidic patch and present a docking domain shorter than H2A, protect only around 120–130 bp of nucleosomal DNA, since they release nucleosomal DNA at the entry and exit of nucleosomes (Arimura et al., 2013; Bao et al., 2004; Doyen et al., 2006; Nekrasov et al., 2012; Soboleva et al., 2011; Syed et al., 2009).

Our in vitro studies also revealed another characteristic of H2A.L.2-containing nucleosomes, explaining how H2A.L.2 could contribute to nucleosome disassembly. Indeed, in the absence of DNA, we found that it was impossible to reconstitute histone octamers containing H2A.L.2, while stable octamers could be formed with H2A (Figure 4D).

These data suggest that all the events that weaken histone-DNA interactions, such as histone post-translational modifications (PTMs) at specific positions (Tropberger and Schneider, 2013), and competition with strong DNA-binding proteins, such as Prms, have the potential to induce an immediate dissociation of TH2B-H2A.L.2 from H3-H4.

We took advantage of this in vitro system to monitor the functional interface between H2A or H2A.L.2-containing nucleosomes and TP1 and TP2.

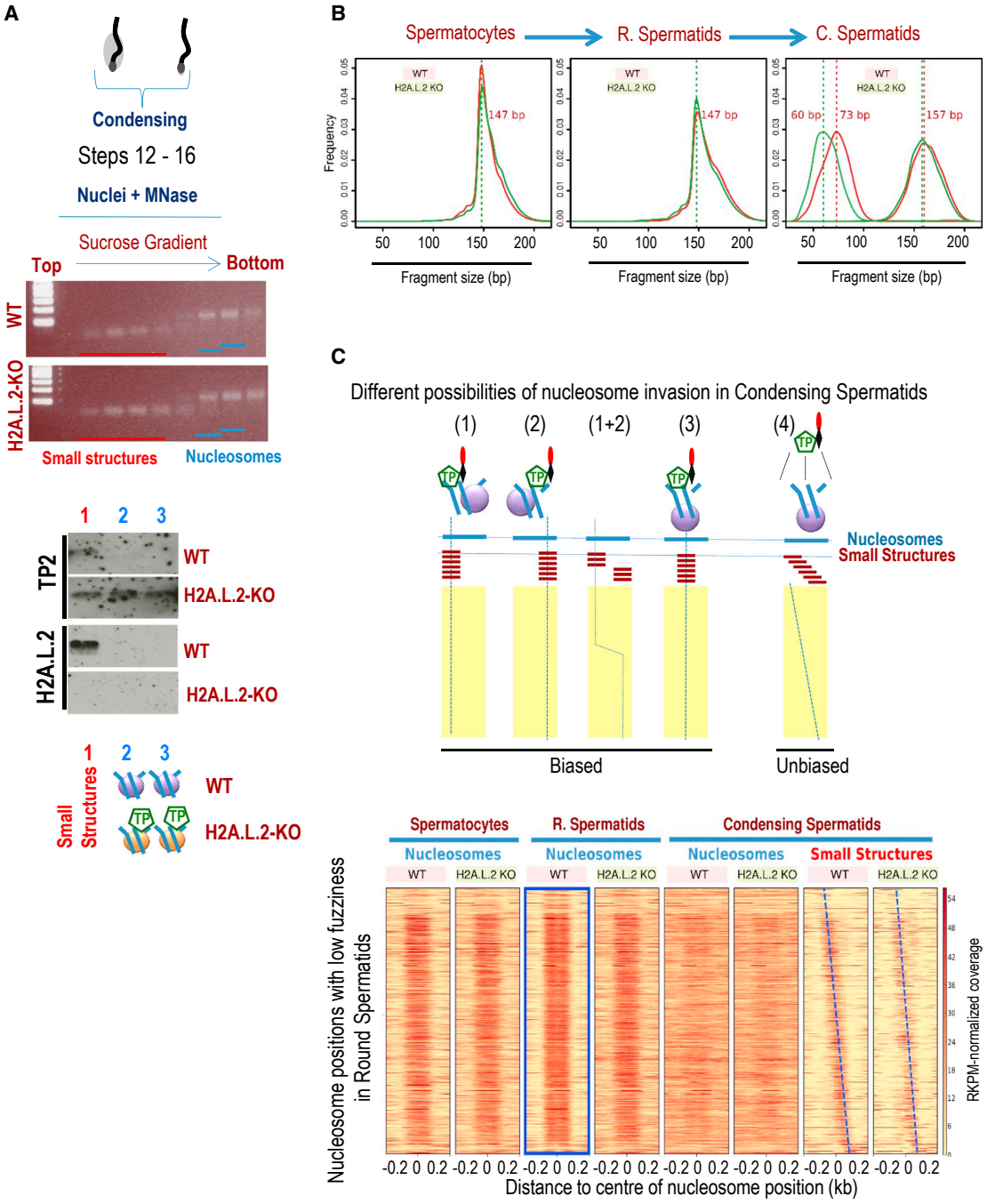


Figure 5. The Absence of H2A.L.2 Affects the Structure of Transitional States In Vivo

(A) Nuclei from condensing spermatids were extensively digested with MNase, and the released structures were loaded on a sucrose gradient to separate the small sub-nucleosomal structures from the nucleosomes. The ethidium bromide-stained gels show different fractions of the gradient collected (upper panels). The indicated fractions were analyzed either after pooling (1) or directly (2 and 3) by immunodetection of TP2 and H2A.L.2. The scheme below represents the nature of complexes in these fractions.

(B) Paired-end sequencing allows the accurate determination of DNA fragment lengths produced by MNase. Standard nucleosome size (147 bp) is found in spermatocytes and round spermatids, both in WT and H2A.L.2-KO samples, whereas the remaining nucleosomes in condensing spermatids produce slightly longer fragments (157 bp). The short MNase fragments observed in the condensing spermatids are smaller in the absence of H2A.L.2 than in WT cells (73 bp versus 60 bp).

(C) Schematic representations depicting various theoretical possibilities for the transformation of nucleosomes of round spermatids into the small structures observed in condensing spermatids. Five different possibilities (among others) were considered here, where the invading TPs/Prms could displace the nucleosomal DNA in biased or unbiased manners (upper panels).

(legend continued on next page)

First, we reconstituted nucleosomes as above but using biotinylated DNA and incubated them with purified TP1 and TP2. Streptavidin beads were used to pull down the corresponding nucleosomes and analyze the presence of associated TP1 and TP2. Figure 4E shows that H2A.L.2-containing nucleosomes have an increased interaction with TP1 compared to H2A-containing nucleosomes (see also Figure S5A). Under the same conditions, TP2 shows a better propensity than TP1 to interact with nucleosomes regardless of the presence of H2A.L.2 (Figure 4E; Figure S5A).

We also analyzed the mode of TP1/TP2 interaction with nucleosomes bearing H2A or H2A.L.2. The reconstituted H2A/H2B and H2A.L.2/TH2B nucleosomes were incubated with increasing amounts of purified TP1 and TP2 and the interaction between TP and nucleosomes was monitored by gel-shift assays (Figure 4F). These experiments revealed different modes of interaction between TPs and canonical nucleosomes or H2A.L.2-containing nucleosomes. In canonical nucleosomes, TPs follow a binding mode, which is reminiscent of the binding of linker histones to nucleosomes, where the loading of additional TP molecules leads to the appearance of defined complexes visible above the nucleosomes (Figure 4F). In contrast, TP loading by H2A.L.2-containing nucleosomes leads to the disappearance of nucleosomes in favor of undefined higher molecular weight complexes without the generation of discrete TP-nucleosome species (Figure 4F). The latter are probably present in different species diluted in a smear above the nucleosome.

We then wanted to know whether the resistance of canonical nucleosomes to TP loading, shown here in vitro, could also be observed in vivo in the absence of H2A.L.2.

Previously, we had shown that extensive MNase digestion of chromatin from condensing spermatids releases remaining nucleosomes as well as transitional structural states (Govin et al., 2007; Montellier et al., 2013). The latter contain TPs, H2A.L.2, and TH2B and are the result of the transformation of nucleosomes at the time of histone eviction (Govin et al., 2007). The same experiment was repeated here to evaluate the role of H2A.L.2. Nuclei from purified condensing spermatids from WT and H2A.L.2-KO mice were incubated with MNase and the resulting released chromatin fragments loaded on a sucrose gradient to separate sub-nucleosomal particles from nucleosomes (Figure 5A). This experiment demonstrated that the transformation of nucleosomes into small sub-nucleosomal particles also takes place in the absence of H2A.L.2 (Figure 5A, upper panel; Figure S5B). Proteins from the different fractions were then analyzed by immunoblotting, confirming the presence of H2A.L.2 and TPs only in the subnucleosomal particles as previously reported (Govin et al., 2007). Interestingly, in the absence of H2A.L.2, TPs were also associated with nucleosomes, which is in drastic contrast with the WT situation, where TPs could only be detected with the sub-nucleosomal structures (Figure 5A, lower panels; Figure S5B).

These data strongly support our observations based on in vitro nucleosome reconstitution and show that, in the absence of H2A.L.2, TPs have a limited capacity to invade the nucleosomes and hence to efficiently mediate the subsequent transformation.

Altogether, these investigations indicate that the incorporation of H2A.L.2 itself is a critical event that "prepares" nucleosomes for the action of additional factors.

Genome-wide Transformation of Nucleosomes into Transitional States in Condensing Spermatids: The Impact of H2A.L.2

To better characterize the transformation of nucleosomes into sub-nucleosomal particles, we sequenced MNase nucleosomal and sub-nucleosomal fragments from condensing spermatids (shown in Figure 5A and Figure S5B). In addition, we sequenced mono-nucleosomal DNA fragments released after extensive MNase digestion of spermatogenic cell populations enriched in meiotic spermatocytes and early post-meiotic round spermatids.

Paired-end sequencing of these fragments allowed us not only to map the position of these fragments on the genome, but also to determine the size distribution of the nucleosomal and sub-nucleosomal fragments in the WT, as well as in H2A.L.2-KO, cells. This analysis showed that, in the absence of H2A.L.2, the nucleosomal fragments present strictly the same size distribution in spermatocytes, round spermatids, and condensing spermatids as in WT cells (Figure 5B). Interestingly, this size distribution analysis also shows that, although sub-nucleosomal particles are generated in the absence of H2A.L.2, they are smaller than sub-nucleosomal fragments observed in the WT condensing spermatids (Figure 5B), suggesting that, in the absence of H2A.L.2, an altered DNA-protein organization results in a modified MNase digestion of the associated DNA.

We also took advantage of this sequencing analysis to better characterize the propensity of various genomic regions to change into transitional states. To this end, we compared the global genomic distribution of the DNA fragments present in the transitional structures of condensing spermatids to the genomic fractions covered by nucleosomes at earlier stages and to the whole genome. This analysis revealed a globally similar genomic distribution for the sub-nucleosomal DNA in condensing spermatids compared to nucleosomal regions at earlier stages (Figure S6), indicating that, in condensing

The lower panels show the Heatmaps of RPKM-normalized coverage around nucleosome positions for the indicated stages and genotypes. The blue frame highlights low fuzziness nucleosome positions in round spermatids defined by the DANPOS peak caller (dynamic analysis of nucleosome position and occupancy by sequencing; see the STAR Methods section). Only the 4,000 positions with lowest fuzziness in round spermatids are depicted. The rows are ordered according to the WT small structure track, using the 5′ to 3′ position of the maximum signal in the −70/+70 bp region around reference nucleosome center. The centers of small structure fragments can be visually approximated by a straight oblique line, depicted with the blue dotted lines on the heatmaps.

Figure S5B shows a more detailed analysis of TP2 distribution in the sub-nucleosomal structures in the material already analyzed in (A). A similar independent experiment as in (A) was performed to also visualize the distribution of TP1.

Figure S6 shows the genomic distribution of the regions associated with well-positioned nucleosomes in spermatocytes and round spermatids or with the small structures in condensing spermatids in WT and H2A.L.2-KO mice.

spermatids, the nucleosomes inherited from previous stages are mainly transformed into transitional states.

The fact that we generated a map of nucleosomal fragments and the corresponding small structures in a stage-specific manner also allows us to consider the mode of transformation of nucleosomes into small structures. In other words, it is possible to map the positions of the 73 bp and 60 bp DNA fragments in condensing spermatids with respect to the corresponding parental nucleosomes in round spermatids. To do so, we first identified nucleosomes with low fuzziness in round spermatids, which correspond to the best-positioned nucleosomes (minimal variation of position between cells) (Figure 5C, lower panel, blue frame, "Round Spermatids"). Interestingly, these nucleosomes have a tendency to occupy the same position in spermatocytes (Figure 5C, lower panel, "Spermatocytes"). The majority of these nucleosomes are transformed into small structures in condensing spermatids, since only a small proportion of these positions is present in the nucleosomal fraction of condensing spermatids (Figure 5C, condensing spermatids, nucleosomes).

In order to precisely map the DNAs corresponding to the small structures in condensing spermatids with respect to their parental nucleosomes in round spermatids, we ordered these fragments from 5′ to 3′ of the parental nucleosome.

Theoretically, the transformation of nucleosomes could occur following different scenarios (depicted in Figure 5C, upper panels). A biased invasion of nucleosomes from either side or from the dyad would result in a biased mapping of the corresponding small structures (Figure 5C, upper panel, biased), whereas an unbiased invasion of nucleosomes would lead to the generation of small structures mapping all along the corresponding parental nucleosomal DNA (Figure 5C, upper panel, unbiased).

Following this relative mapping of the small structures in condensing spermatids, our results suggest an unbiased invasion of the parental nucleosomes (Figure 5C, lower panels, condensing spermatids, small structures, the blue dotted lines highlight the order of the mapped DNA fragments).

Our results are in agreement with a model where Prms would initiate their assembly on DNA in the transforming nucleosomes at any positions in both WT and H2A.L.2-KO condensing spermatids. H2A.L.2 does not seem to modify this mode of nucleosome transformation. However, the observed size difference of MNase-protected DNA in these small structures between the two genotypes demonstrates that the protein organization in these structures is dependent on H2A.L.2. This could be partly explained by the deficiency in mature Prm2 observed in H2A.L.2-KO cells (Figure 3B).

Mechanism of H2A.L.2 Assembly

All the data presented here so far show that H2A.L.2 incorporation is an important event in the transformation of nucleosomes into nucleoprotamines. The important question is whether there is a testis-specific system specialized in mediating the incorporation of H2A.L.2, which would hence trigger the nucleosome transformation process.

To address this question, we took advantage of our previous finding that TH2B preferentially dimerizes with H2A.L.2 and that TH2B is present with H2A.L.2 in the transitional structures (Govin et al., 2007), as well as of our mouse model expressing an endogenous TH2B-tag histone (Montellier et al., 2013). This tag system allowed us to perform a highly reliable and efficient tandem affinity purification (TAP) of soluble TH2B, enabling us to identify its associated chaperone systems.

Soluble TH2B was purified in two independent experiments using first a total germ cell suspension (experiment 1) and second round/elongating spermatid-enriched cell fractions (experiment 2). In both experiments, the extracts were submitted to a TAP-tag protocol involving an anti-Flag immunoprecipitation and elution of the immunoprecipitated materials with the Flag peptide followed by an anti-hemagglutinin (HA) immunoprecipitation (Figure 6A). Mass spectrometry-based proteomic analyses were performed in order to identify and compare the proteins present in the eluates of control and TH2B-tag purifications. All the relevant proteins identified in these two experiments are listed in Table S1.

The outcome of this experiment is that TH2B dimerizes with various H2A variants and hence associates with known H2A variant-specific chaperone systems, suggesting that the H2A part of the TH2B heterodimer is a determinant element in defining the type of chaperone involved.

The analysis of these chaperone systems revealed the presence of the members of the two known H2A.Z chaperone systems, Tip60 (except ANP32E) and SRCAP (Latrick et al., 2016; Obri et al., 2014), as well as of generalist H2A-H2B chaperones, including FACT, Nap1L1, and Nap1L4. Interestingly, we also found MCM2, MCM4, MCM6, and MCM7, previously identified as H3-H4 chaperones in DNA-replicating cells (Groth et al., 2007) (Figure 6A). Since round/elongating spermatids do not replicate their DNA, in this particular context, MCM2–MCM7 may mostly contribute to nucleosome disassembly.

In order to define the specific H2A.L.2 chaperone system and also to test whether the H2A.Z chaperone complexes could also control H2A.L.2 assembly, we prepared soluble extracts from total germ cells and immunoprecipitated H2A.L.2. As expected, H2A.L.2 was found in a complex with TH2B, but neither H2A nor H2B were detected (Figure 6B). We then probed the H2A.L.2 co-immunoprecipitated materials for the presence of RuvBL1 and RuvBL2 present in both Tip60 and SRCAP H2A.Z complexes, as well as for YL1 shown to be the direct chaperone of H2A.Z (Latrick et al., 2016; Obri et al., 2014). Although all these proteins were easily detected in the extract (Figure 6B, Input), none were found in interaction with H2A.L.2. In contrast, when we tested the regular H2A-H2B chaperone, Nap1L4, we found that it co-purifies with H2A.L.2 (Figure 6B).

In a previous work, aiming at the identification of chaperones mediating the many genome-packaging transitions that occur in condensing spermatids, we used a strategy based on an affinity column containing recombinant TP1. Starting with soluble extracts from condensing spermatids, this approach purified, on the one hand, all the acidic proteins binding to TP1 through charge interaction (including chaperones for basic proteins) and, on the other hand, specific TP1 interactants (Govin et al., 2012). This work had identified a group of four chaperones with the highest propensity to act on the elongating spermatid genome, which are Nap1L4, nucleoplasmin 3 (Nmp3), HspA2, and Nasp (Govin et al., 2012). Based on these previous data and in direct connection with the present work, we prepared soluble extracts from total

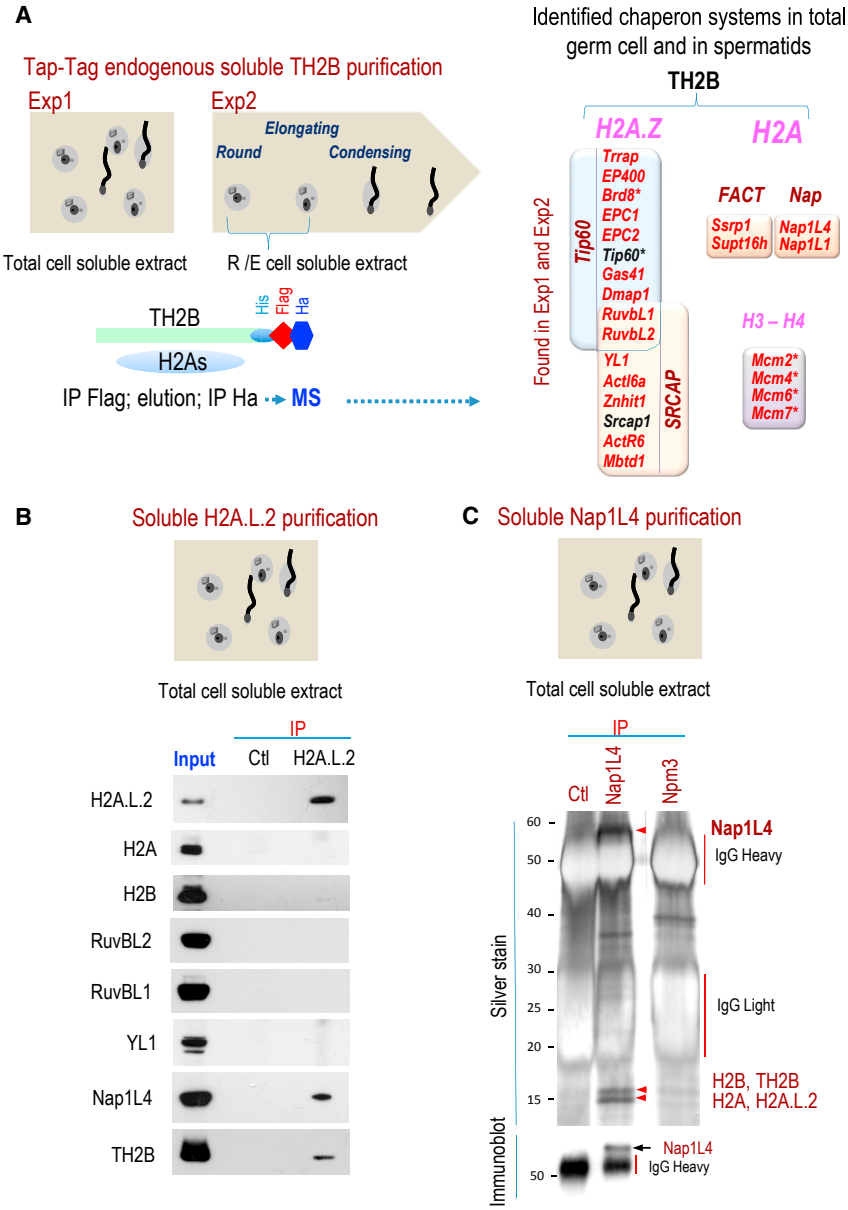


Figure 6. Identification of the Chaperone Systems Involved in the Assembly of H2A.L.2

(A) Testes from WT mice and mice expressing a C-terminally tagged endogenous TH2B were used to prepare total spermatogenic cell soluble extracts or to isolate round/elongating enriched cell fractions and prepare the corresponding soluble extracts. TH2B-tag was then immunoprecipitated following the tandem purification protocol, first with anti-Flag antibody and then with an anti-HA antibody after elution of the first immunoprecipitated materials. Mass spectrometry was then used to identify all the associated proteins. Table S1 presents the list of the chromatin-related identified proteins. Known H2A.Z as well as H2A-H2B chaperones were identified only in the TH2B-tag immunoprecipitates as indicated. Experiment 1 (exp1) was performed on total spermatogenic cells, whereas experiment 2 (exp2) was performed on fractionated round and elongating spermatids. Asterisks (*) indicate the proteins identified only in exp2.

(B) Total spermatogenic cell soluble extracts were prepared and H2A.L.2 immunoprecipitated. Control antibody corresponded to non-specific rabbit Ig. The immunoprecipitated materials were then probed with the indicated antibodies.

(C) The same type of extracts as in (B) were used to immunoprecipitate Nap1L4 and Nucleoplasmin 3 (Npm3). The immunoprecipitated materials were visualized on a silver-stained gel, and the indicated doublet bands in the Nap1L4 immunoprecipitation were identified by MS. Figure S7 shows that both Nap1L4 and Npm3 are predominantly expressed in testis and data on the identification of histones co-immunoprecipitated with Nap1L4.

germ cells and used our anti-Nap1L4 and anti-Npm3 antibodies, both preferentially expressed in testis (Figure S7A), to detect their chaperoned proteins. Interestingly, and in agreement with the data presented above, only anti-Nap1L4 immunoprecipitated a doublet that corresponded to TH2B, H2A.L.2, and H2B-H2A species (Figure 6C; Figure S7). No histone of the H2A or H2B types was found associated with Nmp3.

Altogether, these data show that, in contrast to the specific histone variant H2A.Z, whose assembly requires energy consuming factors to remodel chromatin, H2A.L.2 uses a chaperone system, which normally controls H2A-H2B dynamics to trigger all the subsequent chromatin reorganizations. This finding, along with the biochemical characterization of H2A.L.2 (Figure 4), suggests that the simple exchange of TH2B-H2A by TH2B-H2A.L.2, through the action of Nap1L4, and the presence of Prms and TPs are enough to initiate the process of nucleosome transformation.

DISCUSSION

Since 2004, the date of the publication of the phenotype of TP1/2 double-KO mice by Meistrich laboratory (Shirley et al., 2004; Zhao et al., 2004a, 2004b), no noticeable progress in the understanding of the functional interplay between histones, TPs, and Prms had been made. The identification of H2A.L.2 (Govin et al., 2007) and the subsequent functional investigations reported here are now shedding light on the molecular events driving the transformation of nucleosomes into nucleoprotamines and also provide an insight into the function of TPs.

Histone variants known to open the nucleosome have the potential to create an interaction interface required for the assembly of specific structural non-histone proteins on chromatin. The best-known example is the centromeric H3 variant CENP-A, which organizes only 121 bp of DNA when present in a nucleosome (Tachiwana et al., 2011). The resulting open

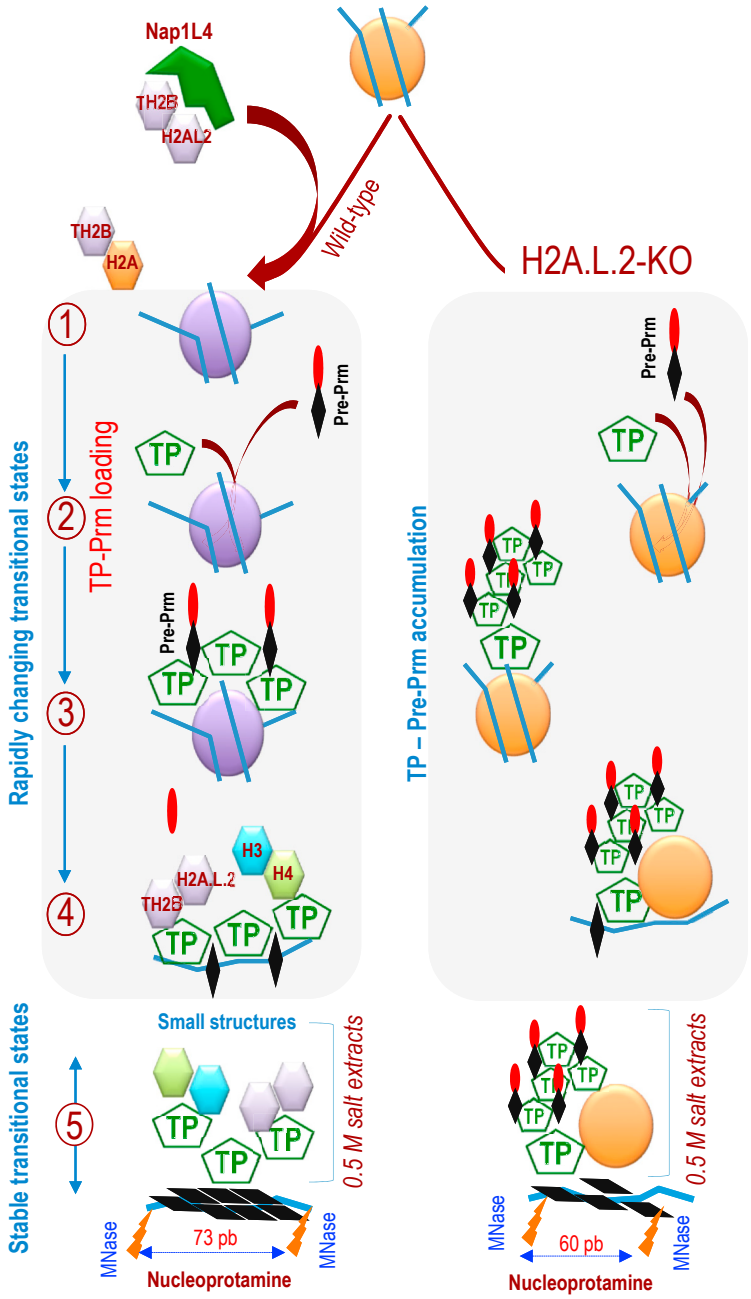


Figure 7. Molecular Model for the Functional Interplay between Histones, TPs, and Prms during the Transformation of Nucleosomes into Nucleoprotamines

In WT condensing spermatids, TPs and H2A.L.2 are co-expressed (Figure 1) and the H2A.L.2-TH2B dimer is loaded onto the nucleosomes by Nap1L4 (Figure 6). The incorporation of H2A.L.2-TH2B into chromatin opens the nucleosomes, releasing the nucleosomal DNA ends and allowing the invasion of nucleosomes by TPs, as suggested by the observation that H2A.L.2 is required for the incorporation of TPs into nucleosomes (Figures 3A, 4, and 5A; Figure S5B). Nucleosome-TPs interact with incoming Prms (Figure 3B; Figure S4) and mediate pre-Prm2 processing (Figures 2D and 3B) (stages 1–3). Prms bind DNA in competition with histone-TP complexes (stage 4). The displaced histones are unable to remain as octamers (supported by the fact that DNA is necessary for the constitution of H2A.L.2-containing octamers; see Figure 4D) (stage 4). Prm-DNA and displaced TP-histone complexes constitute a relatively stable transitional state (stage 5), corresponding to the small subnucleosomal structures, which can be visualized on a sucrose gradient (Figure 5A; Figure S5B) and sequenced (Figures 5B and 5C).

In the absence of H2A.L.2, H2A-containing nucleosomes impair the invasion of TPs (Figures 3, 4, and 5A; Figures S3B and S5B), leading to the accumulation of non-incorporated TPs around chromatin (Figure 3; Figure S3B), which in turn trap incoming Prms (Figure 3B; Figure S4), preventing their assembly and the processing of Prm2 (Figures 2D and 3B). The accumulation of "TP-misassembled Prms" would lead to Prm under-dosage and the generation of modified transitional states (Figure 5B). The fraction of Prms that achieves DNA assembling would be enough to release the "histone-TP" complexes (Figure 3B) but would not be sufficient to sustain a complete genome compaction (Figure 2D).

nucleosomes with flexible DNA ends are thought to facilitate the loading of the non-histone kinetochore proteins (Roulland et al., 2016).

The H2A variant H2A.L.2, expressed at the time of histone replacement, creates nucleosomes, which, like CENP-A nucleosomes, present an open structure with flexible DNA ends that are prone to be loaded with non-histone proteins. These non-histone proteins here are TPs, which are co-synthesized with H2A.L.2 and are loaded onto H2A.L.2-containing nucleosomes. Indeed, we show here that H2A.L.2's primary function is to allow TP loading onto chromatin. However, TP loading does not directly lead to histone replacement but rather generates an interface for Prms' recruitment and processing. Therefore, we can propose that TPs' function is essentially to buffer and orient the action of Prms in displacing histones. Our data, along with Meistrich laboratory's work, support the idea that Prms are the true mediators of histone displacement and the inducers of the transitional states formation revealed by MNase digestion. Indeed, the detailed analysis of these transitional states shows that the absence of H2A.L.2 and TPs mislocalization do not prevent their generation but rather impact their Prm-dependent organization. This interpretation is experimentally supported by the protection of shorter DNA fragments against MNase digestion, as well as defective genome compaction in mature sperm cells, which could result from a sub-optimal concentration of assembled Prms. Interestingly, a similar scenario seems to occur in the total absence of TPs, where Prm processing and assembly could also be suboptimal (Shirley et al., 2004; Zhao et al., 2004a).

In conclusion, this work not only describes, at a molecular level, the functional interconnection between histones, TPs, and Prms during histone-to-protamine replacement (Figure 7), but also highlights the function of TPs and invites biologists to update the classical description of histone/TP/Prm transitions

by the addition of a critical step corresponding to the incorporation of H2A.L.2.

STAR★METHODS

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

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Generation of H2A.L.2-KO Mice
- METHOD DETAILS
  - Spermatogenic Cell Fraction Purification
  - In Vitro Fertilization
  - Protein Profiling by Mass Spectrometry
  - Mass Spectrometry-Based Proteomic Analyses
  - Spermatogenic Cell Chromatin Preparation and Separation of Nucleosomal and Sub-nucleosomal Particles on a Sucrose Gradient
  - Microscopy and In Situ Immunodetection
  - Immunoprecipitations
  - Purification of Recombinant Histones, TPs, and In Vitro Assays
  - Preparation of Histone Complexes
  - Preparation of DNA Fragments for Nucleosome Reconstitution
  - Reconstitution of Nucleosomes
  - Gel Filtration Analysis
  - MNase Treatment Assay
  - TP Binding Assay
  - TP-Nucleosome Interaction by Pull-Down Assay
  - Antibodies' Dilutions
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Bioinformatics
- DATA AND SOFTWARE AVAILABILITY
- ADDITIONAL RESOURCES
  - Non-commercial Antibodies

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and one table and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2017.02.025.

AUTHOR CONTRIBUTIONS

S.B. performed in situ and biochemical analyses, prepared nucleosomal and subnucleosomal fragments for sequencing, and carried out chromatin fractionation on sucrose gradient and subsequent analyses, as well as IH, IF, and spermatogenic cell fractionations. Y.M., H.T., and T.U. conducted all the in vitro reconstitutions and subsequent functional studies under the supervision of H.K. E.M. performed the phenotypical characterizations of H2A.L.2-KO mice, including in vitro and in vivo fertilization tests with help of T.B., electron microscopy and testis histology analysis, and TP and Prm detection by IH. T.B. helped with spermatogenic cell fractionation and H2A.L.2-KO mice maintenance, genotyping, and phenotypical characterizations. J.G. generated anti-Nap1L4 and anti-NPM3 antibodies and performed Nap1L4 purification. M.d.D. generated H2A.L.2-KO mice with the help of P.H. and F.B. in M.G.'s laboratory under his supervision. Y.C. performed the MS protein analyses. F.F. performed UHPLC-MS analyses of histone extracts. S.C. performed transfection IP analyses of TP-Prm interactions. Y.M. performed TAP-tag purification of soluble TH2B, with the help of H.S. G.C., D.P., and S.R. performed all the bioinformatics analyses. N.F.-N. prepared the sequencing libraries and run the sequencing. S.R. helped with the manuscript and figure preparation. S.K. designed and coordinated the whole project and wrote the manuscript. All the authors read and commented the manuscript.

ACKNOWLEDGMENTS

S.K.'s laboratory is supported by a grant from "Foundation pour la Recherche Medicale (FRM)" "analyse bio-informatique pour la recherche en biologie" program (DBI20141231321), as well as by ANR Episperm3 program (ANR-15-CE12-0005-01), and by INCa libre program (RPT13001CCA). Additional supports were from: Fondation ARC "Canc'air" project (RAC16042CLA), Plan Cancer (CH7-INS15B66) and Plan Cancer (ASC16012CSA), and the "Université Grenoble Alpes" ANR-15-IDEX-02. H.S. was a recipient of a Marie Curie Initial Training Network funded by European Commission (FP7-PEOPLE-2011-ITN, PITN-GA-289880) for 3 years and then supported by "La Ligue Nationale Contre Le Cancer" for a fourth year of PhD. Y.M. is supported by a post-doctoral fellowship from Plan Cancer (CH7-INS15B66). G.C. is a bioinformatics expert benefiting from a FRM fellowship. Mice were bred in the High Technology Animal Facility (PHTA) of Grenoble University with the precious help of Cyrielle Colomb and Andrea Piras. We also gratefully acknowledge the help of Karin Pernet in the preparation of samples and their analysis by electron microscopy. High-throughput sequencing was performed at the TGML Platform, supported by grants from Inserm, GIS IBiSA, Aix-Marseille Université, and ANR-10-INBS-0009-10. The proteomic analyses were partially supported by the French National Research Agency ProFi Grant ANR-10-INBS-08-01. This work was supported in part by JSPS KAKENHI Grant Numbers JP25116002 and JP25250023 to H.K. H.K. was supported by the Waseda Research Institute for Science and Engineering, and was also supported by the programs of Waseda University.

Received: September 26, 2016
Revised: January 3, 2017
Accepted: February 24, 2017
Published: March 30, 2017

REFERENCES

Arimura, Y., Kimura, H., Oda, T., Sato, K., Osakabe, A., Tachiwana, H., Sato, Y., Kinugasa, Y., Ikura, T., Sugiyama, M., et al. (2013). Structural basis of a nucleosome containing histone H2A.B/H2A.Bbd that transiently associates with reorganized chromatin. Sci. Rep. *3*, 3510.

Bao, Y., Konesky, K., Park, Y.J., Rosu, S., Dyer, P.N., Rangasamy, D., Tremethick, D.J., Laybourn, P.J., and Luger, K. (2004). Nucleosomes containing the histone variant H2A.Bbd organize only 118 base pairs of DNA. EMBO J. *23*, 3314–3324.

Buchou, T., Tan, M., Barral, S., Vitte, A.L., Rousseaux, S., Arechaga, J., and Khochbin, S. (2017). Purification and analysis of male ferm cells from adult mouse testis. Methods Mol. Biol. *1510*, 159–168.

Cho, C., Willis, W.D., Goulding, E.H., Jung-Ha, H., Choi, Y.C., Hecht, N.B., and Eddy, E.M. (2001). Haploinsufficiency of protamine-1 or -2 causes infertility in mice. Nat. Genet. *28*, 82–86.

Cho, C., Jung-Ha, H., Willis, W.D., Goulding, E.H., Stein, P., Xu, Z., Schultz, R.M., Hecht, N.B., and Eddy, E.M. (2003). Protamine 2 deficiency leads to sperm DNA damage and embryo death in mice. Biol. Reprod. *69*, 211–217.

Chua, E.Y., Vasudevan, D., Davey, G.E., Wu, B., and Davey, C.A. (2012). The mechanics behind DNA sequence-dependent properties of the nucleosome. Nucleic Acids Res. *40*, 6338–6352.

Contrepois, K., Ezan, E., Mann, C., and Fenaille, F. (2010). Ultra-high performance liquid chromatography-mass spectrometry for the fast profiling of histone post-translational modifications. J. Proteome Res. *9*, 5501–5509.

De Vries, M., Ramos, L., Housein, Z., and De Boer, P. (2012). Chromatin remodelling initiation during human spermiogenesis. Biol. Open *1*, 446–457.

Doyen, C.M., Montel, F., Gautier, T., Menoni, H., Claudet, C., Delacour-Larose, M., Angelov, D., Hamiche, A., Bednar, J., Faivre-Moskalenko, C., et al. (2006). Dissection of the unusual structural and functional properties of the variant H2A.Bbd nucleosome. EMBO J. *25*, 4234–4244.

Dyer, P.N., Edayathumangalam, R.S., White, C.L., Bao, Y., Chakravarthy, S., Muthurajan, U.M., and Luger, K. (2004). Reconstitution of nucleosome core particles from recombinant histones and DNA. Methods Enzymol. *375*, 23–44.

Gaucher, J., Reynoird, N., Montellier, E., Boussouar, F., Rousseaux, S., and Khochbin, S. (2010). From meiosis to postmeiotic events: the secrets of histone disappearance. FEBS J. *277*, 599–604.

Goudarzi, A., Shiota, H., Rousseaux, S., and Khochbin, S. (2014). Genome-scale acetylation-dependent histone eviction during spermatogenesis. J. Mol. Biol. *426*, 3342–3349.

Govin, J., Escoffier, E., Rousseaux, S., Kuhn, L., Ferro, M., Thévenon, J., Catena, R., Davidson, I., Garin, J., Khochbin, S., and Caron, C. (2007). Pericentric heterochromatin reprogramming by new histone variants during mouse spermiogenesis. J. Cell Biol. *176*, 283–294.

Govin, J., Gaucher, J., Ferro, M., Debernardi, A., Garin, J., Khochbin, S., and Rousseaux, S. (2012). Proteomic strategy for the identification of critical actors in reorganization of the post-meiotic male genome. Mol. Hum. Reprod. *18*, 1–13.

Groth, A., Corpet, A., Cook, A.J., Roche, D., Bartek, J., Lukas, J., and Almouzni, G. (2007). Regulation of replication fork progression through histone supply and demand. Science *318*, 1928–1931.

Ichikawa, Y., Kagawa, W., Saito, K., Chikashige, Y., Haraguchi, T., Hiraoka, Y., and Kurumizaka, H. (2013). Purification and characterization of the fission yeast telomere clustering factors, Bqt1 and Bqt2. Protein Expr. Purif. *88*, 207–213.

Latrick, C.M., Marek, M., Ouararhni, K., Papin, C., Stoll, I., Ignatyeva, M., Obri, A., Ennifar, E., Dimitrov, S., Romier, C., and Hamiche, A. (2016). Molecular basis and specificity of H2A.Z-H2B recognition and deposition by the histone chaperone YL1. Nat. Struct. Mol. Biol. *23*, 309–316.

Lewis, J.D., Song, Y., de Jong, M.E., Bagha, S.M., and Ausió, J. (2003). A walk though vertebrate and invertebrate protamines. Chromosoma *111*, 473–482.

Lowary, P.T., and Widom, J. (1998). New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. J. Mol. Biol. *276*, 19–42.

Meistrich, M.L., Mohapatra, B., Shirley, C.R., and Zhao, M. (2003). Roles of transition nuclear proteins in spermiogenesis. Chromosoma *111*, 483–488.

Montellier, E., Boussouar, F., Rousseaux, S., Zhang, K., Buchou, T., Fenaille, F., Shiota, H., Debernardi, A., Héry, P., Curtet, S., et al. (2013). Chromatin-to-nucleoprotamine transition is controlled by the histone H2B variant TH2B. Genes Dev. *27*, 1680–1692.

Nekrasov, M., Amrichova, J., Parker, B.J., Soboleva, T.A., Jack, C., Williams, R., Huttley, G.A., and Tremethick, D.J. (2012). Histone H2A.Z inheritance during the cell cycle and its impact on promoter organization and dynamics. Nat. Struct. Mol. Biol. *19*, 1076–1083.

Obri, A., Ouararhni, K., Papin, C., Diebold, M.L., Padmanabhan, K., Marek, M., Stoll, I., Roy, L., Reilly, P.T., Mak, T.W., et al. (2014). ANP32E is a histone chaperone that removes H2A.Z from chromatin. Nature *505*, 648–653.

Roulland, Y., Ouararhni, K., Naidenov, M., Ramos, L., Shuaib, M., Syed, S.H., Lone, I.N., Boopathi, R., Fontaine, E., Papai, G., et al. (2016). The flexible ends of CENP-A nucleosome are required for mitotic fidelity. Mol. Cell *63*, 674–685.

Rousseaux, S., Boussouar, F., Gaucher, J., Reynoird, N., Montellier, E., Curtet, S., Vitte, A.L., and Khochbin, S. (2011). Molecular models for post-meiotic male genome reprogramming. Syst Biol Reprod Med *57*, 50–53.

Shinagawa, T., Takagi, T., Tsukamoto, D., Tomaru, C., Huynh, L.M., Sivaraman, P., Kumarevel, T., Inoue, K., Nakato, R., Katou, Y., et al. (2014). Histone variants enriched in oocytes enhance reprogramming to induced pluripotent stem cells. Cell Stem Cell *14*, 217–227.

Shirley, C.R., Hayashi, S., Mounsey, S., Yanagimachi, R., and Meistrich, M.L. (2004). Abnormalities and reduced reproductive potential of sperm from Tnp1- and Tnp2-null double mutant mice. Biol. Reprod. *71*, 1220–1229.

Soboleva, T.A., Nekrasov, M., Pahwa, A., Williams, R., Huttley, G.A., and Tremethick, D.J. (2011). A unique H2A histone variant occupies the transcriptional start site of active genes. Nat. Struct. Mol. Biol. *19*, 25–30.

Syed, S.H., Boulard, M., Shukla, M.S., Gautier, T., Travers, A., Bednar, J., Faivre-Moskalenko, C., Dimitrov, S., and Angelov, D. (2009). The incorporation of the novel histone variant H2AL2 confers unusual structural and functional properties of the nucleosome. Nucleic Acids Res. *37*, 4684–4695.

Tachiwana, H., Kagawa, W., Osakabe, A., Kawaguchi, K., Shiga, T., Hayashi-Takanaka, Y., Kimura, H., and Kurumizaka, H. (2010). Structural basis of instability of the nucleosome containing a testis-specific histone variant, human H3T. Proc. Natl. Acad. Sci. USA *107*, 10454–10459.

Tachiwana, H., Kagawa, W., Shiga, T., Osakabe, A., Miya, Y., Saito, K., Hayashi-Takanaka, Y., Oda, T., Sato, M., Park, S.Y., et al. (2011). Crystal structure of the human centromeric nucleosome containing CENP-A. Nature *476*, 232–235.

Tanaka, Y., Tawaramoto-Sasanuma, M., Kawaguchi, S., Ohta, T., Yoda, K., Kurumizaka, H., and Yokoyama, S. (2004). Expression and purification of recombinant human histones. Methods *33*, 3–11.

Tropberger, P., and Schneider, R. (2013). Scratching the (lateral) surface of chromatin regulation by histone modifications. Nat. Struct. Mol. Biol. *20*, 657–661.

Zhao, M., Shirley, C.R., Hayashi, S., Marcon, L., Mohapatra, B., Suganuma, R., Behringer, R.R., Boissonneault, G., Yanagimachi, R., and Meistrich, M.L. (2004a). Transition nuclear proteins are required for normal chromatin condensation and functional sperm development. Genesis *38*, 200–213.

Zhao, M., Shirley, C.R., Mounsey, S., and Meistrich, M.L. (2004b). Nucleoprotein transitions during spermiogenesis in mice with transition nuclear protein Tnp1 and Tnp2 mutations. Biol. Reprod. *71*, 1016–1025.

2.2.3 Rôle du facteur NUT dans la vague d'acétylation des histones

Le facteur NUT (NUclear protein in Testis) est une protéine exprimée dans le testicule également impliquée dans des carcinomes hautement agressifs, les *NUT midline carcinoma*. Ces cancers sont caractérisés par une translocation chromosomique qui aboutit dans 90% des cas à la fusion des gènes *BRD4* ou *BRD3*, correspondants à des protéines à double bromodomaine de la famille BET, avec le gène *NUT* (FRENCH, MIYOSHI, KUBONISHI et al. 2003). Jusqu'à l'étude réalisée, la fonction physiologique de cette protéine était inconnue. En revanche, l'activité oncogénique de la protéine de fusion BRD4-NUT entraîne la formation de domaines chromatiniens acétylés après recrutement de l'histone acétyltransférase p300 (REYNOIRD, SCHWARTZ, DELVECCHIO et al. 2010).
Afin de déterminer le rôle de NUT dans des conditions physiologiques, un modèle de souris invalidant son gène a été généré. Les résultats des approches expérimentales comparant ce modèle à un modèle *wild-type* ont fait l'objet d'une publication (ARTICLE 4 ; SHIOTA, BARRAL, BUCHOU et al. 2018). Celles-ci ont montré que Nut recrute p300 ou CBP pour augmenter l'acétylation des lysines 5 et 8 de l'histone 4. L'inactivation de *Nut* induit la stérilité masculine avec un arrêt de la spermatogénèse au stade de remplacement des histones. Dans le modèle moléculaire proposé, l'absence d'induction de vague d'acétylation empêche la fixation de Brdt, dont le rôle est de médier le remplacement des histones.
En ce qui concerne les approches bioinformatiques génomiques, des ChIP-seq pour H4K5ac, RNA-seq et biopuces ont été réalisés sur les spermatides des deux modèles. Les données RNA-seq et biopuce ont indépendamment permis d'identifier les gènes différentiellement exprimés en absence de Nut (FIGURE 6A de l'ARTICLE 4). Alors qu'une diminution d'environ 60% de l'acétylation est observée dans le modèle *Nut KO* par des approches protéomiques (FIGURE 5A de l'ARTICLE 4), l'analyse des données ChIP-seq n'a pu permettre d'observer la même tendance globale de façon aussi marquée, très probablement du fait de l'absence d'un contrôle spike-in fonctionnel pour la normalisation (ORLANDO, CHEN, BROWN et al. 2014).
En revanche, il a été possible d'identifier une légère redistribution locale de l'acétylation au niveau des gènes différentiellement exprimés en accord avec nos précédents résultats (SECTION 2.2.1, GOUDARZI, ZHANG, HUANG et al. 2016) et une redistribution plus marquée au niveau des pics ChIP-seq de Brdt (GSM984200) et BRD4 (GSM1519002) produits par d'autres études (FIGURE 2.2). La fixation préférentielle de Brdt sur les régions où Nut a induit une hyperacétylation est une interprétation de ces résultats en accord avec le rôle connu de Brdt (FIGURE 7 de l'ARTICLE 4). Un rôle similaire pour BRD4 dans un contexte physiologique n'avait pas encore été suggéré et doit encore être caractérisé expérimentalement. Le protocole de ChIP-seq utilisé reposant sur une digestion à la MNase, il a été également possible d'identifier de subtiles différences de sensibilité à l'enzyme suggérant des différences de l'état de la chromatine entre les deux modèles (FIGURE 4 de l'ARTICLE 4).

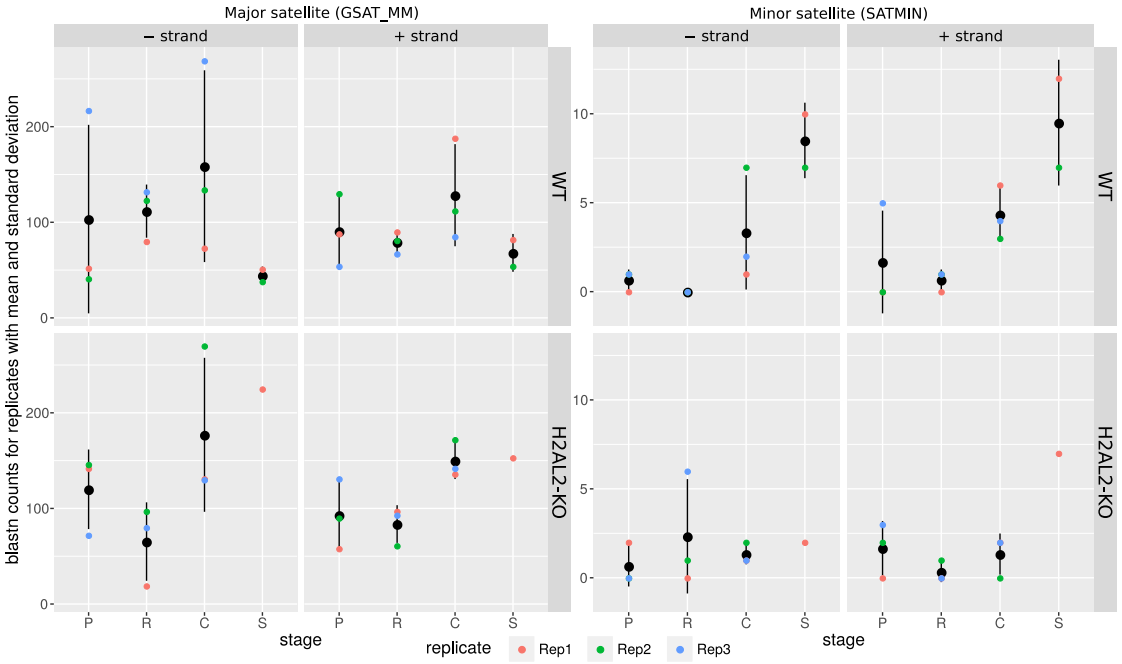


FIGURE 2.1 – Transcription dans les régions satellites péricentromériques en fonction de la présence de H2AL2 pour chaque stade et brin. Aucune différence du niveau transcriptomique entre les deux modèles n'est observée. P : Pachètene. R : Spermatide Ronde. C : Spermatide Condensée. S : Spermatozoïde.

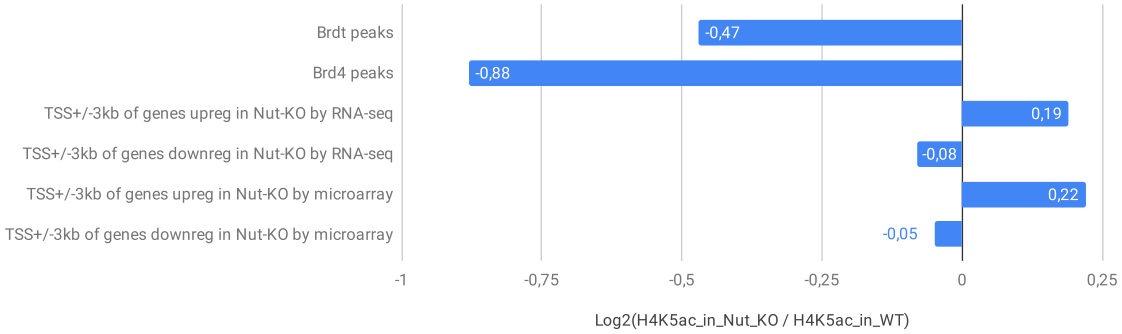


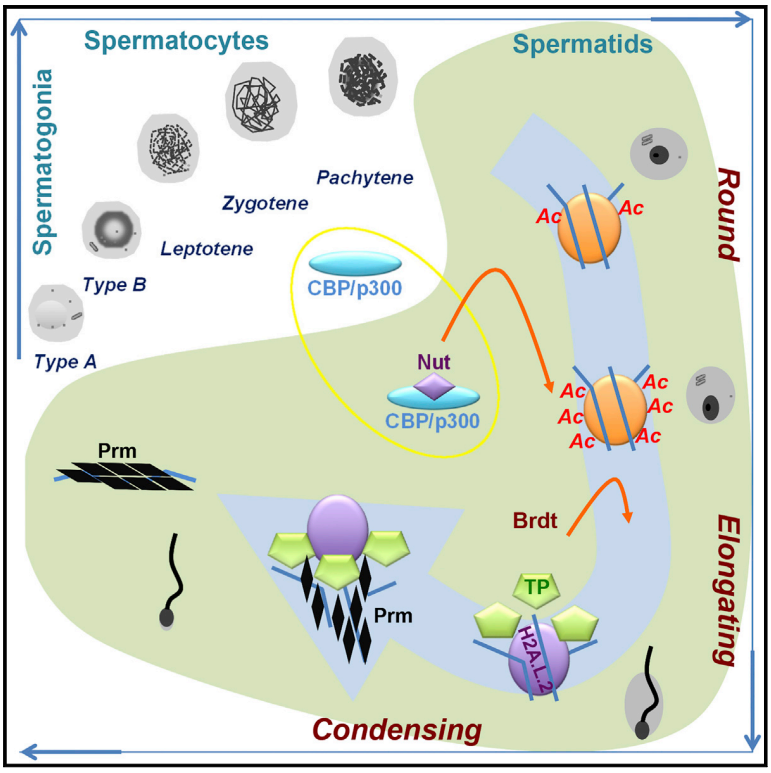
FIGURE 2.2 – Redistribution locale de H4K5ac en l'absence de Nut. Les régions fixées par Brdt et Brd4 dans les spermatides sont des régions hyperacétylées par Nut dans un contexte physiologique. Les régions promotrices des gènes uprégulés dans chacun des modèles, Nut-KO ou WT, sont associées à une H4K5-acétylation plus importante dans ce modèle. Ce résultat est observé que les gènes différentiellement exprimés entre les modèles soient déterminés à partir du RNA-seq ou de biopuces.

Article

Cell Reports

Nut Directs p300-Dependent, Genome-Wide H4 Hyperacetylation in Male Germ Cells

Graphical Abstract



Authors

Hitoshi Shiota, Sophie Barral, Thierry Buchou, ..., Sophie Rousseaux, Yingming Zhao, Saadi Khochbin

Correspondence

saadi.khochbin@univ-grenoble-alpes.fr

In Brief

A transcription-independent histone hyperacetylation is associated with near-total histone replacement during mouse spermatogenesis. Shiota et al. show the oncogenic factor Nut is expressed in post-meiotic male germ cells, where it recruits p300 and/or CBP and enhances histone H4K5 and H4K8 acetylation, leading to histone-to-protamine replacement.

Highlights

- *Nut* is a post-meiotically expressed gene that is critical for male fertility
- Nut recruits p300 and/or CBP to enhance histone H4K5 and H4K8 acetylation
- Nut-mediated histone hyperacetylation is required for histone-to-protamine transition

Data Resources

GSE111931
GSE118969
PXD008727

Shiota et al., 2018, Cell Reports *24*, 3477–3487
September 25, 2018
https://doi.org/10.1016/j.celrep.2018.08.069

CellPress

Cell Reports
Article

Nut Directs p300-Dependent, Genome-Wide H4 Hyperacetylation in Male Germ Cells

Hitoshi Shiota,[1,10] Sophie Barral,[1,10] Thierry Buchou,[1,10] Minjia Tan,[2] Yohann Couté,[3] Guillaume Charbonnier,[4] Nicolas Reynoird,[1] Fayçal Boussouar,[1] Matthieu Gérard,[5] Mingrui Zhu,[2] Lisa Bargier,[4] Denis Puthier,[4] Florent Chuffart,[1] Ekaterina Bourova-Flin,[1] Sarah Picaud,[6] Panagis Filippakopoulos,[6] Afsaneh Goudarzi,[1] Ziad Ibrahim,[7] Daniel Panne,[7,8] Sophie Rousseaux,[1] Yingming Zhao,[9] and Saadi Khochbin[1,11,*]

[1]CNRS UMR 5309, Inserm U1209, Université Grenoble Alpes, Institute for Advanced Biosciences, Grenoble 38700, France
[2]Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, P.R. China
[3]Université Grenoble Alpes, CEA, Inserm U1038, CEA, BIG-BGE, Grenoble 38000, France
[4]TGML, platform IbiSA, Aix Marseille Université, Inserm U1090, TAGC, Marseille, France
[5]CEA, iBiTec-S, Gif-sur-Yvette 91191, France
[6]Structural Genomics Consortium & Ludwig Institute for Cancer Research, Old Road Campus Research Building, Roosevelt Drive, Oxford OX3 7DQ, UK
[7]EMBL Grenoble, BP 181, 71 Avenue des Martyrs, 38042 Grenoble Cedex 9, France
[8]Leicester Institute of Structural and Chemical Biology, Department of Molecular and Cell Biology, University of Leicester, Lancaster Road, Leicester LE1 7RH, UK
[9]Ben May Department of Cancer Research, The University of Chicago, Chicago, IL 60637, USA
[10]These authors contributed equally
[11]Lead Contact
*Correspondence: saadi.khochbin@univ-grenoble-alpes.fr
https://doi.org/10.1016/j.celrep.2018.08.069

SUMMARY

Nuclear protein in testis (Nut) is a universal oncogenic driver in the highly aggressive NUT midline carcinoma, whose physiological function in male germ cells has been unclear. Here we show that expression of Nut is normally restricted to post-meiotic spermatogenic cells, where its presence triggers p300-dependent genome-wide histone H4 hyperacetylation, which is essential for the completion of histone-to-protamine exchange. Accordingly, the inactivation of *Nut* induces male sterility with spermatogenesis arrest at the histone-removal stage. Nut uses p300 and/or CBP to enhance acetylation of H4 at both K5 and K8, providing binding sites for the first bromodomain of Brdt, the testis-specific member of the BET family, which subsequently mediates genome-wide histone removal. Altogether, our data reveal the detailed molecular basis of the global histone hyperacetylation wave, which occurs before the final compaction of the male genome.

INTRODUCTION

In mammals, unique physiological, genome-wide histone hyperacetylation has been observed, associated with a near-total histone eviction that occurs during the late stages of spermatogenesis. In this context, an outstanding issue is the understanding of the causes and consequences of this histone hyperacetylation and, more precisely, of its role in histone eviction (Goudarzi et al., 2014). After meiosis, haploid cells, named spermatids, engage in one of the most dramatic known chromatin remodeling and genome reorganization processes, which essentially takes place in cells known as elongating and condensing spermatids. It consists of the genome-wide removal of histones and their replacement by small basic proteins, transition proteins (TPs) and protamines (Prms) (Barral et al., 2017; Gaucher et al., 2010). Spermatids are produced following two successive meiotic divisions of spermatocytes, which are continuously generated from the progenitor cells known as spermatogonia.

During the past decades, we have developed strategies to discover the molecular basis of this genome-wide histone hyperacetylation taking place in late haploid male germ cells (Goudarzi et al., 2014). These strategies included a series of structural and functional studies of the testis-specific bromodomain-containing BET factor Brdt. These investigations showed that Brdt's first bromodomain specifically recognizes histone H4 bearing the simultaneous acetylation of K5 and K8 (Goudarzi et al., 2016; Miller et al., 2016; Morinière et al., 2009; Sasaki et al., 2009) and somehow mediates the eviction of histones (Gaucher et al., 2012).

In the frame of these studies, we also became interested in a testis-specific gene of unknown function, nuclear protein in testis (*NUT*), which is involved in a cancer known as NUT midline carcinoma (NMC). NMC is characterized by a series of chromosomal translocations that fuse the *NUT* gene located on chromosome 15q14 to partner genes, most frequently one of the double bromodomain proteins of the BET family, *BRD4* (67% of cases) or *BRD3* (25% of cases). Therefore, in more than 90% of NMC cases, NUT is fused to a double bromodomain BET factor (French, 2012).

Because of the testis-specific nature of *NUT* expression and its cooperation with somatic members of BETs, BRD3 and BRD4 in NMC, we hypothesized that a possible cooperation between NUT and BRDT in the physiological setting of spermatogenesis could occur and be involved in the establishment and management of histone hyperacetylation. Molecular investigations of the

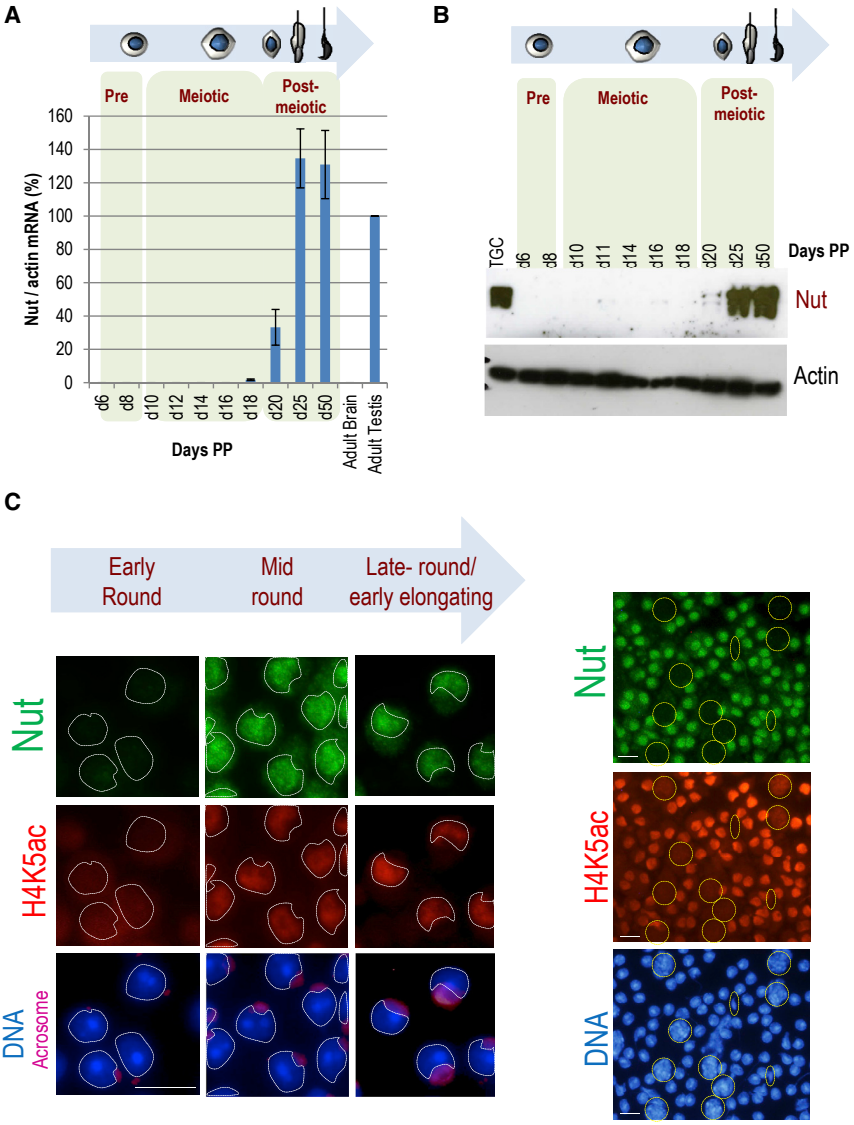


Figure 1. Nut Is Exclusively Expressed in Post-meiotic Spermatogenic Cells and Is Associated with Enhanced Histone H4 Hyperacetylation

(A) RNA was prepared from testes harvested at the indicated times after birth, and the expression of *Nut* was monitored by qRT-PCR, normalized with respect to *actin* expression and represented as the percentage of *Nut* expression in adult testis. Adult testis and adult brain RNAs were used as positive and negative controls, respectively. The histograms represent four experiments. Values represent averages ± SDs (error bars).

(B) In another set of experiments, protein extracts were prepared from testes harvested as described in (A). The corresponding immunoblot was successively probed with the anti-Nut and anti-actin antibodies.

(C) In the left panels, cells from seminiferous tubule preparations were used to co-detect in the same cells Nut (rabbit polyclonal antibodies) and H4K5ac (mouse monoclonal antibody). Lectin was used to simultaneously visualize the acrosomes (in pink). The right panels show images of Nut-expressing round spermatids mixed with spermatocytes and a few condensed spermatids (indicated) taken at lower magnification. The staining was performed as for left panels, except that lectin was not used.

Scale bars: 10 μm. See Figure S1 for tissue-specific patterns of *Nut/NUT* expression in human and mouse tissues.

oncogenic activity of the fusion protein BRD4-NUT revealed that it mediates the creation of hyperacetylated chromatin domains. In cells expressing BRD4-NUT, a feed-forward loop drives histone acetylation propagation (Reynoird et al., 2010). This BRD4-NUT-driven histone acetylation is constrained by cellular deacetylases (Reynoird et al., 2010; Schwartz et al., 2011) and remains limited to the nuclear topologically associated domains (TADs) (Alekseyenko et al., 2015), leading to the creation of hyperacetylated chromatin domains.

The discovery of the molecular basis of the activity of the fusion protein BRD4-NUT in NMC and of a role for NUT in inducing enhanced histone acetylation in NMC cancer cells prompted us to propose that Nut could also have a role in inducing the genome-wide histone hyperacetylation in post-meiotic male germ cells. To explore this hypothesis, we designed a series of experiments aiming to address the role of Nut during spermatogenesis at a molecular level.

The investigations reported here reveal mechanisms underlying the obscure process of histone H4 hyperacetylation and its molecular link to the genome-wide histone eviction. These data also demonstrate that the functional cooperation of Nut, p300 and/or CBP, and a BET factor that is created following the *BRD4-NUT* gene fusion in NMC also exists in the natural setting of Nut's action, the post-meiotic spermatogenic cells. The sequence of events triggered by this interaction becomes oncogenic when it operates "off context."

RESULTS

Nut Is a Testis-Specific Factor Exclusively Expressed in Post-Meiotic Spermatogenic Cells

The fusion partner of *BRD4*, *NUT*, was identified as a gene expressed in testis. However, until now, no information was available on its tissue-specific activity or on the pattern of its expression in testis during male germ cell differentiation.

Using publicly available transcriptomic data from different human and mouse tissues, we found that in both human and mouse, *NUT/Nut* is exclusively expressed in testis (Figure S1). We prepared RNA and protein extracts from developing mouse testes to precisely define the timing of

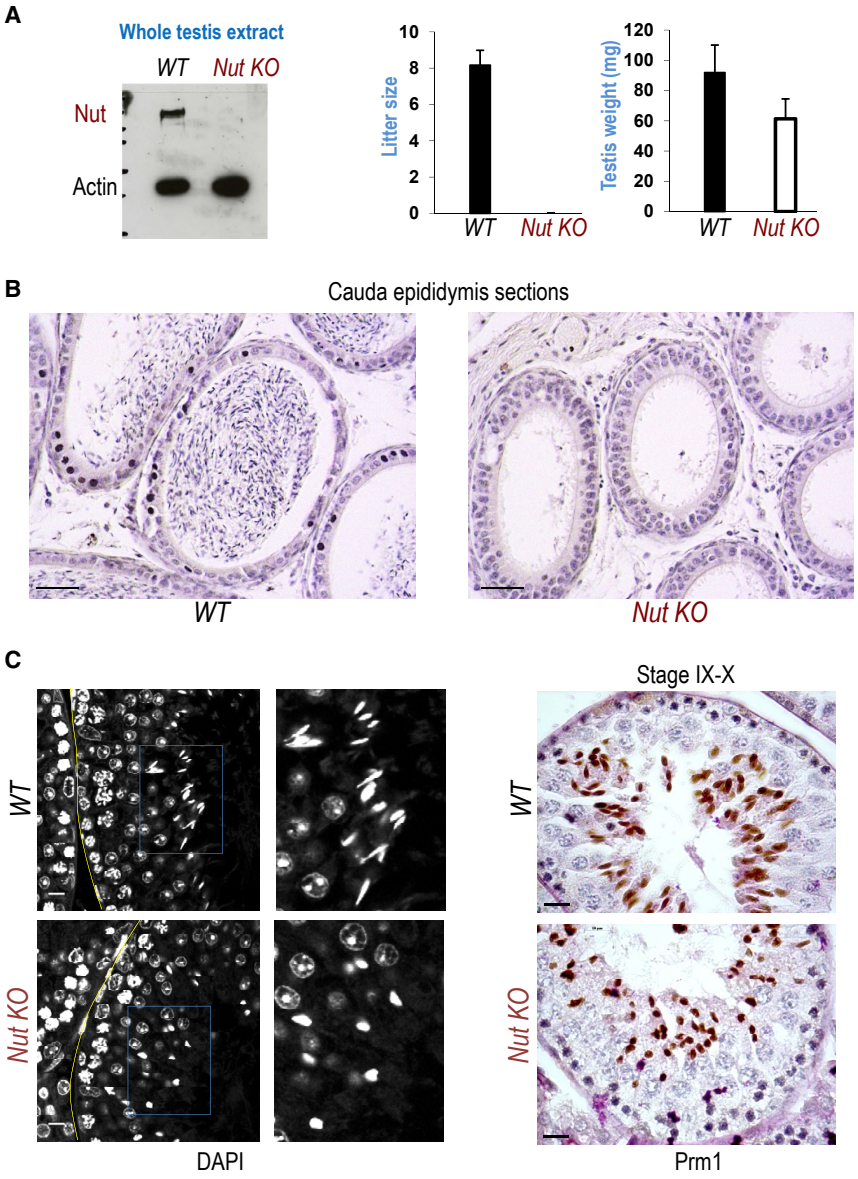


Figure 2. Inactivation of the *Nut* Gene Leads to Spermatogenesis Arrest and Male Infertility

(A) Protein extracts from wild-type and *Nut KO* mouse testes were probed with the anti-Nut and anti-actin antibodies as indicated (left panel). The right panels show litter sizes obtained after crossing wild-type female mice with wild-type (n = 12) or *Nut KO* (n = 5) male mice. The average weights of testes of wild-type (n = 16) and *Nut KO* (n = 19) mice were determined. Litter sizes and testis weights are represented as histograms. Each histogram represents averaged values ± SDs (error bars).

(B) Histological preparations of cauda epididymis from wild-type or *Nut KO* mice are shown as indicated. Scale bar: 50 μm.

(C) Histological sections of paraffin-embedded testes from wild-type and *Nut KO* mice testes at stages IX and X of spermatogenesis were stained with DAPI (left panels). The regions corresponding to the insets are shown at a higher magnification on the right. The right panels show the same types of sections as in the left panels but immunostained with an anti-Prm1 antibody and counterstained with hematoxylin. Scale bars: 10 μm.

See Figure S2 for the *Nut KO* strategy.

Nut gene expression and Nut protein accumulation during spermatogenesis.

Post-meiotic haploid round spermatids first appear at 20 days post-partum (dpp). Traces of *Nut* mRNA (Figure 1A) and protein (Figure 1B) also appear at day 20. Both *Nut* mRNA and protein strongly accumulate at 25 dpp, when late round and early elongating spermatids are produced. *In situ* co-detection of Nut, along with the acrosome on spermatogenic cell preparations, shows that Nut is not detectable in early round spermatids, as defined by the size of their acrosomes, but instead accumulates in mid-round, late round, and early elongating spermatid populations (Figure 1C, left panel).

The observed timing of Nut accumulation in round and early elongating spermatids corresponds to when histone H4 hyperacetylation is known to first occur (Goudarzi et al., 2014). This observation prompted us to co-detect Nut and H4K5ac (acetyl), a critical histone mark in post-meiotic cells (Gaucher et al., 2012; Goudarzi et al., 2016). Figure 1C (left panels) shows that the accumulation of Nut is associated with the induction of histone H4K5 acetylation in differentiating spermatids. Additional *in situ* co-detection of Nut and H4K5ac confirms this strict correlation between the accumulation of Nut and the occurrence of enhanced H4K5 acetylation. Only background H4K5ac could be detected in spermatocytes that do not express Nut compared to Nut-expressing round spermatids (Figure 1C, right panel).

The question therefore arose of whether Nut could be a direct player in the induced acetylation of histone H4 in early elongating spermatids.

Nut Is Essential for the Completion of Spermiogenesis and Mature Sperm Production

To evaluate the function of Nut in post-meiotic spermatogenic cells, we set up a strategy based on the deletion of *Nut* exon 2 to generate $Nut^{KO/KO}$ mice (*Nut* knockout [*Nut KO*] mice) (Figure S2). Western blots using our antibody on total testis extracts demonstrated the absence of Nut in spermatogenic cells of *Nut KO* mice (Figure 2A, left panel). *Nut KO* mice harbored smaller testes and were sterile (Figure 2A, middle and right panels). Cauda epididymis from *Nut KO* mice showed the absence of spermatozoa (Figure 2B), demonstrating an arrest

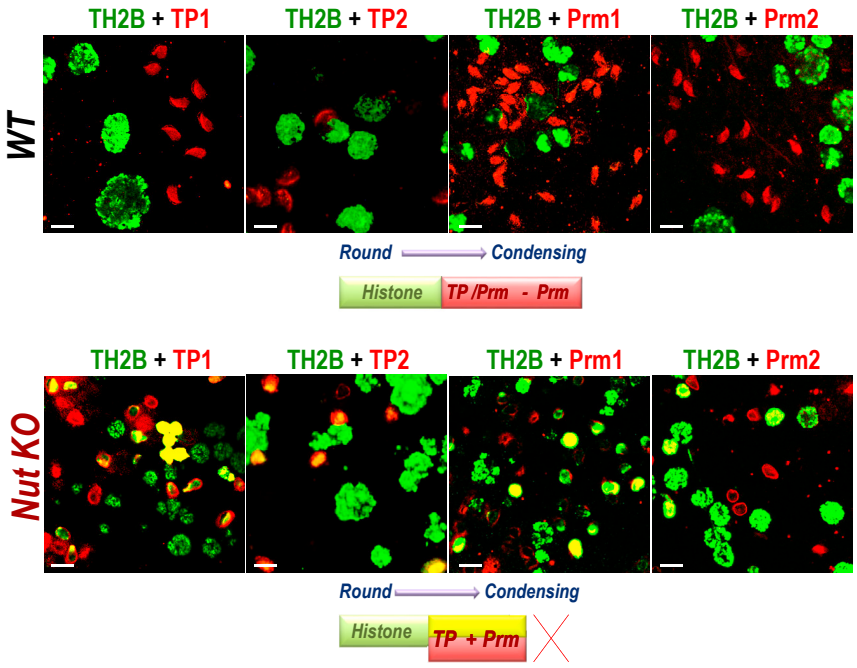


Figure 3. Nut Is Required for Histone-to-Protamine Replacement

TH2B (in green) was co-detected, along with TP1, TP2, Prm1, and Prm2 (in red) in wild-type and *Nut KO* male germ cells. See also Figure S3, showing Prm1 detection in both *Nut KO* and *Brdt ΔBD1/ΔBD1* elongating and condensing spermatids. Scale bars: 10 μm.

of spermatogenesis at earlier stages. To precisely define spermatogenesis stages affected by the absence of Nut, we prepared histological sections of paraffin-embedded testes from wild-type (*WT*) and *Nut KO* mice. In *Nut KO* mice testes, spermatogenesis seems to occur normally until the appearance of condensing spermatids. Although round spermatids could be easily detected in both wild-type and *Nut KO* testes, no condensing spermatids were found in *Nut KO* testes (Figure 2C, DAPI panels). Immunohistochemistry on these sections using an anti-Prm1 antibody showed that in the absence of Nut, the Prm1-expressing spermatids do not undergo nuclear elongation and mostly remain round (Figure 2C, Prm1 panels).

Because Nut accumulation is associated with the induction of histone H4 acetylation (Figure 1C), which is associated with the replacement of histones by TPs and protamines (Gaucher et al., 2012; Goudarzi et al., 2016), we focused on the process of histone-to-protamine replacement. Accordingly, TH2B, which is the major spermatogenic cell histone H2B variant (Montellier et al., 2013), was co-detected, along with histone-replacing proteins TP1, TP2, Prm1, and Prm2. Figure 3 shows that although TH2B was never observed in wild-type cells expressing histone-replacing proteins, in *Nut KO* spermatids, TH2B was co-detected with TP1, TP2, Prm1, and Prm2. These observations demonstrate that even though the histone-replacing proteins are expressed in the absence of Nut, they are incapable of replacing histones.

To further demonstrate the defective histone replacement in the *Nut KO* spermatids, we made the hypothesis that the accumulation of TPs and protamines in the absence of histone replacement should leave an imprint on nucleosomes, especially on the more accessible DNA linker regions. The interaction of these highly basic proteins with linker DNA should protect DNA linker regions against micrococcal nuclease (MNase) digestion and therefore generate DNA fragments longer than the canonical 147 base pairs (bp). To test this hypothesis, we prepared nuclei from fractionated round-elongating spermatids from wild-type and *Nut KO* testes and submitted them to extensive digestion with MNase to generate nucleosomal monomers. We then used paired-end sequencing to accurately determine DNA fragment lengths at the highest (base pair) resolution following two independent experiments (different mice and different fractionations). Both experiments revealed an enrichement of nucleosomal DNA fragments with lengths longer than the canonical 147 bp in *Nut KO* spermatids compared to wild-type spermatids (Figures 4A and 4B).

Altogether, the co-immunodetection of TH2B and histone-replacing proteins, as well as the accumulation of nucleosomes with longer linker DNA in *Nut KO* spermatids, strongly support an impairment of histone replacement in the absence of Nut.

The defective histone replacement observed in the absence of Nut was reminiscent of the impaired TP and protamine assembly previously detected in mouse spermatogenic cells expressing a Brdt mutant lacking its first bromodomain (BD1) (Gaucher et al., 2012) (Figure S3), suggesting a possible functional link between Nut and Brdt's first bromodomain in histone replacement.

Nut Sustains Site-Specific Histone Acetylation

Because Brdt's BD1 binds to histone H4 acetylated on K5 and K8 (Goudarzi et al., 2016; Miller et al., 2016; Morinière et al., 2009; Sasaki et al., 2009), and Brdt's BD1 is necessary for histone-to-protamine replacement (Gaucher et al., 2012), the impairment of histone removal in the absence of Nut (Figures 3 and S3) could be attributed to the absence of functional Brdt or to a defective histone acetylation specifically at H4K5 and H4K8.

A western blot on extracts from fractionated post-meiotic round and early elongating spermatids showed no change in the level of Brdt expression in the absence of Nut (Figure 5A). This result, along with our observation that Nut accumulation in mid- and late round and early elongating spermatids corresponds to the induction of histone hyperacetylation (Figure 1C), prompted us to perform a comparative analysis of the level of histone acetylation on histones extracted from round and early elongating spermatids from wild-type and *Nut KO* testes.

Consequently, we sought an unbiased approach to quantitatively monitor histone acetylation in the wild-type and *Nut KO*

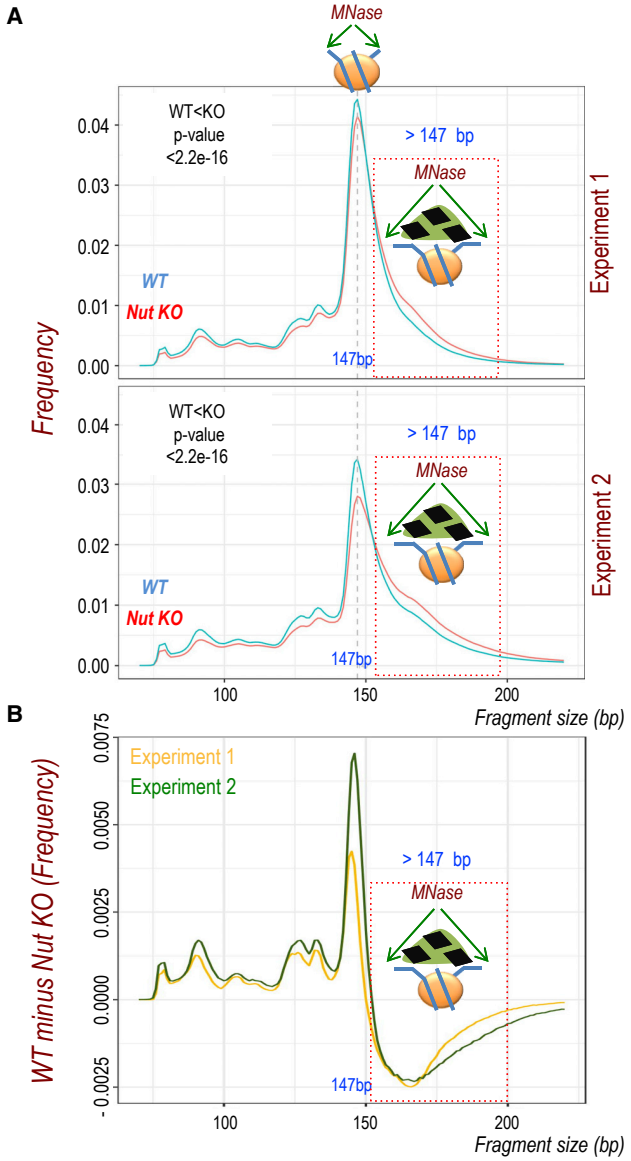


Figure 4. Genome-wide High-Resolution Studies Show Chromatin Changes in *Nut KO* Spermatids

(A) Paired-end sequencing was used for the high-resolution determination of nucleosomal DNA fragment lengths protected against extensive MNase digestion of nuclei prepared from fractionated spermatids from *wild-type* and *Nut KO* testes. Standard MNase fragment size distributions (mode at 147 bp) are observed for the two independent experiments. A more prominent shift toward DNA lengths > 147 bp is observed in *Nut KO* spermatids. The length shift significance was statistically assessed using a one-sided Wilcoxon rank-sum test with continuity correction.

(B) Comparison by subtraction of *wild-type* versus *Nut KO* distribution frequency of fragment size, highlighting the enrichment of standard nucleosomes and sub-nucleosomal structures in *wild-type* spermatids and the enrichment of longer nucleosomal DNAs in the chromatin of *Nut KO* spermatids. The results of two independent experiments are shown. For each experiment, round spermatid fractions were obtained from 3 to 5 mice (6 to 10 testes) and nuclei were prepared and extensively digested by MNase to reduce chromatin to nucleosome monomers. The nucleosomal DNA obtained from each experiment was then purified and used for paired-end sequencing.

round and early elongating spermatids. *In vitro* isotopic labeling followed by high-performance liquid chromatography-tandem mass spectrometry (HPLC-MS/MS) analysis was performed to quantify histone post-translational modifications (PTMs). Histone extracts were digested with trypsin, and the resulting peptides were propionylated using light ($^{12}C_6$) and heavy ($^{13}C_6$) propionic anhydride. The light and heavy propionylated tryptic peptides were mixed in equal amounts and analyzed by HPLC-MS/MS.

The quantification of histone acetylation revealed a decrease in histone H4 and H2A acetylation, with a remarkable downregulation of acetylation at H4K5 and H4K8, as well as of H2AK5 and H2AK9 (Figure 5A, lower panel; Tables S1 and S2).

The downregulation of acetylation, especially at the critical H4K5 and H4K8 sites, in *Nut KO* round and early elongating spermatids compared to the wild-type spermatids was confirmed by western blotting, with the corresponding antibodies on independently prepared round and early elongating spermatid-enriched fractions (Figure 5B). Because none of the H3 tail lysines analyzed here by mass spectrometry were significantly affected by the absence of Nut, H3K9 acetylation was additionally checked by immunoblotting, showing only a slight decrease in *Nut KO* spermatids (Figure 5B).

These results show that the expression of *Nut* in round and early elongating spermatids (Figure 1C) is required for histone H4 hyperacetylation. In addition, the impairment of Nut-mediated acetylation of H4 demonstrates the inability of Brdt's BD1 to bind chromatin, explaining the similarity between *Nut KO* and *Brdt*$^{delta\ BD1/delta\ BD1}$ phenotypes (Figure S3).

Nut Controls the Expression of a Subset of Genes Expressed in Spermatids

Our previous work on NMC identified p300 and/or CBP as the major cellular histone acetyltransferases (HATs) recruited by the NUT moiety of the BRD4-NUT fusion protein in a NMC cell line or upon transfection (Reynoird et al., 2010). We therefore hypothesized that when expressed in its physiological setting, spermatids, Nut could interact with p300 and/or CBP and use a similar mechanism to enhance histone acetylation. To verify this hypothesis, we performed two independent sets of experiments. First, the round and early elongating spermatid fractions from wild-type and *Nut KO* testes were used to obtain and compare the corresponding transcriptomes and to identify genes whose expression depends on Nut. The absence of Nut was found to be associated with the downregulation of 1,184 genes and the upregulation of 499 genes (using a cutoff of 1.5 for the absolute log2 fold change value and a t test p value, after adjustment for multiple tests by Benjamini-Hochberg, of $p < 0.05$) (Figure 6A).

To ensure the observed changes in gene expression in the absence of Nut were not due to changes in the cell-type composition of the *Nut KO* round and elongating cell fraction compared to the same fraction of wild-type cells, we performed an additional analysis. By comparing wild-type meiotic and post-meiotic cell populations, we identified genes that are normally activated in wild-type post-meiotic cells. The observation that only a fraction of these genes was downregulated in *Nut KO* post-meiotic cells was supportive of a specific action of Nut on gene

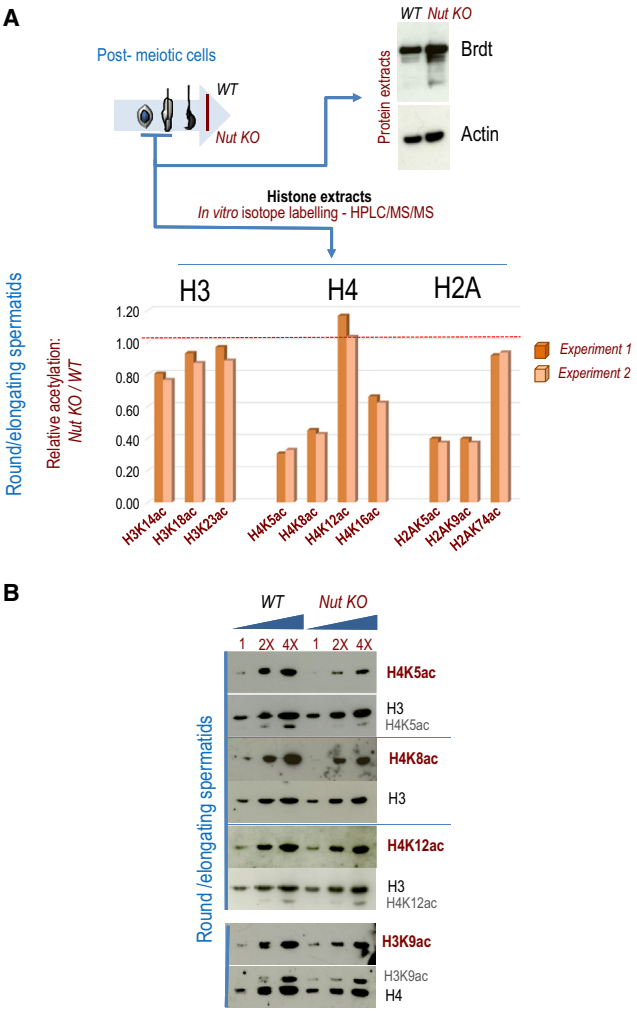


Figure 5. Nut Sustains Histone H4 Hyperacetylation in Spermatids

(A) Spermatogenic cell suspensions from wild-type and *Nut KO* testes were fractionated to prepare cell suspensions enriched in round and elongating spermatids. In one experiment, the corresponding extracts were used to detect Brdt and actin as indicated. In another experiment, histone extracts from these fractions (wild-type and *Nut KO*) were analyzed by *in vitro* isotopic labeling followed by HPLC-MS/MS. The relative abundance of identified site-specific acetylation was determined and expressed as the ratio *Nut KO*/wild-type. H3, H4, H2B, and H2A peptides were used to normalize the *Nut KO*/wild-type samples. For more information on the normalization process, refer to Table S2 and its legend and to Quantification of Histone PTMs.

(B) Independent cell fractionation experiments were set up to prepare extracts from round and elongating spermatid-enriched fractions. Increasing amounts of extracts (1×, 2×, and 4×) were used for immunoblots as indicated. H4K5ac, H4K8ac, and H4K12ac were visualized with the corresponding specific antibodies in wild-type and *Nut KO* cell extracts as indicated. Each of these blots was re-probed with anti-H3 antibody as indicated. H3K9ac was also detected (lowest panel) and the corresponding blot was re-probed with anti-H4 antibody (the remaining H4K5ac, H4K12ac, and H3K9ac signals after anti-H3 and anti-H4 re-probing are indicated in gray letters).

expression as opposed to the loss of particular cell types. The expression of a significant number of the genes, 317, including the Prm1-encoding gene, that are normally activated in wild-type post-meiotic cells did not significantly change in the absence of Nut (Figures S4A and S4B), demonstrating that our transcriptomic data from wild-type and *Nut KO* spermatids could be compared, because they were obtained from two comparable cell populations.

We then used a series of stage-specific transcriptomes that were produced in wild-type male germ cells from our previous studies (Boussouar et al., 2014; Montellier et al., 2013) and the present work to establish the normal pattern of these Nut-regulated genes during spermatogenesis. Figure S4C shows that the genes that require Nut to be active (downregulated genes in *Nut KO*) are normally activated in post-meiotic cells, with the highest activity in late round and early elongating spermatids.

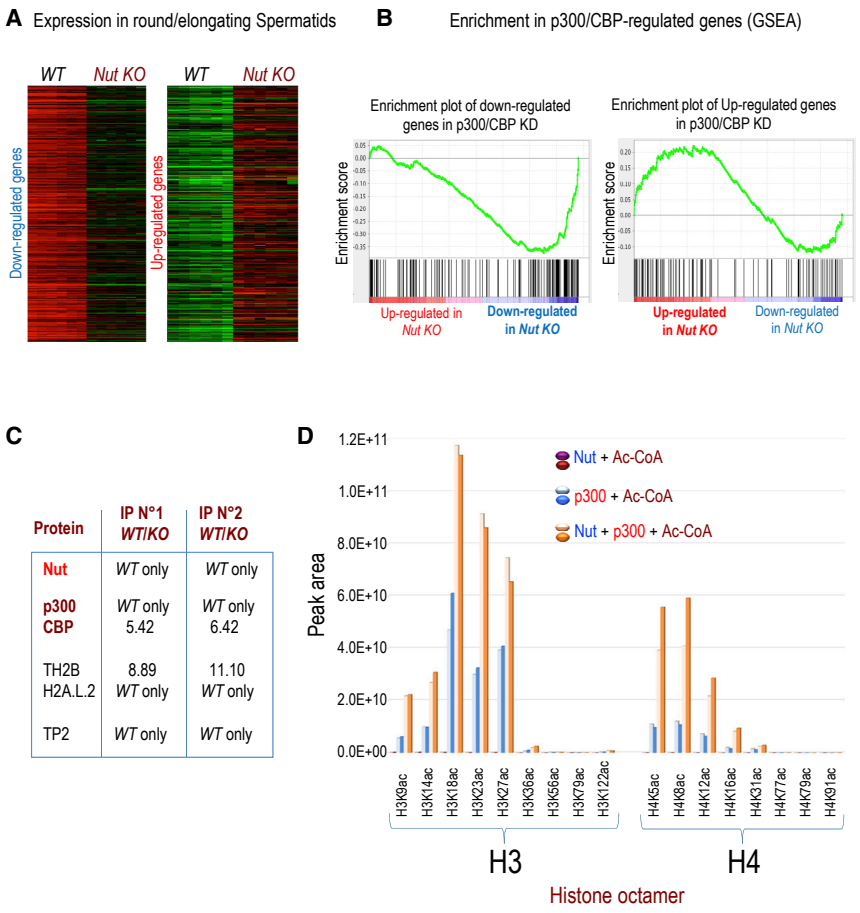
In a previous work, we identified genes that are specifically responsive to post-meiotic p300 and/or CBP variations in spermatids (Boussouar et al., 2014). We used these data to define whether Nut regulates gene expression through p300 and/or CBP by comparing our two Nut- and p300 and/or CBP-dependent transcriptomes, both obtained in comparable post-meiotic cell populations. We looked for a significant enrichment of Nut-regulated genes in the lists of p300 and/or CBP-regulated genes, which would suggest the involvement of p300 and/or CBP in gene regulation by Nut. Gene set enrichment analysis (GSEA) was used to look for the expression in *Nut KO* spermatids of genes whose expression in spermatids had previously been found to be affected by a targeted post-meiotic deletion of *p300* and *CBP* (Boussouar et al., 2014). Although following this latter approach, the inheritance of both p300 and CBP from preceding stages led to only a small decrease in the enzyme levels in post-meiotic cells, it was enough to observe a change in the expression of a specific set of genes whose expression is highly dependent on, and sensitive to, p300 and CBP dosage (Boussouar et al., 2014).

Figure 6B shows a clear enrichment of the p300 and/or CBP-regulated genes among the genes up- and downregulated in *Nut KO* mice, supporting the idea that Nut-p300 and/or CBP, while directing the global histone acetylation, also affects the expression of late active genes before the general shutdown of transcription in elongating spermatids.

p300 and/or CBP Are the Only Acetyltransferases Present in the Nut Interactome

The question arises of whether Nut uses CBP and p300 in spermatids as it does in NMC cells or if other HATs are also involved. To answer this question, in an additional set of experiments, we prepared extracts from fractionated wild-type and *Nut KO* round and early elongating spermatids, which were submitted to anti-Nut immunoprecipitation and a proteomic analysis in two independent assays.

The aim of this experiment was two-fold: first, to confirm the presence of p300 and CBP in association with Nut, and second, to see whether additional cellular acetyltransferases interact with Nut in post-meiotic cells undergoing histone hyperacetylation. To maximize the chance of detecting any HAT associated with Nut, we used non-stringent immunoprecipitation conditions and performed the experiment twice on independent mice and fractionated cells. As expected from non-stringent immunoprecipitation conditions, many proteins were identified as



Protein	IP N°1 WT/KO	IP N°2 WT/KO
Nut	*WT* only	*WT* only
p300 CBP	*WT* only 5.42	*WT* only 6.42
TH2B H2A.L.2	8.89 *WT* only	11.10 *WT* only
TP2	*WT* only	*WT* only

Figure 6. Nut Is Involved in the Control of Gene Expression by p300 and/or CBP and Uses p300 to Acetylate Histones

(A) Round and elongating spermatid fractions were obtained as in Figures 5A and 5B, but this time RNA was purified and used to compare the transcriptomes of *Nut KO* cells and wild-type cells. The heatmaps show genes that are down-regulated (left panel, n = 1,184) and upregulated (right panel, n = 499) in *Nut KO* compared to wild-type round and elongating spermatids. See Figure S4 and its legend.

(B) Gene set enrichment analysis (GSEA) (http://software.broadinstitute.org/gsea/) in the *Nut KO* versus wild-type transcriptomic experiment (obtained in A) was performed for two lists of genes (used as gene sets) previously established as downregulated or upregulated in round and elongating spermatids after the conditional post-meiotic knockout of *p300* and *CBP* (Boussouar et al., 2014). Genes that are down- or upregulated after double *p300* and/or *CBP KO* are significantly enriched in genes that are, respectively, down- or upregulated in *Nut KO* spermatids.

(C) Spermatogenic cell fractions identical to those shown in (A) and Figure 5 were used to perform the immunoprecipitation of Nut. Mass spectrometry was then used to identify Nut-associated proteins, meaning proteins present or significantly enriched in the wild-type cell extracts and absent from *Nut KO* cell extracts. The wild-type/*KO* ratio of proteins corresponding to the only identified HATs (p300 and CBP), as well as proteins known to be involved in histone-to-protamine replacement, is shown. The indicated values represent protein enrichment in wild-type cell immunoprecipitations over *Nut KO*. The list of identified proteins functionally linked to chromatin and/or transcription and/or RNA and/or metabolism is presented in Table S3.

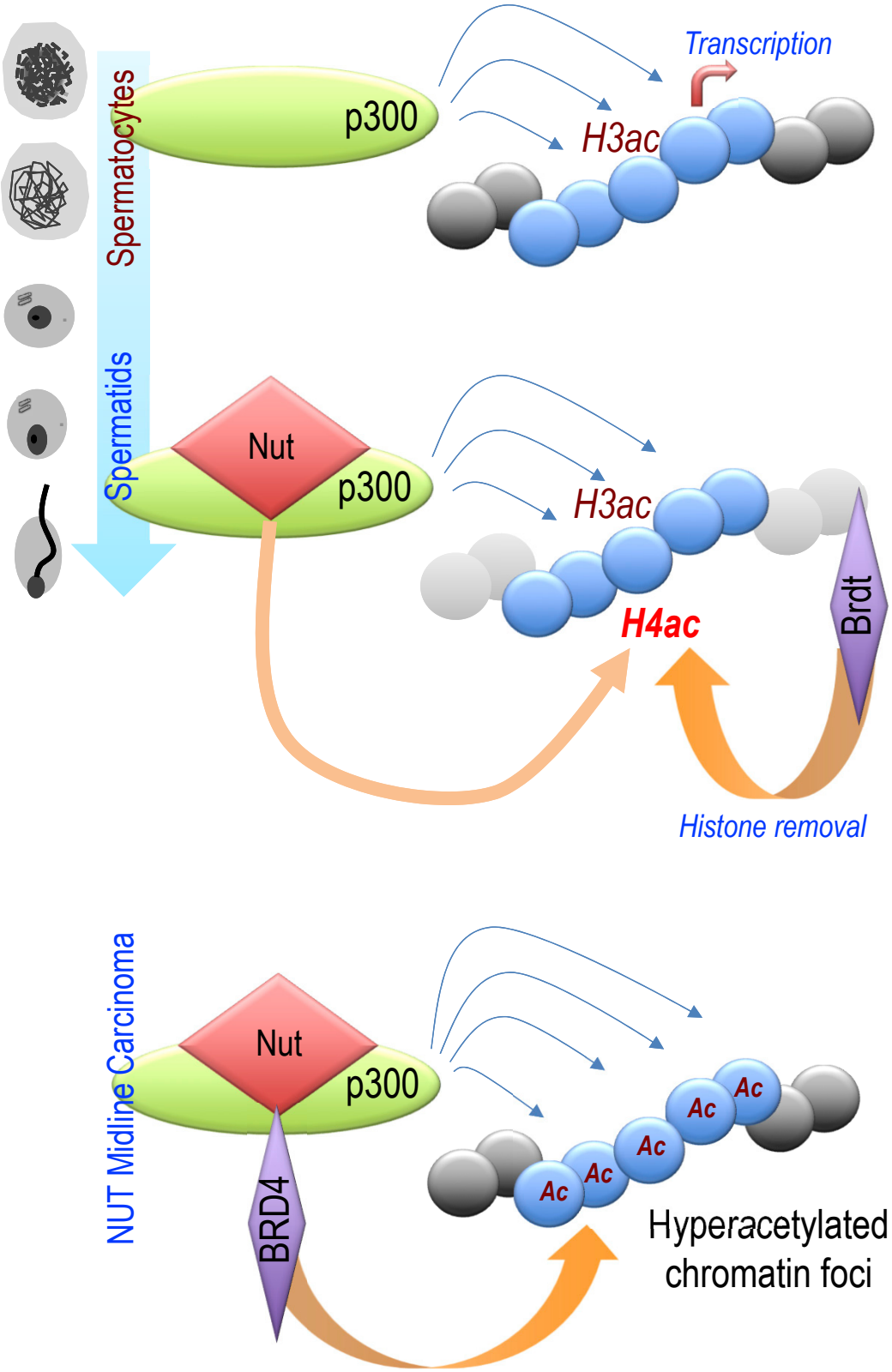
(D) Reconstituted histone octamers were incubated with acetyl-CoA, as well as purified p300 (residues 340–2,094: $p300_{short}$) and the p300-interacting domain of NUT as indicated. HPLC-MS/MS was used to quantify site-specific H3 and H4 acetylation by label-free quantification *in vitro*. Five high-abundance unmodified peptides were used to normalize the amount of each histone. The peak area of histone octamer was then normalized, and the data corresponding to two experiments are shown (with different color codes as indicated). See related Figure S5.

specifically enriched in *Nut* wild-type cells and were observed in both experiments. However, in an effort to focus on proteins relevant to the present study, especially HATs, we chose to highlight the transcription and/or chromatin and/or RNA and/or metabolism-related proteins. Table S3 lists the proteins that were only found in the wild-type cell extracts or enriched at least 5 times in these extracts compared to *Nut KO* cell extracts. Despite the non-stringent immunoprecipitation conditions, p300 and CBP were the only known HATs found to be exclusively or significantly enriched in the Nut immunoprecipitates in wild-type extracts compared to *Nut KO* extracts (Figure 6C). In the Nut-dependent proteome, we also found TH2B-H2A.L.2 and transition protein 2 (TP2), shown to allow protamines to efficiently invade nucleosomes at the time of histone-to-protamine replacement (Barral et al., 2017) (Figure 6C). Because the generation of these nucleosomal transitional states also involves H4 acetylation and Brdt's BD1, these findings reinforce the hypothesis that Nut could also be a player in the process of acetylation-dependent histone replacement.

A proteomic analysis of BRD4-NUT fusion partners from a human cell line generated similar data, with p300 and CBP as the only BRD4-NUT-associated HATs (Alekseyenko et al., 2017), confirming our early finding that p300 and CBP are the major HAT partners of NUT (Reynoird et al., 2010). Altogether, these data are in agreement with our comparative transcriptomic analysis (Figure 6B), pointing to p300 and/or CBP as the only cellular HAT partners of Nut.

Nut Enhances p300-Induced H3 and H4 Acetylation

We previously demonstrated that p300 interacts with a specific fragment of NUT, which stimulates its HAT activity on purified histone H3 (Reynoird et al., 2010). Here, to have a comprehensive vision of the action of Nut on histone acetylation, and more specifically to evaluate the action of NUT on site-specific histone acetylation by p300, we incubated purified recombinant p300 with the p300-interacting fragment of NUT and histone octamers (Figure S5). Isotopic labeling following HPLC-MS/MS was used to quantitatively measure the impact of NUT on



(legend on next page)

p300-mediated histone H3 and H4 acetylation. In histone octamers, p300 more efficiently acetylated histone H3 compared to histone H4 (p300 + Ac-coenzyme A [CoA]) (Figure 6D). For both histones, acetylation occurred at the tails, with no significant acetylation of the internal lysines (Figure 6D). The addition of the p300-interacting fragment of NUT enhanced H3 and H4 tail acetylation (Figure 6D; Tables S4 and S5; Figure S5). These data were confirmed after independent HAT assays following immunodetection with two site-specific antibodies against H3K9ac and H4K5ac (Figures S5B and S5C).

Altogether, these data show that Nut directly recruits p300 and/or CBP to control both histone H3 and H4 acetylation, although *in vivo*, in early elongating spermatids, the most important impact is observed on H4 acetylation. In contrast to the action of Nut on H3 and H4 acetylation observed *in vitro* (Figure 6D), although the absence of Nut affected H4 acetylation in round and elongating spermatids, none of the considered H3 acetyl acceptor sites seemed to be significantly affected by the absence of Nut (Figure 5).

DISCUSSION

Quantitative measurements of Nut-dependent histone acetylation in round and early elongating spermatids reported here demonstrated that Nut has a histone-specific and site-specific action on chromatin acetylation, particularly targeted on H2A and H4, including H4K5 and H4K8 acetylation.

A wave of histone hyperacetylation had long been described in elongating spermatids. In this context, H4 acetylation seems to play a central role, because it provides a specific binding site for the essential factor involved in histone removal, Brdt (Goudarzi et al., 2014, 2016).

All investigations reported here point to p300 and/or CBP as major HATs, which in addition to a role in maintaining a basal acetylation of H3, become involved in H4 hyperacetylation in post-meiotic cells because of the expression of Nut. However, an important question was whether Nut could control the activity of other HATs in post-meiotic spermatogenic cells to direct the observed H4 hyperacetylation. Our Nut interactomic analysis shows that p300 and CBP are the only cellular Nut-interacting HATs. In an immunopurification of BRD4-NUT followed by mass spectrometry identification of the associated proteins, p300 and CBP were the only HATs found in a complex with BRD4-NUT in somatic cells (Alekseyenko et al., 2017), which highlights the conserved mode of action of Nut both in spermatogenic cells and after its ectopic expression in NMC cancer cells.

Although the action of Nut on H4 acetylation seems to be dramatic, its transcriptional regulatory role only concerns a subset of genes expressed in transcriptionally active late spermatids. Our data show that in round and early elongating spermatids, most Nut-regulated genes are those previously found to be highly sensitive to the cellular levels of p300 and CBP in elongating spermatids before the transcriptional shutdown (Boussouar et al., 2014). It is therefore possible that the enhancement of H4 acetylation by Nut-p300 and/or CBP, in addition to the basal role of these HATs in H3 acetylation (Figure 6) in elongating spermatids, in preparation for histone replacement, stimulates a subset of genes that are normally active immediately before the general transcriptional inactivation.

These data therefore depict Nut's function in post-meiotic germ cells. Nut appears as a stimulator of p300 and CBP, which, in addition to H3 acetylation that normally occurs in spermatogonia and spermatocytes, drives cell-type-specific H4 hyperacetylation in post-meiotic cells in preparation for histone eviction. This gene is exclusively expressed in cells in which global H4 hyperacetylation naturally occurs, strongly supporting this conclusion. In agreement with a role of histone hyperacetylation in histone removal and a role of Nut in H4 hyperacetylation, we observed that in the absence of Nut, the acetylation of histone H4 is downregulated and histone removal is impaired.

The similarity of phenotypes between *Nut* KO and *Brdt*$^{deltaBD1/deltaBD1}$ mouse models also supports a role of Nut in histone acetylation that occurs upstream of histone removal. In both cases, although TPs and protamines are produced, they are unable to replace histones. This observation strongly suggests that Nut and Brdt function in the same molecular pathway, leading to histone-to-protamine replacement. The molecular mechanism linking Nut activity to histone acetylation, specifically that of H4K5 and H4K8, explains how Nut and Brdt are functionally related. In *Nut KO*-elongating spermatids, because of insufficient acetylation of H4, Brdt's BD1 remains non-functional, leading to the same phenotype as observed in *Brdt*$^{deltaBD1/deltaBD1}$ cells.

These studies also place Nut at the heart of the molecular mechanisms controlling histone-to-protamine replacement. Nut could link histone hyperacetylation to histone exchange and nucleosome invasion by protamines (Barral et al., 2017). Our Nut-interactomic approach also identified TH2B, H2A.L.2, and TP2 as Nut partners. Therefore, p300 could be part of a machinery that would couple histone H4 acetylation, Brdt binding, and the exchange of TH2B-H2A for TH2B-H2A.L.2 in a series of events that are essential for nucleosome opening and TP-controlled protamine invasion of nucleosomes and histone replacement (Barral et al., 2017). In agreement with this hypothesis, published data have reported the ability of p300 to cooperate with Nap1 to mediate nucleosome disassembly (Asahara et al., 2002; Luebben et al., 2010; Sharma and Nyborg, 2008).

Therefore, at the time of p300-Nut interaction and histone H4 hyperacetylation, nucleosomes adopt an increasingly open configuration and are largely remodeled. It is possible that under these conditions, histone accessibility to HATs such as p300 becomes similar to that of histones in octamers rather than histones

Figure 7. Model for the Cooperation between Nut-p300 and/or CBP-BET in Physiological and Pathological Settings

Along spermatogenesis, the ubiquitous HATs p300 and/or CBP drive transcription, mainly through histone H3 acetylation. Upon its expression in late round and elongating spermatids, Nut interacts with p300 and/or CBP and stimulates its HAT activity, leading to a genome-wide increase in H4 acetylation, specifically at H4K5 and H4K8. Brdt uses its first bromodomain to recognize the two indicated sites on H4 to bind chromatin and to eventually replace histones (upper panel). In NMC cells, where a chromosomal translocation fuses *BRD4* to *NUT*, p300 recruitment by the fusion protein creates an obligatory functional loop involving cycles of NUT-stimulated histone acetylation and BRD4-NUT binding (lower panel).

in regular nucleosomes. In our *in vitro* HAT assays, histone octamers were used instead of nucleosomes. In our hands, nucleosomes seemed to be poor substrates for p300. However, in spermatids *in vivo*, in contrast to our *in vitro* conditions, Nut seems to be indispensable for H4 acetylation, whereas its absence seems to have only minor effects on H3 acetylation. It is possible that *in vivo*, in spermatids, H3 acetylation is inherited from earlier stages and that only H4 acetylation occurs *de novo* at the time of Nut synthesis. This hypothesis is in good agreement with our previous observations that in post-meiotic cells, CBP and p300 are mostly inherited from earlier stages (Boussouar et al., 2014). This possibility could therefore explain why *in vivo* Nut only affects H4 acetylation, which occurs *de novo* in these cells, while most H3 acetylation comes from the preceding stages (Figure 7).

Hence, our data support the hypothesis that Nut is a factor required by CBP and p300 to include H4 as a substrate *in vivo* and ensure H4 hyperacetylation in spermatids. In contrast, its impact on the level of H3 acetylation *in vivo* is relatively modest, because these enzymes already efficiently acetylate H3 in male germ cells at earlier stages of spermatogenesis (Figure 7).

In summary, we found that Nut's major function is to direct histone H4 hyperacetylation, leading to the cooperative action of p300, CBP, and the BET factor Brdt in haploid male germ cells. In NMC cells, a chromosomal translocation induces NUT-mediated forced cooperation between p300 and/or CBP and BRD4 to induce hyperacetylated histone chromatin foci (Figure 7). An extension of this work would be to explore the similarities between NMC cells and elongating spermatids. Rescue experiments with *Nut* re-expression under the control of different promoters, active at earlier stages of spermatogenesis, could be performed in the context of mice bearing *Nut KO* alone or both *Nut KO* and *Brdt*$^{deltaBD1/deltaBD1}$, which would demonstrate the cooperation between Nut and BD1 in the process of histone hyperacetylation.

This work also highlights a concept that cells can use particular auxiliary factors in a tissue- and cell-type-specific manner to enhance the activity of chromatin-modifying enzymes that are ubiquitously expressed and thus achieve a large-scale, tissue-specific chromatin modification such as H4 hyperacetylation in post-meiotic spermatogenic cells. The ectopic activation of such factors could initiate an oncogenic molecular circuit as it does in NMC cells expressing the BRD4-NUT fusion protein.

STAR★METHODS

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

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures and five tables and can be found with this article online at https://doi.org/10.1016/j.celrep.2018.08.069.

ACKNOWLEDGMENTS

The S.K. laboratory is supported by a grant from the Foundation pour la Recherche Medicale (FRM), "analyse bio-informatique pour la recherche en biologie" program, as well as by the ANR Episperm3 program and by INCa libre program (RPT13001CCA). Additional support came from the Université Grenoble Alpes (ANR-15-IDEX-02) LIFE and SYMER programs, as well as from Fondation ARC "Canc'air" project (RAC16042CLA) and Plan Cancer (CH7-INS15B66 and ASC16012CSA). Z.I. and D.P. were supported by a Worldwide Cancer Research foundation grant (16-0280) to D.P. H.S. was a recipient of a Marie Curie Initial Training Network funded by European Commission (FP7-PEOPLE-2011-ITN and PITN-GA-289880) for 3 years and then supported by La Ligue Nationale Contre Le Cancer for a 4th year of PhD pursuit. G.C. is a bioinformatic expert benefiting from a FRM fellowship. Mice were bred in the High Technology Animal Facility (PHTA) of Grenoble University with the help of Natacha Pitoun and Kevin Escot. The TGML Platform is supported by grants from Inserm, GIS IBiSA, and Aix-Marseille Université (ANR-10-INBS-0009-10). We are grateful to Hortense Vachon for her help with library preparation and sequencing. The proteomic analyses were partially supported by the French National Research Agency ProFi grant (ANR-10-INBS-08-01). P.F. and S.P. were supported by a Wellcome Trust Cancer Development Fellowship (095751/Z/11/Z). Y.C. thanks the support of the discovery platform and informatics group at EDyP. Most computations presented in this paper were performed using the GRICAD infrastructure (https://gricad.univ-grenoble-alpes.fr).

AUTHOR CONTRIBUTIONS

H.S., S.B., and T.B. performed most experiments. M.T., M.Z., and Y.Z. performed quantitative measurements of *in vivo* and *in vitro* histone acetylation. F.B., T.B., and N.R. designed the *Nut KO* strategy and prepared the necessary materials and controlled the various steps toward mouse production, which was carried out in the M.G. laboratory. Y.C. performed the characterization of the Nut interactome. A.G., Z.I., and D. Panne performed recombinant p300 purification and *in vitro* HAT assays. S.P. and P.F. did Nut fragment purification for anti-Nut antibody production. L.B. constructed and sequenced the library for micrococcal nuclease sequencing (MNase-seq) experiments. G.C., D. Puthier, F.C., E.B.-F., and S.R. performed the bioinformatics analyses. S.K. conceived and coordinated the project and wrote the manuscript. All authors discussed the data and read and commented on the manuscript.

DECLARATION OF INTERESTS

Y.Z. is a shareholder and a member of the scientific advisory board of PTM BioLabs, Co. (Chicago). The other authors declare no competing interests.

Received: April 20, 2018
Revised: July 19, 2018
Accepted: August 23, 2018
Published: September 25, 2018

REFERENCES

Alekseyenko, A.A., Walsh, E.M., Wang, X., Grayson, A.R., Hsi, P.T., Kharchenko, P.V., Kuroda, M.I., and French, C.A. (2015). The oncogenic BRD4-NUT chromatin regulator drives aberrant transcription within large topological domains. Genes Dev. *29*, 1507–1523.

Alekseyenko, A.A., Walsh, E.M., Zee, B.M., Pakozdi, T., Hsi, P., Lemieux, M.E., Dal Cin, P., Ince, T.A., Kharchenko, P.V., Kuroda, M.I., and French, C.A. (2017). Ectopic protein interactions within BRD4-chromatin complexes drive oncogenic megadomain formation in NUT midline carcinoma. Proc. Natl. Acad. Sci. USA *114*, E4184–E4192.

Asahara, H., Tartare-Deckert, S., Nakagawa, T., Ikehara, T., Hirose, F., Hunter, T., Ito, T., and Montminy, M. (2002). Dual roles of p300 in chromatin assembly and transcriptional activation in cooperation with nucleosome assembly protein 1 *in vitro*. Mol. Cell. Biol. *22*, 2974–2983.

Barral, S., Morozumi, Y., Tanaka, H., Montellier, E., Govin, J., de Dieuleveult, M., Charbonnier, G., Couté, Y., Puthier, D., Buchou, T., et al. (2017). Histone variant H2A.L.2 guides transition protein-dependent protamine assembly in male germ cells. Mol. Cell *66*, 89–101.

Boussouar, F., Goudarzi, A., Buchou, T., Shiota, H., Barral, S., Debernardi, A., Guardiola, P., Brindle, P., Martinez, G., Arnoult, C., et al. (2014). A specific CBP/p300-dependent gene expression programme drives the metabolic remodelling in late stages of spermatogenesis. Andrology *2*, 351–359.

Buchou, T., Tan, M., Barral, S., Vitte, A.-L., Rousseaux, S., Arechaga, J., and Khochbin, S. (2017). Purification and analysis of male germ cells from adult mouse testis. Methods Mol. Biol. *1510*, 159–168.

Casabona, M.G., Vandenbrouck, Y., Attree, I., and Couté, Y. (2013). Proteomic characterization of Pseudomonas aeruginosa PAO1 inner membrane. Proteomics *13*, 2419–2423.

Chen, Y., Kwon, S.W., Kim, S.C., and Zhao, Y. (2005). Integrated approach for manual evaluation of peptides identified by searching protein sequence databases with tandem mass spectra. J. Proteome Res. *4*, 998–1005.

French, C.A. (2012). Pathogenesis of NUT midline carcinoma. Annu. Rev. Pathol. *7*, 247–265.

Garcia, B.A., Mollah, S., Ueberheide, B.M., Busby, S.A., Muratore, T.L., Shabanowitz, J., and Hunt, D.F. (2007). Chemical derivatization of histones for facilitated analysis by mass spectrometry. Nat. Protoc. *2*, 933–938.

Gaucher, J., Reynoird, N., Montellier, E., Boussouar, F., Rousseaux, S., and Khochbin, S. (2010). From meiosis to postmeiotic events: the secrets of histone disappearance. FEBS J. *277*, 599–604.

Gaucher, J., Boussouar, F., Montellier, E., Curtet, S., Buchou, T., Bertrand, S., Hery, P., Jounier, S., Depaux, A., Vitte, A.-L., et al. (2012). Bromodomain-dependent stage-specific male genome programming by Brdt. EMBO J. *31*, 3809–3820.

Goudarzi, A., Shiota, H., Rousseaux, S., and Khochbin, S. (2014). Genome-scale acetylation-dependent histone eviction during spermatogenesis. J. Mol. Biol. *426*, 3342–3349.

Goudarzi, A., Zhang, D., Huang, H., Barral, S., Kwon, O.K., Qi, S., Tang, Z., Buchou, T., Vitte, A.-L., He, T., et al. (2016). Dynamic competing histone H4 K5K8 acetylation and butyrylation are hallmarks of highly active gene promoters. Mol. Cell *62*, 169–180.

Liu, P., Jenkins, N.A., and Copeland, N.G. (2003). A highly efficient recombineering-based method for generating conditional knockout mutations. Genome Res. *13*, 476–484.

Luebben, W.R., Sharma, N., and Nyborg, J.K. (2010). Nucleosome eviction and activated transcription require p300 acetylation of histone H3 lysine 14. Proc. Natl. Acad. Sci. USA *107*, 19254–19259.

Luger, K., Rechsteiner, T.J., and Richmond, T.J. (1999). Expression and purification of recombinant histones and nucleosome reconstitution. Methods Mol. Biol. *119*, 1–16.

Miller, T.C.R., Simon, B., Rybin, V., Grötsch, H., Curtet, S., Khochbin, S., Carlomagno, T., and Müller, C.W. (2016). A bromodomain-DNA interaction facilitates acetylation-dependent bivalent nucleosome recognition by the BET protein BRDT. Nat. Commun. *7*, 13855.

Montellier, E., Boussouar, F., Rousseaux, S., Zhang, K., Buchou, T., Fenaille, F., Shiota, H., Debernardi, A., Héry, P., Curtet, S., et al. (2013). Chromatin-to-nucleoprotamine transition is controlled by the histone H2B variant TH2B. Genes Dev. *27*, 1680–1692.

Morinière, J., Rousseaux, S., Steuerwald, U., Soler-López, M., Curtet, S., Vitte, A.-L., Govin, J., Gaucher, J., Sadoul, K., Hart, D.J., et al. (2009). Cooperative binding of two acetylation marks on a histone tail by a single bromodomain. Nature *461*, 664–668.

Panne, D., McWhirter, S.M., Maniatis, T., and Harrison, S.C. (2007). Interferon regulatory factor 3 is regulated by a dual phosphorylation-dependent switch. J. Biol. Chem. *282*, 22816–22822.

Reynoird, N., Schwartz, B.E., Delvecchio, M., Sadoul, K., Meyers, D., Mukherjee, C., Caron, C., Kimura, H., Rousseaux, S., Cole, P.A., et al. (2010). Oncogenesis by sequestration of CBP/p300 in transcriptionally inactive hyperacetylated chromatin domains. EMBO J. *29*, 2943–2952.

Sasaki, K., Ito, T., Nishino, N., Khochbin, S., and Yoshida, M. (2009). Real-time imaging of histone H4 hyperacetylation in living cells. Proc. Natl. Acad. Sci. USA *106*, 16257–16262.

Schwartz, B.E., Hofer, M.D., Lemieux, M.E., Bauer, D.E., Cameron, M.J., West, N.H., Agoston, E.S., Reynoird, N., Khochbin, S., Ince, T.A., et al. (2011). Differentiation of NUT midline carcinoma by epigenomic reprogramming. Cancer Res. *71*, 2686–2696.

Shang, E., Nickerson, H.D., Wen, D., Wang, X., and Wolgemuth, D.J. (2007). The first bromodomain of Brdt, a testis-specific member of the BET sub-family of double-bromodomain-containing proteins, is essential for male germ cell differentiation. Development *134*, 3507–3515.

Sharma, N., and Nyborg, J.K. (2008). The coactivators CBP/p300 and the histone chaperone NAP1 promote transcription-independent nucleosome eviction at the HTLV-1 promoter. Proc. Natl. Acad. Sci. USA *105*, 7959–7963.

Vizcaíno, J.A., Csordas, A., del-Toro, N., Dianes, J.A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., et al. (2016). 2016 update of the PRIDE database and its related tools. Nucleic Acids Res. *44* (*D1*), D447–D456.

2.2.4 Dynamique des remaniements du paysage nucléosomal

Cette section contient des résultats confidentiels au moment de la soutenance.

3 Développements méthodologiques pour une recherche reproductible en bioinformatique

3.1 metaworkflow : Une expérience méthodologique afin de déterminer des conventions de codage permettant de développer un flux de travail en Snakemake reproductible et automatisé pouvant produire tous types d'analyses basées sur des outils en ligne de commande

3.1.1 Analyse du gestionnaire de flux de travail Snakemake

3.1.1.1 Principe et fonctionnalités

Snakemake est un gestionnaire de flux de travail inspiré de l'outil de compilation GNU Make. Il récupère de ce dernier le concept de réentrabilité et le principe d'inférence d'un flux de travail complet à partir d'un ensemble de règles définissant des actions à effectuer sur des fichiers d'entrée (input) pour produire des fichiers de sortie (output). Par rapport à son prédecesseur, Snakemake propose en plus :

— une syntaxe plus accessible proche de pseudocode et dérivé de Python (FIGURE 3.1a),

— un système qui supporte l'inférence de chemin de fichiers avec de multiples *wildcards* au lieu d'une seule pour GNU Make,

— une fonctionnalité de génération du graphe orienté acyclique correspondant (FIGURE 3.1b),

— une scalabilité automatique à partir des tâches parallélisables,

— l'orchestration des tâches en local, via soumission sur un cluster de calcul et dans le cloud via kubernetes,

— le déploiement d'environnements logiciels contrôlés via conda et singularity,

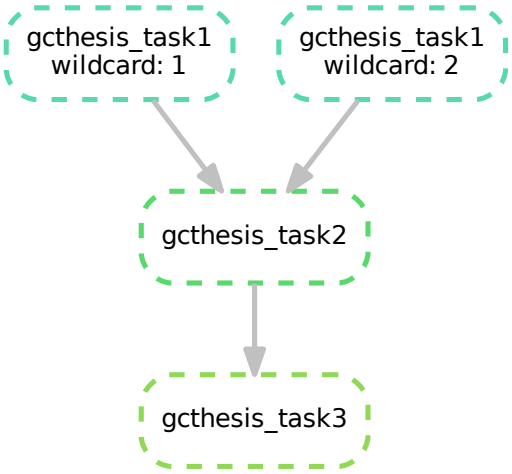
— et de nombreuses autres fonctionnalités qu'il serait trop long de lister.

Snakemake possède également

— un support réactif grâce à une communauté riche,

```
rule gcthesis_task1:
   output: "date_{wildcard}.txt"
   shell: "date +%s%N > {output}"

rule gcthesis_task2:
   input:
      "date_1.txt",
      "date_2.txt"
   output: "cat_dates.txt"
   shell: "cat {input} > {output}"

rule gcthesis_task3:
   input: "cat_dates.txt"
```

(a) Source d'un flux de travail snakemake simple

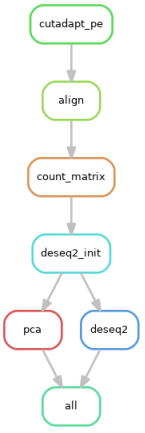


(b) Diagramme orienté acyclique (DAG) de ses tâches

FIGURE 3.1 – Exemple de flux de travail avec Snakemake. Les wildcards correspondent à des chaines de caractères variables qui peuvent être définies entre accolades dans l'input et l'output de chaque règle. Dans l'exemple, la règle `gcthesis_task1` contient une wildcard nommée ici sobrement wildcard bien qu'elle pourrait prendre un autre nom. Par défaut, Snakemake sait qu'il peut produire à partir de cette règle n'importe quel fichier de la forme `date_{wildcard}.txt` où wildcard peut être remplacé par n'importe quelle chaîne de caractères non nulle. On peut noter que cet exemple est automatisé mais non reproductible car l'action effectuée dans la première règle est dépendante du moment d'exécution.

- rna-seq-star-deseq2
- dna-seq-gatk-variant-calling
- accel-amplicon-trimming
- single-cell-rna-seq
- single-cell-drop-seq
- SnakeChunks
- Sequana

(a) Liens vers les sources des flux



(b) DAG du premier flux

FIGURE 3.2 – Exemples de flux de travail développés et mis à jour par la communauté scientifique pour la génomique.

— un développement actif et des mises à jours fréquentes (version stable 5.4.5 sortie en avril 2019),

— de nombreux exemples de flux de travail pour l'analyse de données génomiques (FIGURE 3.2).

3.1.1.2 Points délicats pour la prise en main

Le principe d'inférence d'un flux de travail à partir des règles, propre à Snakemake et à tous les outils inspirés de GNU Make, impose au développeur une façon de penser particulière "en arrière" qui peut sembler contre-intuitive. Ainsi, pour exécuter un flux de travail constitué de trois étapes enchaînées, il faut directement demander l'exécution de la dernière et compter sur le système de résolution des dépendances pour retrouver l'enchainement des étapes correct. En cas d'erreur dans la formulation des règles, il est difficile pour un non-initié d'identifier d'où vient l'erreur. Les bioinformaticiens préférant formaliser leur pensée de façon chronologique à l'image de l'enchainement des flux de données des "pipes" UNIX pourront lui préférer NextFlow.

Une force de l'approche "en arrière" est qu'elle permet de s'assurer avant le début de l'exécution du flux de travail que l'ensemble des enchaînements des règles est cohérent et qu'il n'existe aucune ambiguité possible entre règles pour la génération de tous les fichiers intermédiaires et finaux. L'apparition de ces ambiguités est une difficulté que va rencontrer un développeur Snakemake à mesure que son flux de travail se complexifie (FIGURE 3.3). Si Snakemake propose des mécanismes pour résoudre les ambiguités (FIGURE 3.4), ils ne peuvent pas résoudre tous les cas de figures (FIGURE 3.5). Dans ces cas, il semble indispensable de réécrire ces règles pour en limiter la portée (FIGURE 3.6) ou de se passer d'une partie des règles conflictuelles.

La plus grande difficulté en tant que développeur de flux de travail Snakemake est de trouver des patterns pour les outputs possédant le juste équilibre entre généralité et spécificité de façon à pouvoir appliquer une règle dans de multiple contextes, tout en n'entrant jamais en conflit avec une autre règle. Des bonnes pratiques de développement de flux de travail Snakemake sont données par son créateur mais contournent plus que ne résolvent cette problématique en préconisant de séparer chaque type d'analyse dans des flux de travail séparés (FIGURE 3.2). Cette solution sera définie pour la suite du propos comme étant le paradigme de développement de référence *"Un type d'analyse par flux de travail"*, ou plus simplement *paradigme de référence*.

3.1.1.3 Inconvénients perçus du paradigme de développement de référence *"Un type d'analyse par flux de travail"*

Pour ma part, cette solution n'était pas satisfaisante pour plusieurs raisons. La principale est liée à mon objectif de recherche reproductible automatisée. Dans un contexte biologique où la plupart des types d'analyses de données doivent être intégrées entre elles à une étape du processus de recherche, il est important

```
rule gzip:
   "Compress any file to gz"
   input:
      "{filepath}"
   output:
      "{filepath}.gz"

rule wget:
   "Download any file from Internet"
   output:
      "wget/{url}"

rule deepTools_computeMatrix:
   "Produce a count matrix"
   output:
      "deepTools/{matrix}.txt.gz"
```

FIGURE 3.3 – Exemple de règles Snakemake ambigües. Les patterns de fichiers commençant par `wget/` ou `deepTools/` et terminant par `gz` peuvent être produit par la règle `wget` ou `deepTools_computeMatrix` mais aussi par la règle `gzip`. Snakemake refusera d'exécuter le flux de travail avec le message d'erreur : `AmbiguousRuleException`. Par exemple, demander le fichier `deepTools/test.txt.gz` est ambigu car il peut être produit soit en remplaçant `matrix` par `test` dans la règle `deepTools_computMatrix`, soit en remplaçant `filepath` par `deepTools/test.txt` dans la règle `gzip`.

```
ruleorder: deepTools_computeMatrix > wget > gzip

rule gzip:
   "Compress any file to gz"
   input:
      "{filepath}"
   output:
      "{filepath}.gz"

rule wget:
   "Download any file from Internet"
   output:
      "wget/{url}"

rule deepTools_computeMatrix:
   "Produce a count matrix"
   output:
      "deepTools/{matrix}.txt.gz"
```

FIGURE 3.4 – Exemple d'application du mécanisme de priorité pour résoudre les règles ambigües dans Snakemake. En cas d'ambiguité, Snakemake choisira si possible prioritairement la règle `deepTools_computeMatrix` puis la règle `wget` et enfin la règle `gzip`.

```
rule gunzip:
   "Uncompress any gz file"
   input:
      "{filepath}.gz"
   output:
      "{filepath}"

rule gzip:
   "Compress any file to gz"
   input:
      "{filepath}"
   output:
      "{filepath}.gz"
```

FIGURE 3.5 – Exemple de règles circulaires que le mécanisme de priorité ne peut résoudre. Un fichier de sortie de la règle **`gunzip`** peut être utilisé par la règle **`gzip`** pour produire son fichier d'entrée.

```
rule gunzip:
   "Uncompress any file to gunzip directory"
   input:
      "{filepath}.gz"
   output:
      "gunzip/{filepath}"

rule gzip:
   "Compress any file to gzip directory"
   input:
      "{filepath}"
   output:
      "gzip/{filepath}.gz"
```

FIGURE 3.6 – Exemple de restriction de la portée de règles pour résoudre des règles ambigües. Les règles gunzip et gzip ne sont plus conflictuelles car chacune ne peut produire des fichiers que dans un dossier qui leur est propre. Il n'est plus possible de compresser ou décompresser un fichier dans son dossier.

de disposer d'un flux de travail qui peut orchestrer l'exécution de l'ensemble des tâches de façon optimisée. En isolant *de facto* les flux de travail pour des types de données différentes, le paradigme de référence prive d'une grande partie de la scalabilité de Snakemake dans de nombreux cas de figures usuels. Chaque flux de travail doit alors être exécuté séquentiellement, par exemple en liant leur exécution dans un script shell. Alternativement, Snakemake propose une mécanique de sous-flux de travail pour ces applications mais elle impacte tout autant les performances (FIGURE 3.7).

Un autre inconvénient du paradigme de référence est qu'il incite à écrire des flux de travail peu flexibles, conçus pour effectuer une série d'actions définies par un fichier de configuration yaml à partir de fichiers d'entrée formatés selon l'exemple-type proposé par le créateur du flux de travail. Si le format de certains fichiers à traiter diffère légèrement par rapport à ce qui est attendu, par exemple le suffixe des noms de fichiers ou le type de compression des fichiers, alors il encourage à effectuer manuellement des modifications dessus ce qui entraîne un risque d'erreur et une perte de traçabilité. Encore une fois, il serait possible d'écrire un sous-flux de travail préparant les données sous le format attendu, et un autre pour orchestrer l'enchaînement des sous-flux de travail, mais c'est fastidieux et surtout nuisible pour les performances pour la raison évoquée précédemment.

Enfin, une composante majeure du travail de bioinformaticien n'est pas d'appliquer des flux de travail clés-en-main mais de développer, d'ajuster et d'améliorer des flux de travail en cherchant de meilleures combinaisons d'outils et de paramètres. Pourtant, le paradigme de référence ne propose pas de mécanique simple et efficiente en stockage et en ressources de calcul pour comparer et substituer différents outils et paramètres au sein d'un flux de travail en cours de développement (FIGURE 3.8).

Face à ces inconvénients, il pourrait être possible de penser que la philosophie de Snakemake n'est pas adaptée pour le développement de flux de travail complexes. Un gestionnaire de flux de travail utilisant une philosophie différente comme NextFlow ou un gestionnaire basé sur la spécification CWL pourrait peut-être proposer un cadre plus robuste. Je n'ai cependant pas exploré ces pistes et préféré une autre : Chercher à définir un groupe de conventions de développement de règles Snakemake qui permettent le juste équilibre entre portée généralisée et risque limité de conflits entre règles. Afin de tester la robustesse des règles face à ces risques, il semblait naturel d'envisager d'obtenir un unique flux de travail capable de produire tous les types d'analyses de données. Cette solution sera définie pour la suite du propos comme étant le paradigme de développement alternatif *"Un flux de travail pour tous les types d'analyses"*, ou plus simplement *paradigme alternatif*.

```
# sw = subworkflow
#    = sous-flux de travail
rule sw_task:
   output:
      "sw_task_output.txt"
   shell:
      "sleep 10; touch {output}"




# mw = main workflow

rule mw_targets:
   input:
      "mw_sw_dep_task_output.txt",
      "mw_indep_task_output.txt"

rule mw_sw_dep_task:
   input:
      "sw_task_output.txt"
   output:
      "mw_sw_dep_task_output.txt"
   shell:
      "sleep 10; touch {output}"

rule mw_indep_task:
   output:
      "mw_indep_task_output.txt"
   shell:
      "sleep 20; touch {output}"
```

(a) Flux de travail autonome

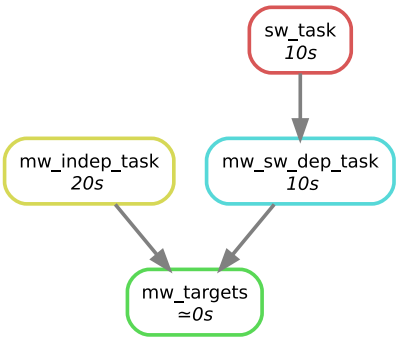
subworkflow

```
rule sw_task:
   output:
      "sw_task_output.txt"
   shell:
      "sleep 10; touch {output}"
```

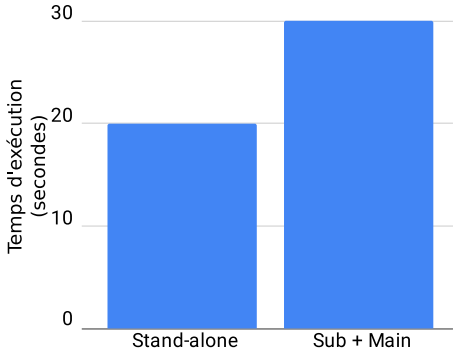
main workflow

```
subworkflow sw:
   snakefile:
      "subworkflow.smk"

rule mw_targets:
   input:
      "mw_sw_dep_task_output.txt",
      "mw_indep_task_output.txt"

rule mw_sw_dep_task:
   input:
      sw("sw_task_output.txt")
   output:
      "mw_sw_dep_task_output.txt"
   shell:
      "sleep 10; touch {output}"

rule mw_indep_task:
   output:
      "mw_indep_task_output.txt"
   shell:
      "sleep 20; touch {output}"
```

(b) Flux de travail séparées en sous-flux



(c) Graphe Acylique Orienté



(d) Banc d'essai

FIGURE 3.7 – Exemple minimaliste de performances entre un flux de travail autonome et un flux de travail reposant sur un sous-flux de travail. Bien que les deux approches effectuent les mêmes étapes, le flux de travail autonome est plus rapide dans le cas d'une exécution parallélisée car la tâche **`mw_indep_task`** peut commencer dès l'exécution du flux de travail en parallèle de la tâche **`sw_task`**. Dans le cas de l'utilisation d'un sous-flux de travail, cette tâche ne peut commencer que lorsque le sous-flux de travail est terminé, ce qui génère des goulots d'étranglement impactant les performances. Les différences restent anecdotiques pour cet exemple mais peuvent atteindre plusieurs ordres de grandeurs pour des cas réels contenant des dizaines d'échantillons de plusieurs types.

```
configfile: "config.yaml"

rule task_to_benchmark:
    input:
        "preprocess_task_output.txt"
    output:
        "task_to_benchmark_output.txt"
    params:
        config["task_to_benchmark"]["args"]
    shell:
        "touch {params.args} {output}"

rule preprocess_task:
    output:
        "preprocess_task_output.txt"
    shell:
        "sleep 10; touch {output}"
```

(a) Flux de travail pour banc d'essai

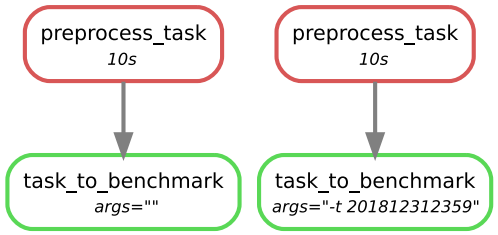
config1.yaml

```
task_to_benchmark:
  args: ""
```

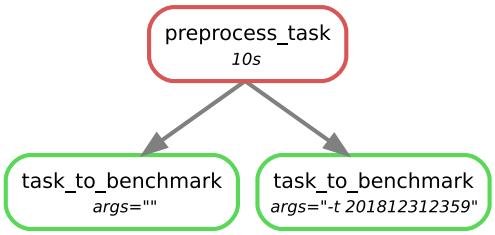
config2.yaml

```
task_to_benchmark:
  args: "-t 201812312359"
```

(b) yaml de configuration pour chaque variante à tester



(c) Graphes acycliques orientés obtenus



(d) Graphe acyclique orienté désiré

FIGURE 3.8 – Exemple minimaliste de banc d'essais. Selon le paradigme de référence, un flux de travail (3.8a) doit se configurer par un fichier de configuration yaml (3.8b). Comme le produit du banc d'essais `task_to_benchmark_output.txt` possède le même nom, il est nécessaire de dupliquer le flux de travail dans deux dossiers séparés, ce qui implique également d'effectuer en double l'étape de préprocess en amont de la tâche à benchmarker (3.8c). Des ressources de calculs sont inutilement gâchées dans le processus, tout comme l'espace de stockage qui doit contenir deux versions de `preprocess_task_output.txt`. Idéalement, un flux de travail optimisé pour les bancs d'essai devrait pouvoir se passer de cette duplication (3.8d).

3.1.2 Expérimentations sur un paradigme alternatif *"Un flux de travail pour tous les types d'analyses"*

En étant basé sur Python, Snakemake dispose d'un panel de possibilités pour décrire un flux de travail bien plus riche que ce que les exemples proposés jusqu'à présent ont pu dévoiler. Pourtant, cette richesse et cette liberté de possibilités peuvent être génératrice de conflits entre règles et sont sources d'indécisions pour le développeur. Un sous-ensemble de ces possibilités semble être suffisant pour décrire tous les types possibles de règles. Ses principes clés détaillés ci-dessous sont :

1. Toutes les règles peuvent et doivent être contraintes à écrire leurs sorties dans leur propre dossier.
2. Toutes les règles peuvent accepter un pattern global pour les fichiers d'entrée.
3. Toutes les combinaisons de paramètres peuvent être généralisées grâce à une fonction Python `params_extra` et un tableau associatif.
4. Toutes les règles peuvent être généralisées en fonction des types de fichiers en entrée et sortie.
5. Des fichiers d'origines diverses peuvent être rassemblés si besoin dans un même dossier et renommés grâce à une règle `ln_alias` et un tableau associatif.

3.1.2.1 Principes clés du paradigme alternatif

Toutes les règles peuvent et doivent être contraintes à écrire leurs sorties dans leur propre dossier. Déjà évoqué précédemment, les problèmes d'ambiguités sont la principale source d'incompatibilités entre des flux de travail développés par différentes personnes. En bioinformatique, il est commun de trouver par exemple des règles pour certains aligneurs qui produisent toutes des fichiers selon un pattern sensiblement proche de `"aligned/{sample}.bam"` ou `"{sample}.bam"`. Il est alors impossible de simplement intégrer ces différentes règles dans son propre flux de travail sans avoir à les modifier. Cet exemple illustre pourquoi il est préférable d'éviter de choisir des préfixes de fonctions génériques comme `aligned`. Une alternative qui permet d'anticiper à l'avance l'absence de conflits avec de futures règles est d'utiliser le nom de l'outil comme première partie du préfixe, suivi de la sous-commande de l'outil s'il en possède une. Les outils différents possédant un même nom sont rarissimes voire inexistants si l'on se restreint à des domaines spécifiques comme la bioinformatique. Le risque d'ambiguités est alors grandement diminué. Il n'est cependant jamais éliminé car plusieurs développeurs peuvent avoir pensé à des implémentations différentes pour un même outil. Dans ce cas, il est nécessaire de fusionner ces implémentations ou de les discriminer en employant un nom de fonction alternatif pour l'une d'elle.

```
rule samtools_index:
   input:
      bam="out/{filler}.bam",
   output:
      bam="out/samtools/index/{filler}.
       bam",
      bai="out/samtools/index/{filler}.
       bam.bai"
   shell:
      "ln -srf {input.bam} {output.bam};
       "
      "samtools index {output.bam}"
```

(a) Outil écrivant son fichier de sortie dans le dossier du fichier d'entrée

```
rule thisTool:
   input:
      "out/{filler}"
   output:
      "out/thisTool/{filler}"
   shell:
      """
      WDIR=`pwd`
      OUTDIR=`dirname {output}`
      cd $OUTDIR
      thisTool $WDIR/{input}
      """
```

(b) Outil écrivant ses fichiers dans le dossier où il est exécuté

FIGURE 3.9 – Exemples d'outils devant être contraints à écrire dans leur propre dossier. `-srf` est la combinaison d'arguments pour `ln` qu'il est préférable d'utiliser par défaut à l'intérieur de règles. `-s` crée un lien symbolique à la place d'un lien matériel car la création de ce dernier met à jour l'horodatage du fichier cible incitant Snakemake à relancer les règles en dépandant à la prochaine exécution. `-f` écrase la cible si elle existe. `-r` permet la création de liens symboliques relatif au dossier de sortie plutôt qu'au dossier d'exécution.

La grande majorité des outils s'accordent bien avec ce principe car ils disposent soit d'un argument permettant de définir l'emplacement du ou des fichiers de sorties, soit ils retournent leur résultat dans la sortie standard qui peut être redirigée vers le nom de fichier souhaité. Deux cas de figures rencontrés semblent plus problématiques. En effet, certains outils sont conçus pour produire un fichier de sortie dans le même dossier que leur fichier d'entrée. C'est notamment le cas des outils d'indexation de fichiers tel `samtools index`. Il suffit dans ce cas de créer un lien symbolique entre le fichier d'entrée et le dossier de sortie souhaité (FIGURE 3.9a). D'autres outils produisent des fichiers dans le dossier dans lequel ils ont été exécutés. Se déplacer dans le dossier de sortie avant d'exécuter un tel outil permet de résoudre cette limitation (FIGURE 3.9b).

Appliquer un tel concept dans un flux de travail peut rapidement produire un très grand nombre de dossiers dans la racine du projet. Pour éviter de noyer les autres composantes d'un flux de travail comme le code source ou les fichiers d'entrées, on peut convenir d'un unique dossier pour contenir tous les produits de l'exécution de règles. Nommé `out` dans mon cas, il permet également de disposer d'un mécanisme simplissime pour nettoyer une analyse terminée des éléments non essentiels à son archivage. Il suffit de supprimer le dossier `out`.

Toutes les règles peuvent accepter un pattern global pour les fichiers d'entrée. Conséquence de l'application du premier principe, il n'est pas nécessaire de res-

```
rule targets:
   input:
      "out/task3/s1.txt",
      "out/task3/s2.txt"
      # Impossible to bypass
      # task2 with this workflow

rule task1:
   output:
      "out/task1/{filler}"

rule task2:
   input:
      "out/task1/{filler}"
   output:
      "out/task2/{filler}"

rule task3:
   input:
      "out/task2/{filler}"
   output:
      "out/task3/{filler}"
```

(a) Flux de travail restreignant les fichiers d'entrée

```
rule targets:
   input:
      "out/task3/task2/task1/s1.txt",
      "out/task3/task2/task1/s2.txt",
      "out/task3/task1/s1.txt",
      "out/task3/task1/s2.txt"

rule task1:
   output:
      "out/task1/{filler}"

rule task2:
   input:
      "out/{filler}"
   output:
      "out/task2/{filler}"

rule task3:
   input:
      "out/{filler}"
   output:
      "out/task3/{filler}"
```

(b) Flux de travail ne restreignant pas les fichiers d'entrée

FIGURE 3.10 – Exemples et conséquences de flux de travail restreignant la portée des fichiers d'entrée. En restreignant cette portée, les fichiers finaux possèdent des chemins plus courts mais on se prive d'un mécanisme intéressant pour contourner certaines règles optionnelles si besoin. En ne restreignant pas la portée des fichiers d'entrée, il est également pratique de générer naturellement des chemins de fichiers qui permettent de conserver la traçabilité complète des étapes ayant menées à un fichier.

treindre la portée des fichiers d'entrées. Le faire permet d'obtenir des chemins de fichiers courts (FIGURE 3.10a) mais prive de mécanismes pour contourner certaines étapes optionnelles et d'une traçabilité explicite des différentes étapes ayant produit un fichier (FIGURE 3.10b). `filler` est un nom arbitraire pour la wildcard par défaut des règles du fait de l'empilement des noms des étapes qu'il contient.

On peut noter toutefois dans les exemples que les fichiers d'entrées sont quand même restreints à chercher dans le dossier `out`. Ce choix à pour objectif d'éviter des chemins redondants du type `out/task3/out/task2/out/task1/s1.txt`. Pour intégrer des fichiers de données d'entrées privées (SECTION 1.4.3.1), il est possible de faire une exception pour des règles applicant l'outil de création de liens `ln` (FIGURE 3.11).

```
rule ln_srf_parent_dir:
   input:
      "../{filler}"
   output:
      "out/ln/updir/{filler}"
   shell:
      "ln -srf '{input}' '{output}'"
```

(a) Intégration de données à partir du niveau du flux de travail dans l'arborescence du système

```
rule ln_srf_abspath:
   input:
      "/{filler}"
   output:
      "out/ln/abspath/{filler}"
   shell:
      "ln -srf '{input}' '{output}'"
```

(b) Intégration de données à partir de la racine dans l'arborescence du système

FIGURE 3.11 – Exemples des règles intégrant des données d'entrées privées. D'autres variantes seraient possibles si besoin bien que seule la première soit suffisante pour toutes mes applications. Notons que l'encadrement des `input` et `output` par des guillemets simples permet d'éviter l'erreur survenant lorsqu'un nom de fichier contient des caractères pouvant perturber le shell.

Toutes les combinaisons de paramètres peuvent être généralisées grâce à une fonction Python `params_extra` et un tableau associatif. De nombreux outils peuvent avoir des résultats très différents en fonction de paramètres additionnels. Dans ce cas il est intéressant de pouvoir produire les fichiers de sortie pour ces différents paramètres avec si possible un nombre restreint de règles. La solution la plus généralisée consiste à utiliser une fonction Python pour traduire une wildcard de paramètres intercalée dans le pattern de sortie entre l'outil, sa sous-commande optionnelle et le filler selon un pattern générique du type `out/{tool}{extra}/{filler}` où `tool` contient le nom de l'outil et sa fonction optionnelle et `extra` contient un descriptif encodé de la liste des paramètres optionnels. Le choix du nom `extra` pour cette wildcard est inspiré de règles du créateur de Snakemake où celle-ci contient l'ensemble des paramètres optionnels, par exemple ici. La façon d'encoder la liste des paramètres optionnels n'est limitée que par la complexité de la fonction python que l'on souhaite appliquer. Dans mon cas, cette fonction `params_extra` est accessible ici. Comme elle repose sur certains principes non-abordés, son fonctionnement exact n'est pas détaillé pour l'instant. Schématiquement, elle effectue deux actions. Premièrement elle remplace chaque "_" (tiret bas) par un " " (espace) dans la wildcard `extra`. Deuxièmement, elle remplace certains mots-clés contenus dans la wildcard `extra` par leur valeur renseignée dans un tableau associatif (TABLE 3.1). La chaîne de caractères ainsi obtenue peut être appliquée directement dans le shell d'une règle (FIGURE 3.12).

Toutes les règles peuvent être généralisées en fonction des types de fichiers en entrée et sortie. Jusqu'à présent, toutes les règles présentées sont simples dans le sens où elles requièrent un fichier d'entrée, un outil, sa fonction et des

tool	id	value
thisTool	idFromTable	`--itsArg -i -j --key value`
thisTool	otherIdFromTable	`--special_arg_with_space`
anotherTool	idFromTable	`not used for thisTool`

TABLE 3.1 – Exemple de tableau associatif utilisé par la fonction `params_extra`.

```
rule targets:
   input:
      "out/thisTool/itsFun/s1.extout",
      # without optional argument

      "out/thisTool/itsFun_--itsArg/s1.extout",
      # with "--itsArg" argument

      "out/thisTool/itsFun_-i_-j/s1.extout",
      # with "-i" and "-j" arguments

      "out/thisTool/itsFun_--key_value/s1.extout",
      # with "--key value" argument

      "out/thisTool/itsFun_idFromTable/s1.extout",
      # with all theses args deduced from table

      "out/thisTool/itsFun_otherIdFromTable/s1.extout"
      # with "--special_arg_with_space" deduced from table

rule thisTool_itsFun:
   input:
      "out/{filler}.extin"
   output:
      "out/{tool}{extra}/{filler}.extout"
   wildcard_constraints:
      tool="thisTool/itsFun", # tool can only take one value here.
      extra = "[^\/]*" # extra can take any string without "/" or none.
   params:
      params_extra # see mw-lib/src/snakemake/functions/params_extra.py
   shell:
      "thisTools itsFun {params} {input} > {output}"
```

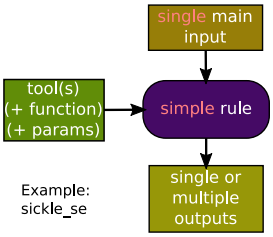
FIGURE 3.12 – Exemple de flux de travail pouvant tester toutes les combinaisons possibles d'arguments d'un outil à partir d'une seule règle et de la fonction `params_extra`. La fonction `itsFun` de l'outil `thisTool` est appliquée sur le fichier `out/s1.extin` avec différents paramètres. Comme la fonction `params_extra` n'a accès qu'aux valeurs prises par les différentes wildcards d'une règle, le préfixe de l'outil est remplacé par une wildcard `tool` ne pouvant prendre qu'une seule chaîne de charactères définie par le mot-clé `wildcard_constraints`. Dans la TABLE 3.1, c'est donc la valeur `--itsArg -i -j --key value` qui est retournée et non `not used for thisTool` car ce dernier est déclaré pour l'outil `anotherTool`.

paramètres pour produire un fichier de sortie. Toutes les règles peuvent être classifiées en quatres types et ces exemples font partie de la première catégorie (FIGURE 3.13a). Une variante consiste à nécessiter non pas un mais un nombre invariable de fichiers d'entrées de même type. C'est par exemple le cas lorsque l'on manipule des données de séquençage "paired-end" et certaines règles prennent toujours deux fichiers d'entrées, un pour chaque brin séquencé (FIGURE 3.13a).

Lorsque des fichiers d'entrées de types différents sont nécessaires, on parle de règles intégratives. Si la règle intégrative nécessite un seul fichier pour un des types, alors ce type peut être fixé comme étant le principal et il est spécifié comme les fichiers des règles simples par extension de la wildcard `filler`. Le contenu des fichiers supplémentaires est spécifié par des identifiants à la manière des paramètres spécifiés par un tableau associatif (FIGURE 3.13c). Dans le cas où il n'existe pas de fichier d'entrée principal, tous les fichiers d'entrées sont spécifiés par un tableau associatif (FIGURE 3.13d). Dans ces exemples, la conversion des identifiants vers les chemins des fichiers d'entrées est réalisée par une fonction lambda Python qui va consulter les valeurs correspondant aux identifiants fournis dans le dictionnaire de configuration qui est chargé à l'exécution du flux de travail à partir d'un tableau associatif ou plusieurs tableaux associatifs.

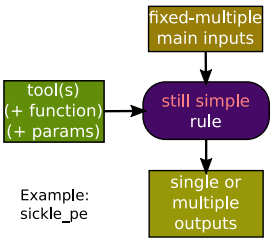
Des fichiers d'origines diverses peuvent être rassemblés si besoin dans un même dossier et renommés grâce à une règle `ln_alias` et un tableau associatif. Intégrer des fichiers provenant de sources différentes est une activité routinière en analyse de données. Les règles intégratives proposent un mécanisme pour l'effectuer de manière entièrement reproductible et automatisée. Cependant, de nombreux outils utilisent le nom de fichier, ou le chemin complet de chaque fichier comme nom d'échantillon. Intégrer des données de sources diverses peut alors produire des figures avec des noms d'échantillons hétérogènes. Un mécanisme simple basé sur la combinaison de principes déjà présentés permet d'homogénéiser de tels fichiers. Additionnellement, cela permet d'appliquer un type particulier de règles simples constituées de deux fichiers d'entrées, un échantillon test et un contrôle, sans passer par une règle de type intégrative (FIGURE 3.14 et TABLE 3.2).

Les règles snakemake doivent rester bêtes et disciplinées mais Python peut rajouter une couche d'intelligence optionnelle dans le flux de travail. La combinaison des principes précédents permet d'effectuer tous les types d'analyses possibles à condition de demander expressément les chemins de fichiers correspondant à l'enchaînement des étapes souhaitées, et d'avoir complété les tableaux associatifs correspondants si nécessaire. Le flux de travail ne fait aucun choix par lui-même et se contente de produire le seul enchaînement d'étapes possibles d'après les spécifications fournies. Ce comportement est très pratique en phase de développement de règles et d'analyses nouvelles car les sources d'erreurs et de bugs sont facilement identifiables et les étapes concernées isolables.



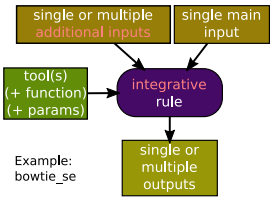

```
rule simple:
    input:
        "out/{filler}.extin"
    output:
        "out/{tool}{extra}/{filler}.extout"
    params:
        params_extra
    wildcard_constraints:
        tool="simple/"
    shell:
        "simple {params} {input} > {output}"
```

(a) Principe et syntaxe pour règle simple



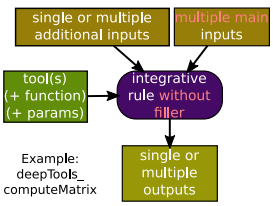

```
rule still_simple:
    input:
        "out/{filler}_1.extin",
        "out/{filler}_2.extin"
    output:
        "out/{tool}{extra}/{filler}.extout"
    params:
        params_extra
    wildcard_constraints:
        tool="stillSimple/"
    shell:
        "stillSimple {params} {input} > {output}"
```

(b) Principe et syntaxe pour règle à entrées multiples fixes




```
rule integrative_with_filler:
    input:
        main="out/{filler}.extin",
        suppl=lambda wildcards: eval(config['ids'][wildcards.
      suppl_id])
    output:
        "out/{tool}{extra}_{suppl_id}/{filler}.extout"
    params:
        params_extra
    wildcard_constraints:
        tool="integrative/"
    shell:
        "integrative {params} {input} > {output}"
```

(c) Principe et syntaxe pour règle intégrative avec filler




```
rule integrative_without_filler:
    input:
        type1=lambda wildcards: eval(config['ids'][wildcards.
      type1_id]),
        type2=lambda wildcards: eval(config['ids'][wildcards.
      type2_id])
    output:
        "out/{tool}{extra}_{type1_id}_{type2_id}.extout"
    params:
        params_extra
    wildcard_constraints:
        tool="integrative/"
    shell:
        "integrative {params} {input} > {output}"
```

(d) Principe et syntaxe pour règle intégrative sans filler

FIGURE 3.13 – Types de règles

```
rule targets:
   input:
      "out/thisTool/sampleVScontrol/source1/s1_VS_c1.extout",
      "out/thisTool/sampleVScontrol/source2/smp_VS_ctr.extout",
      # Impossible to compare directly s1 versus ctr, or smp versus c1 as above
      # because they do not belong to the same directory.
      # However ln_alias rule and an alias table allow to gather them in the
      # same directory, and rename smp and ctr to consistent names relative
      # to the samples from the source1:
      "out/thisTool/sampleVScontrol/ln/alias/integVarSources/s1_VS_c2.extout",
      "out/thisTool/sampleVScontrol/ln/alias/integVarSources/s2_VS_c1.extout",

rule thisTool_sampleVScontrol:
   input:
      sample="out/{filler}/{sample}.extin",
      control="out/{filler}/{control}.extin"
   output:
      "out/{tool}{extra}/{filler}/{sample}_VS_{control}.extout"
   params:
      params_extra
   wildcard_constraints:
      tool = "thisTool/sampleVScontrol",
      sample  = "[\w_-]+",
      control = "[\w_-]+"
   shell:
      "thisTool -s {input.sample} -c {input.control} -o {output}"

rule source1:
   output:
      "out/source1/{filler}.extin"

rule source2:
   output:
      "out/source2/{filler}.extin"

rule ln_alias:
   input:
      lambda wildcards: config['ids'][wildcards.id]
   output:
      "out/ln/alias/{id}"
   shell:
      "ln -srf '{input}' {output}"
```

FIGURE 3.14 – Exemple de flux de travail intégrant des fichiers de sources différentes.

id	value
integVarSources/s1.extin	out/source1/s1.extin
integVarSources/c1.extin	out/source1/c1.extin
integVarSources/s2.extin	out/source2/smp.extin
integVarSources/c2.extin	out/source2/ctr.extin

TABLE 3.2 – Exemple de tableau associatif utilisé par la règle `ln_alias`.

Lorsqu'un flux de travail concernant un type d'analyse atteint un certain niveau de maturité et que l'on souhaite l'appliquer en routine, il peut être intéressant que le flux de travail soit en mesure de générer ses propres cibles à partir d'un ensemble de métadonnées sur les échantillons à analyser. Ceci peut-être fait simplement grâce à une ou plusieurs fonctions Python intégrées dans le code source du flux de travail Snakemake et qui s'exécutent à son démarrage.

3.1.2.2 Résultats et retour d'expériences

Après plusieurs années à appliquer, ajuster et tester ce paradigme alternatif, voici succintement les résultats obtenus en termes de bénéfices et d'inconvénients relativement au paradigme de référence.

Bénéfices du paradigme alternatif Du fait de l'exploitation d'un unique graphe orienté acyclique représentant l'ensemble du flux de travail, les tâches sont mieux parallélisées ce qui permet d'obtenir plus rapidement des résultats pour l'intégration de données de différentes sources et types, ainsi que pour le benchmarking d'outils et de paramètres.

Du fait du partage d'un ensemble de règles communes à plusieurs types d'analyses et à la généralisation des règles autant que possible, le code est moins redondant ce qui facilite le maintien à jour des meilleures pratiques pour l'ensemble des types d'analyses. Le développement logiciel est alors moins redondant et moins frustrant pour le bioinformaticien. Il est également plus familier grâce aux principes déterminant à l'avance quasi-algorithmiquement comment doivent être nommés les patterns d'entrées et de sorties. Les règles ambigües sont évitées.

Enfin, une traçabilité supérieure de l'ensemble des fichiers est assurée. La simplicité d'intégrations de fichiers avec des formatages différents en entrée du flux de travail incite à automatiser toutes ces étapes préliminaires. Il en est de même pour les étapes postliminaires.

Inconvénients du paradigme alternatif Le paradigme alternatif n'est cependant pas exempt de défauts. Premièrement, l'exploration du dossier contenant toutes les données processées requiert des connaissances bioinformatiques sur les outils utilisés afin de se retrouver dans les méandres des chemins de fichiers (FIGURE 3.15). Partager les chemins vers les données processées n'est pas suffisant pour certains collaborateurs qui attendent une structure hiérarchique orientée "type d'analyse" et non "outils". Dans ce cas, le principe de renommage de fichiers avec une fonction d'alias peut être appliqué pour produire une ou plusieurs structures hiérarchiques alternatives mais cela demande une réflexion et des discussions avec les collaborateurs pour que les livrables puissent correspondre à leurs attentes (FIGURE 3.18).

Un point délicat à prendre en compte dans le paradigme alternatif est le risque de bugs liés à l'utilisation de chemins de fichiers trop longs. Si les systèmes de fichiers usuels sur MacOS et Linux ne présentent pas de limitations pratiques, ce

n'était pas le cas pour Windows avant la mise à jour anniversaire de Windows 10. Sous ces anciens systèmes d'exploitation, les chemins de fichiers étaient limités à 260 caractères ce qui peut être rapidement dépassé en cas d'enchaînements d'étapes d'analyses dans le flux de travail du type `out/taskN/taskN-1/taskN-2/.../task2/task1/sample.ext`. Certains outils comme R ou LaTeX peuvent ne pas accepter de noms de fichiers dépassant cette limite. Dans tous les cas où cette limitation deviendrait problématique, l'utilisation de la fonction d'alias pour redéfinir des chemins plus courts est possible (FIGURE 3.14 et TABLE 3.2).

Si le paradigme alternatif apporte plus de puissance au développeur initié, il demande également plus de connaissances que le paradigme de référence. Ce manque d'accessibilité le limite aux utilisateurs ayant déjà intégré le fonctionnement de Snakemake et rencontré les limitations de l'approche de référence pour des projets complexes.

Enfin, le paradigme alternatif est par postulat initial peu pratique pour le partage d'analyses dans un contexte à plusieurs projets impliquant des collaborations différentes. Puisqu'il n'y a qu'un seul flux de travail, tout le code source est partagé par tous les projets. Ceci est problématique lorsqu'une partie de ce code source contient des informations spécifiques d'un projet, voire confidentielles. Des ajustements pouvant être réalisés pour résoudre cette limitation sont abordés dans la section suivante.

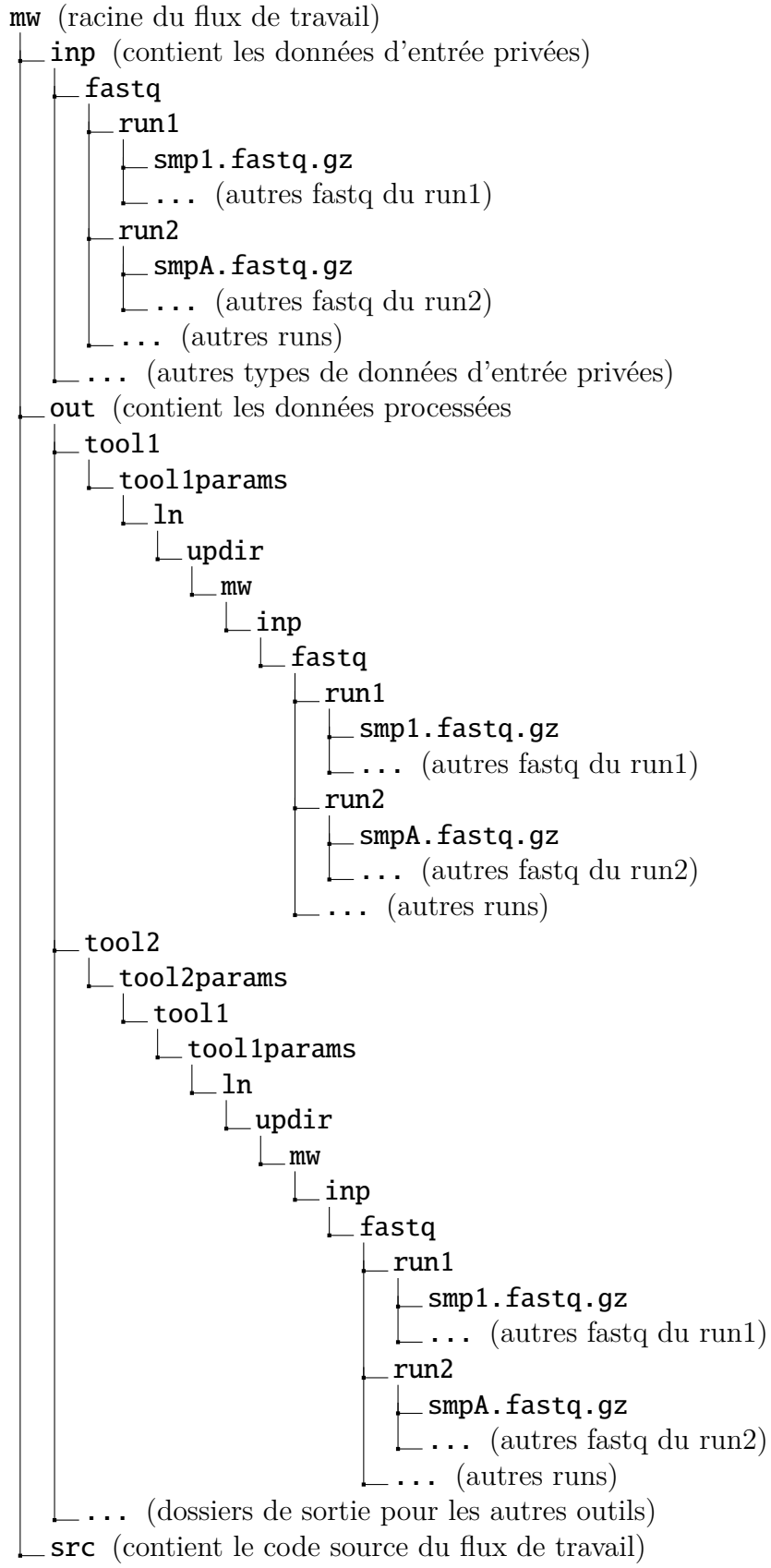


FIGURE 3.15 – Exemple d'arborescence des données produites selon le paradigme alternatif

3.1.3 Composants logiciels de la solution développée

Initialement, un composant contenant l'ensemble des éléments nécessaires aux analyses de tous les projets a été développé (FIGURE 3.16a). Il est possible de créer différentes versions d'un flux de travail contenant à chaque fois les éléments nécessaires pour un projet spécifique. Cependant l'intégration de ces versions en tant que branches d'un dépôt git est problématique car ce dernier ne peux contenir à la fois des branches publiques et des branches privées.

Une solution plus simple consiste à séparer dans différents dépôts git les éléments utiles pour plusieurs projets de ceux spécifiques d'un seul projet. Le flux de travail complet est alors constitué d'une librairie de règles Snakemake et de fonctions Python qu'il est possible de partager publiquement, puis de différents composants liés à un projet particulier et partagés uniquement aux collaborateurs concernés. Un composant privé peut être scindé en deux au moment de la publication d'un article de façon à rendre publique le code source nécessaire à la réalisation des analyses présentées dans l'article.

Dans l'implémentation actuelle, les différents composants correspondent à différents dépôts git installés au même niveau d'une arborescence de travail (FIGURE 3.16b). Ces dépôts ne sont pas explicitement liés mais il pourrait être possible d'expliciter ces liens en déclarant le composant principal `mw` comme sous-module d'un composant spécifique d'une analyse en utilisant la fonction `git submodule`. Des ajustements mineurs du code contenu dans le fichier Snakemake principal permettent d'inclure de façon transparente le code source de l'ensemble des composants installés. Les différents composants développés selon ce concept sont présentés ci-dessous.

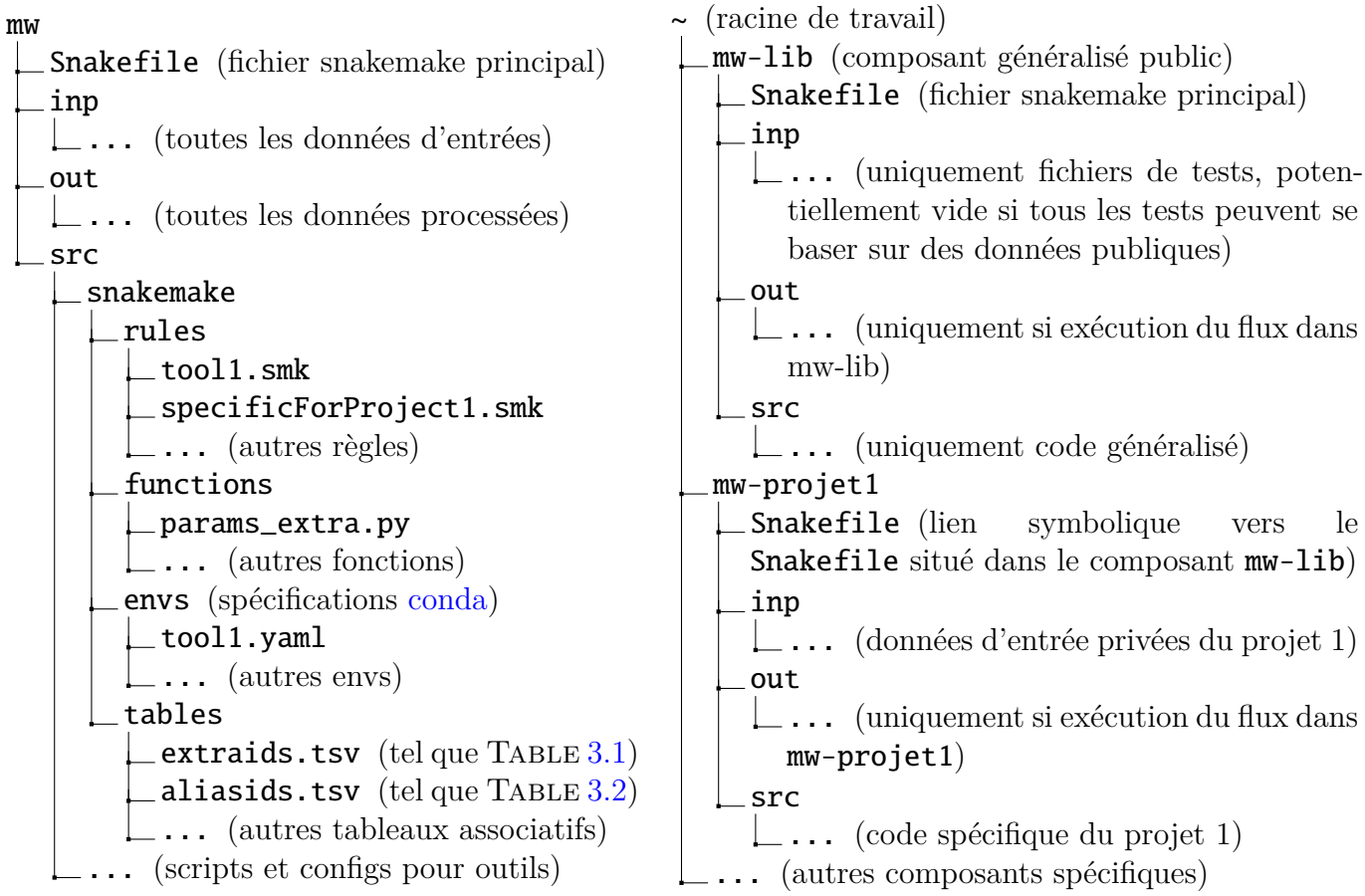
3.1.3.1 mw-lib : Une librairie de fonctions Pythons et règles Snakemake génériques partagées par des projets et analyses diverses

Plus de 150 paires d'outils-fonctions sont implémentées en tant que règles généralisées dans le dépôt principal :
«««< HEAD https://github.com/guillaumecharbonnier/mw-lib. ======= https://github.com/guillaumecharbonnier/mw-lib/src/snakemake/rules. »»»> refs/remotes/origin/master

3.1.3.2 mw-gcthesis : Un module pour la génération de cette thèse

Le code source de cette thèse et permettant sa génération est disponible dans un module dédié : https://github.com/guillaumecharbonnier/mw-gcthesis. Si le lien vers le dépôt renvoie vers une erreur 404, c'est que la période d'embargo sur certains résultats non publiés n'est pas terminée. En tant que membre de mon jury de thèse, vous pouvez si vous le souhaitez me demander par e-mail un accès privé à ce dépôt.

Pour compiler la thèse, un environnement contenant conda, snakemake et ses dépendances est nécessaire. Pour ce faire, installez miniconda en suivant



(a) Arborescence unique

(b) Arborescence éclatée en composants

FIGURE 3.16 – Organisation des composants en arborescences parallèles complémentaires. Une arborescence unique pour le paradigme alternatif (à gauche) peut être scindée en autant d'arborescences parallèles contenant chaque projet spécifique, ainsi qu'une librairie de fonctions générales. Les fichiers de code source sont chargés séquentiellement pour tous les composants présents à la racine de travail. Des fichiers de mêmes noms présents dans différents composants peuvent coexister. Il est par exemple commun que chaque projet possède son propre `aliasids.tsv` pour la clarté du code et la confidentialité.

les instructions officielles http://conda.pydata.org/miniconda.html avec les dépôts bioconda et conda-forge configurés dans cet ordre :

```
conda config --add channels defaults
conda config --add channels bioconda
conda config --add channels conda-forge
```

Ensuite, installez l'environnement de compilation et activez le :

```
conda create -n compileThesis conda snakemake git coreutils rpy2 pandas numpy
    biopython xlrd
conda activate compileThesis
```

Enfin installez les composants mw nécessaires et lancez la compilation :

```
git clone https://github.com/guillaumecharbonnier/mw-lib
git clone https://github.com/guillaumecharbonnier/mw-gcthesis
cd mw-gcthesis
./compile-thesis.sh
```

La rédaction d'une thèse est un bon exemple pour illustrer les possibilités d'intégration de rapports et analyses rédigés à partir de données d'entrées. Snakemake intègre une fonctionnalité de production de rapports basée sur le format reStructuredText. Il est également possible d'intégrer une règle de compilation de code R-Markdown via knitr parmi les étapes finales du flux de travail. Dans le cadre d'un document demandant un formatage avancé, il est possible d'intégrer une règle pour la compilation du code LaTeX (FIGURE 3.17).

3.1.3.3 mw-sst : Un module dédié à l'analyse de données de routine

Salvatore Spicuglia's Tree (SST) est un projet dont l'objectif est de proposer une interface simple d'accès pour le biologiste afin d'appliquer des traitements de routine sur les nouvelles données générées dans l'équipe. Le biologiste a besoin de renseigner uniquement des métadonnées pour chaque échantillon (TABLE 3.3) et obtient des données traitées rangées selon une architecture simplifiée (FIGURE 3.18). Le processus est presque entièrement automatisé et l'expertise bioinformatique est uniquement nécessaire pour transférer les données séquencées sur le serveur de calcul et compléter les métadonnées (FIGURE 3.19).

3.1.3.4 Autres composants

Des modules spécifiques à chaque analyse contenant des portions de code non généralisés ont été isolés :

— *mw-thymus* (privé) et *mw-cieslak2019* (public) pour les analyses présentées dans la SECTION 2.1.

— *mw-sk* (privé) pour les analyses présentées dans la SECTION 2.2 en collaboration avec l'équipe de Saadi Khochbin.

— *mw-broad* (public) pour les analyses synthétisées dans l'ARTICLE 8.

```
rule tectonic:
    input:
        tex="out/{filler}.tex",
        dep=latex_smi_dep
    output:
        pdf="out/tectonic/{filler}.pdf
      "
    conda:
        "../envs/tectonic.yaml"
    shell:
        """
        OUTDIR=`dirname {output.pdf}`
        INDIR=`dirname {input.tex}`
        for DIR in `ls -d */`;
        do
            ln -srf $DIR $INDIR/
            ln -srf $DIR $OUTDIR/
        done
        tectonic --outdir $OUTDIR -p {
       input.tex}
        """
```

(a) Règle Snakemake pour compilation LaTeX avec Tectonic

```
def latex_smi_dep(wildcards):
    filepaths = []
    filepaths.append("out/" +
     wildcards['filler'] + ".tex")
    pattern = re.compile(r"^(?:(?!%).)
     *?\\def\\smi{([^\}]+)}")
    for filepath in filepaths:
        if filepath.endswith(".tex") or
     filepath.endswith(".cls"):
            print(filepath + ' is a tex
     file')
            if os.path.isfile(filepath):
                with open (filepath, "rt")
     as infile:
                    for line in infile:
                        match = pattern.
     search(line)
                        if match:
                            print(line)
                            filepath = match.
     group(1)
                            filepaths.append(
     filepath)
    return(filepaths)
```

(b) Fonction Python pour analyse des dépendances dans le code LaTeX

```
\documentclass{article}
\begin{document}

\def\smi{out/wget/https/upload.wikimedia.org/wikipedia/commons/0/00/
    Dnaconformations.png}
\includegraphics{\smi}

\def\smi{out/wget/https/dl.dropboxusercontent.com/s/5bi8retzyn4dqz0/animal-cell-
    structure-en.pdf}
\includegraphics{\smi}

\end{document}
```

(c) Exemple d'application dans le code LaTeX

FIGURE 3.17 – Intégration de LaTeX dans Snakemake. En déclarant ses figures par l'intermédiaire de la variable "smi" (SnakeMake Input), il est possible d'automatiser au sein du même flux de travail les étapes d'analyses des données et de rédaction de rapports. Les figures requises dans le rapport et non présentes dans le dossier du projet sont générées par le flux de travail. Tectonic est une distribution LaTeX présentant des fonctionnalités plus intéressantes que Texlive pour une utilisation avec Conda.

métadonnée	description
sample_name	Obligatoire. Nom d'échantillon tel qu'utilisé dans l'arborescence des fichiers. Doit être unique. Certains caractères spéciaux entrant en conflit avec les règles de nommage de fichiers sont interdits.
control_name	Optionnel. Nom d'un échantillon à utiliser comme contrôle pour le *peak calling*.
quantile_norm_name	Optionnel. Nom d'un échantillon à utiliser comme référence pour produire des bigiwgs normalisés par les quantiles.
run	Optionnel. Numéro de run de séquençage.
type	Obligatoire. Type d'expérience (RNA, ChIP, ATAC, ...)
specie	Obligatoire. Nom d'espèce. Seulement *mouse* et *human* implémentés pour l'intant. Les deux assemblages de génomes les plus récents pour ces espèces sont utilisés en parallèle.
se_or_pe	Obligatoire. Indique si la librairie est *single-end* ou *paired-end*.
origin	Obligatoire. Indique l'origine de l'échantillon parmi ces choix : *tgml*, *blueprint*, *sra*, *merge_fastq* et *merge_bam*.
process	Obligatoire. *yes* ou *no*. Permet d'ignorer l'échantillon de la ligne. Utile lorsque les métadonnées d'un échantillon sont incomplètes et génèrent un bug lors de l'exécution.
merged_or_unmerged	Obligatoire si l'échantillon provient du TGML. Indique si les quatre voies du NextSeq500 ont été fusionnées en un seul fichier ou pas.
compressed_fastq	Obligatoire si l'échantillon provient du TGML. Indique l'algorithme de compression des fastq en entrée
srr_id	Identifiant SRR de l'échantillon s'il provient de SRA.
blueprint_fastq_prefix	Chemin et préfixe vers l'échantillon s'il est formaté tel que fourni par le projet Blueprint.
tgml_fastq_prefix	Chemin et préfixe vers l'échantillon s'il est formaté tel que fourni par le TGML.
exp_name	Optionnel. Nom d'expérience utilisé comme filtre
sample_merge_list	Obligatoire si l'échantillon est une fusion d'autres échantillons. Contient la liste des échantillons à utiliser pour la fusion.
cell_type	Optionnel. Type cellulaire. Utilisé comme filtre.
chip_target	Optionnel. Cible de ChIP. Utilisé comme filtre.
others	D'autres métadonnées peuvent être complétées.

TABLE 3.3 – Types de métadonnées récoltées dans `Sequencing_summary.xlsx`.

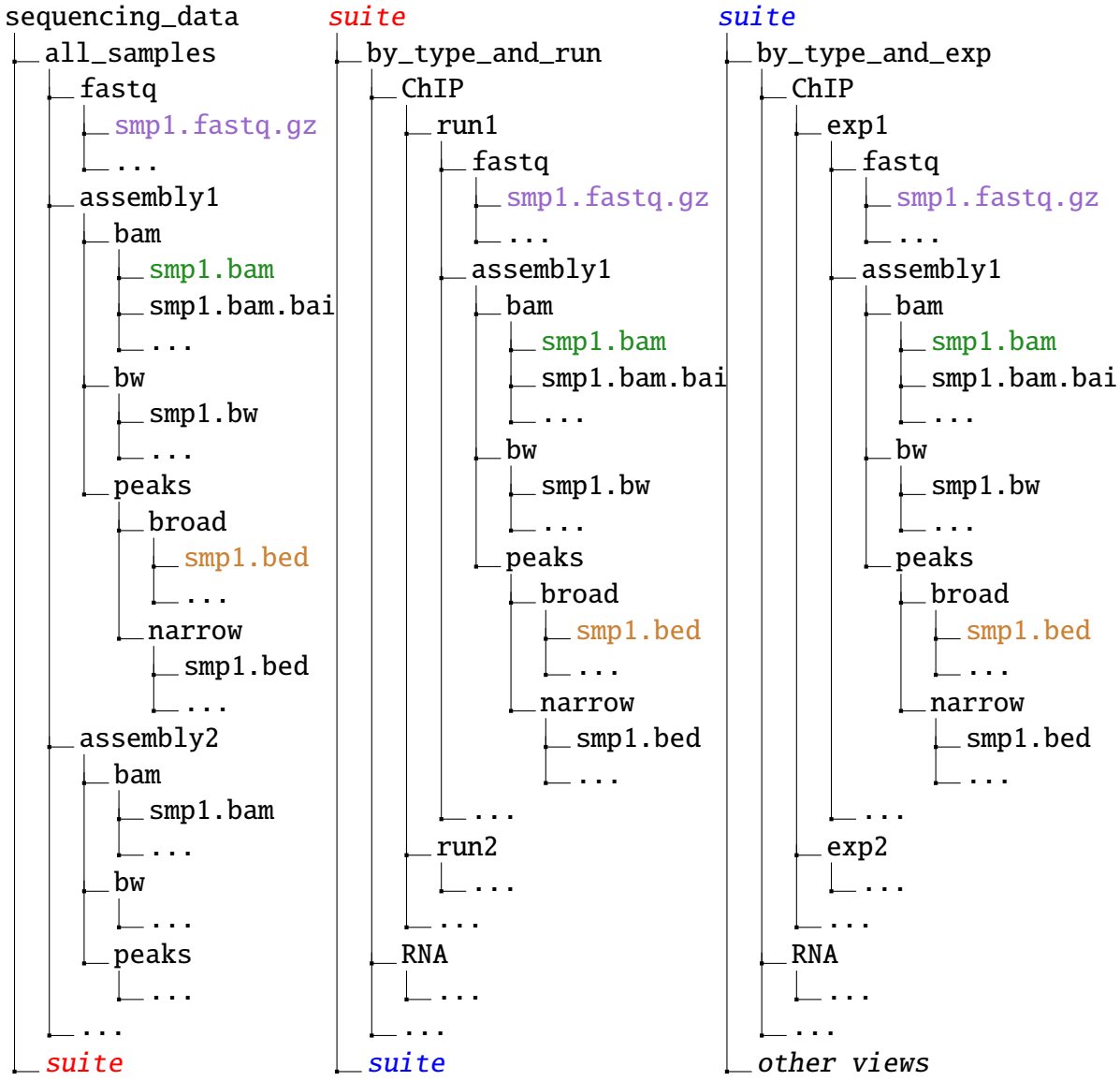


FIGURE 3.18 – Arborescence accessible pour le biologiste dans le projet mw-sst. L'architecture est sectionnée en trois parties au niveau des connecteurs *suite* de même couleur. Les noms de fichiers de même couleur correspondent au même fichier et visible selon différentes vues sans duplication grâce à l'utilisation de liens matériels. Trois vues sont implémentées : tous les échantillons mélangés, par type puis par run, par type puis par expérience. D'autres vues peuvent être créées par combinaisons des métadonnées disponibles.

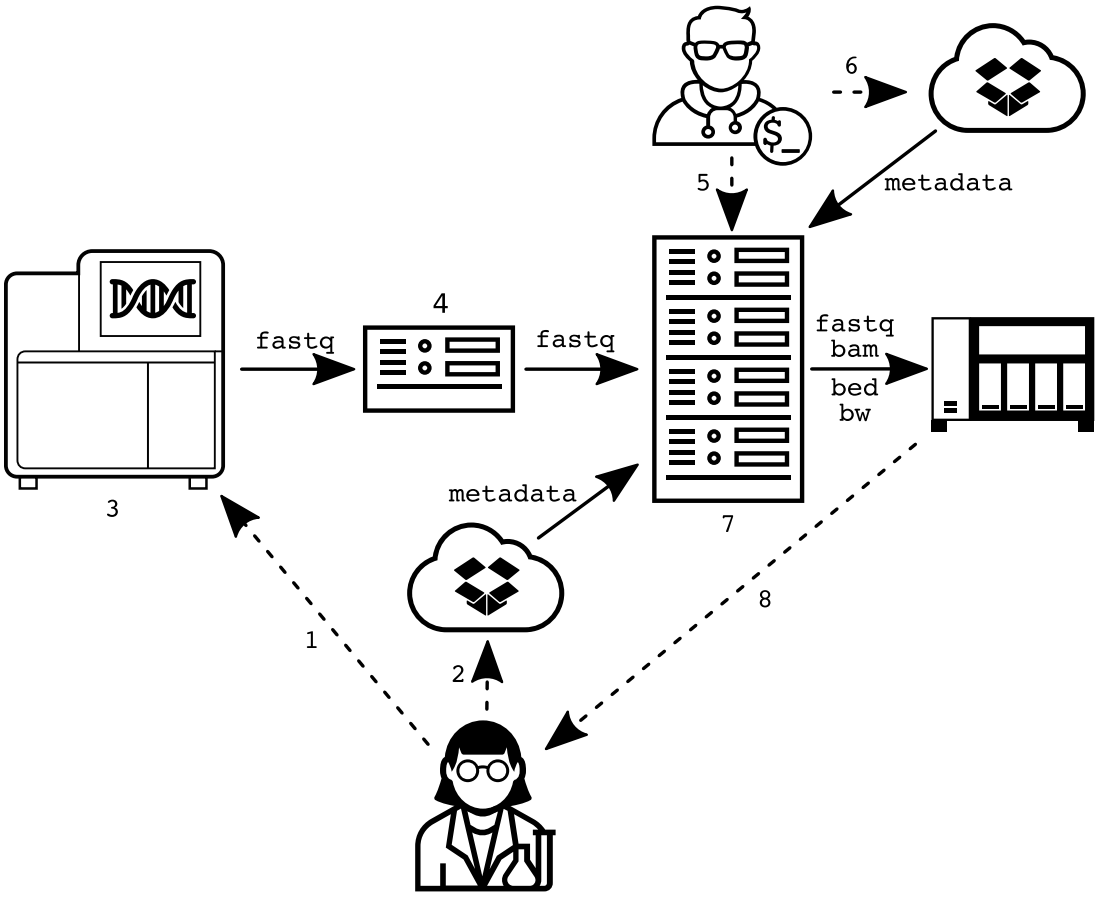


FIGURE 3.19 – Flux de travail de routine pour l'analyse de données de séquençage. Les flèches pleines correspondent à des flux de données. Les flèches en pointillées correspondent à des interactions homme-machine.

1. Un biologiste fournit ses échantillons à la plateforme de séquençage.
2. Il remplit les métadonnées conçernant ses nouveaux échantillons dans `Dropbox/Commun Labo/Sequencing_TGML/Sequencing_summary.xlsx`.
3. Un séquenceur produit des fichiers images convertis en séquences nucléotidiques sous forme de fichiers fastq.
4. Un opérateur de la plateforme met à disposition les fichiers fastq sur un serveur SFTP.
5. Un bioinformaticien transfère les fichiers sur le serveur de calcul Sacapus.
6. Il complète les chemins des fichiers de chacun des échantillons dans `Sequencing_summary.xlsx` et corrige si nécessaire les informations remplies par le biologiste.
7. Le flux de travail s'exécute automatiquement toutes les heures. Il récupère le fichier de métadonnées dans le cloud Dropbox, génère le graphe orienté acyclique correspondant, produit les nouvelles cibles si nécessaire et effectue une synchronisation des données processées sur un NAS.
8. Le biologiste peut consulter et explorer les données de son expérience.

3.2 pygtftk : Une boîte à outils pour manipuler des annotations génomiques au format GTF

Il existe plusieurs formats standards pour stocker des régions génomiques sous formes de coordonnées annotées : BED6, BED12, GTF/GFF2, GFF3. Le *General Transfer Format* (GTF) représente un équilibre entre la simplicité du format BED et la complexité du GFF3, qui permet d'encoder et d'accéder à des anotations génomiques riches (gènes, exons, ontologies, *etc.*) de manière plus efficiente qu'avec ce dernier en contrepartie de l'absence de hiérarchisation explicite des annotations. La plupart des outils existants pour manipuler des fichiers au format GTF se contentent de le convertir en un autre format, objet GRanges pour la librairie R/Bioconductor rtrackalayer (Lawrence, Gentleman et Carey 2009), ou base de données SQLite pour gffutils, ce qui offre des propriétés intéressantes mais en échange d'un coût en temps d'importation relativement important.

Durant ma thèse, j'ai eu l'opportunité de pouvoir participer à mon premier projet collaboratif de développement informatique aux côtés d'experts de la programmation en langages C et Python, respectivement Fabrice Lopez et Denis Puthier, afin de développer pygtftk, une boîte à outils en ligne de commande efficiente pour manipuler des annotations génomiques, centrée autour du format GTF. Ces outils ont fait l'objet d'une publication (Article 5 ; Lopez, Charbonnier, Kermezli et al. 2019) et ont été utilisés à de nombreuses reprises au cours des différents projets présentés dans cette thèse pour simplifier les étapes de génération et de manipulation des annotations d'intérêt pour chaque analyse.

Genome analysis

Explore, edit and leverage genomic annotations using Python GTF toolkit

F. Lopez [1,2], G. Charbonnier [1], Y. Kermezli [1,3], M. Belhocine [4], Q. Ferré [1], N. Zweig [5], M. Aribi [3], A. Gonzalez [1], S. Spicuglia [1,6], D. Puthier [1,*]

[1] Aix Marseille Univ, INSERM, UMR U1090, TAGC, Marseille, France , [2] Aix Marseille Univ, INSERM, TAGC UMR U1090, BCF-C platform, Marseille, France, [3] Tlemcen University, The Laboratory of Applied Molecular Biology and Immunology, Algeria, [4] Molecular Biology and Genetics Laboratory, Dubai, United Arab Emirates, [5] Aix Marseille Univ, [6] Equipe Labellisée LIGUE contre le Cancer.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: While Python has become very popular in bioinformatics, a limited number of libraries exist for fast manipulation of gene coordinates in Ensembl GTF format.
Results: We have developed the GTF toolkit Python package (pygtftk), which aims at providing easy and powerful manipulation of gene coordinates in GTF format. For optimal performances, the core engine of pygtftk is a C dynamic library (libgtftk) while the Python API provides usability and readability for developing scripts. Based on this Python package, we have developed the gtftk command line interface that contains 57 sub-commands (v0.9.10) to ease handling of GTF files. These commands may be used to (*i*) perform basic tasks (*e.g.* selections, insertions, updates or deletions of features/keys), (*ii*) select genes/transcripts based on various criteria (*e.g.* size, exon number, TSS location, intron length, GO terms) or (*iii*) carry out more advanced operations such as coverage analyses of genomic features using bigWig files to create faceted read-coverage diagrams. In conclusion, the pygtftk package greatly simplifies the annotation of GTF files with external information while providing advance tools to perform gene analyses.
Availability: pygtftk and gtftk have been tested on Linux and MacOSX and are available from https://github.com/dputhier/pygtftk under the MIT license. The libgtftk dynamic library written in C is available from https://github.com/dputhier/libgtftk
Contact: denis.puthier@univ-amu.fr

1 Introduction

Several formats exist to store genomic features. The standard BED format stores basic information (chromosome, start, end, name, score and strand) related to generic genomic features (BED6) or composite genomic features (BED12). The GTF/GFF2 format (thereafter referred as GTF) can describe more exhaustively defined genomic features (genes, transcripts, exons...) by taking advantage of the 'attributes' column which contains a set of keys/values to store various kinds of annotations. Some composition relationships are implicitly declared in the GTF file making it possible to describe, for instance, the exons of the transcripts corresponding to a gene. This relationship is more explicit in the GFF3 format that can be viewed as a directed acyclic graph with nodes corresponding to features (gene, transcript, exon...) and edges corresponding to part-of relationships. Only few libraries are specifically dedicated to GTFs and most of them propose very focused tasks. The GenomeTools suite is a collection of bioinformatic tools based on the libgenometools C library that handle GTF and GFF3 formats (Gremme *et al.*, 2013). However, this library extends well beyond these annotation formats and the developing framework may appear rather complicated for naive developers as it requires deep knowledge of C programming language. Regarding R/Bioconductor, the rtracklayer provides fast access to the GTF/GFF by providing the user with a GRanges object (Lawrence *et al.*, 2009).

While Python language has gained lot of popularity among bioinformaticians, only a handful of tools are available for manipulating GTF files. The gffutils package can parse and store GTF/GFF files into SQLite databases. The creation of a subsequent hierarchical models of genomic features while highly useful can be relatively time consuming. We developed the pygtftk package with the objective to provide a fast and readable way to load and manipulate GTF files within Python scripts. This package comes with the gtftk command line interface (CLI) that provide various operations to write workflows based on GTF files.

```
from pygtftk.gtf_interface import  GTF
from pygtftk.utils import get_example_file

# Create a GTF instance
cod_pot = get_example_file('mini_real', 'tab')[0]
gn_info = get_example_file('mini_real', 'genome')[0]
gtf = GTF(get_example_file('mini_real', 'gtf.gz'))

# Get a BedTool object containing the TSSs
# of the selected transcripts
tss = gtf.select_by_key('gene_biotype', 'lincRNA'
).select_by_transcript_size(min=200
).select_by_number_of_exons(min=2
).add_attr_from_file(feat='transcript',key='transcript_id',
                              has_header=True, new_key='coding_pot',
                              inputfile=cod_pot
).eval_numeric('coding_pot < 0.2', na_omit='.,?'
).get_tss(name=['transcript_id', 'gene_name', 'gene_id']
).slop(s=True, l=1000,r=1000, g=gn_info)
```

Fig. 1. Use case for the pygtftk package. These few lines of codes are used to extract the promoter region ([-1000, 1000] around the TSS) of LincRNAs, with the conditions that the transcripts have size greater than 200nt, at least two exons and a coding potential (assessed by CPAT and joined from an external file) below 0.2. (Wang et al., 2013)

2 Implementation

2.1 The core libgtftk C library

The core of the package is written in C and exposed through a dynamic library called libgtftk. The GTF format is represented without hierarchical relationships to maximize performances. More complex operations are carried out by the libgtftk Python client.

2.2 The pygtftk Python package

The GTF class of pygtftk comes with a large number of methods. Most of these methods return a new GTF object so that they can be chained intuitively. This object can also produce two additional objects from the gtftk library including: a TAB object (representation of a matrix) and a FASTA object (representation of a FASTA file). The GTF object is integrated within the scientific Python ecosystem and can produce *pybedtools.BedTool* objects, *Bio.SeqRecord* generators or a *pandas.DataFrame* (Quinlan, 2014; Cock *et al.*, 2009; McKinney, 2010). A typical use case is proposed in Figure 1 where the transcription start site (TSS) coordinates of lincRNAs are extracted with the conditions that *(i)* the transcript size is above 200nt, *(ii)* the number of exons is greater than 2 *(iii)* and the coding potential (imported from a separated file) is lower than 0.2. The TSSs are then obtained using the *get_tss()* method returning a *pybedtools.BedTool* object that can be used to extend coordinates by 1000 nucleotides in the 5' and 3' directions. Regarding performances, the human genome annotation in GTF format from Ensembl release 92 ($\sim 2.7.10^6$ lines) is loaded in about 30 seconds while the creation of a hierarchical model using gffutils takes about 11 minutes (performed on Intel(R) Xeon(R) CPU E5-2640 v3, 2.60GHz). In addition, the search engine is also highly optimized since it takes 0.6 seconds to select all lincRNAs from the human genome.

2.3 The gtftk command-line interface

The pygtftk package provides a gtftk CLI with 57 subcommands. These subcommands can be used to: *(i)* download GTF files, *(ii)* edit them, *(iii)* mine the GTF files in various ways (*e.g.* select transcripts by genomic/exonic/intronic size, number of exons, associated GO term...), *(iv)* annotate the GTF files (e.g. flagging divergent/convergent/overlapping transcripts...), *(v)* convert them to other formats or *(vi)* perform epigenomic analyses by producing faceted coverage diagrams through the plotnine Python package (*i.e.* the recently developed Python port of ggplot2).

3 Conclusion

The pygtftk package and the associated gtftk CLI provides a new way to easily handle gene coordinates with Python. They are regularly updated and users familiar with Python and/or command-line programs should quickly get comfortable and productive with (py)gtftk. As the GTF/GFF format is now also used for storing regulatory features and variants, this paves the way for future developments of (py)gtftk that could be an interesting framework for the integration of heterogeneous genomic data (Zerbino *et al.*, 2018; Reese *et al.*, 2010).

Acknowledgements

We thank Jacques van Helden for helpful discussion.

Funding

G.C. was supported by a fellowship from the "Fondation pour la Recherche Médicale" (FRM). S.S. and D.P. were supported by recurrent funding from INSERM and Aix Marseille Univ and by the Foundation for Cancer Research ARC (ARC PJA 20151203149) and A*MIDEX (ANR-11-IDEX-0001-02), Plan Cancer 2015 (C15076AS) and Ligue contre le Cancer Equipe Labellisée. Y.K., was supported, by the Franco-Algerian partenariat Hubert Curien (PHC) Tassili (15MDU935).

References

Cock, P. J. *et al.* (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11), 1422–1423.

Gremme, G. *et al.* (2013). GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans Comput Biol Bioinform*, **10**(3), 645–656.

Lawrence, M. *et al.* (2009). rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*, **25**(14), 1841–1842.

McKinney, W. (2010). Data structures for statistical computing in python. In S. van der Walt and J. Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56.

Quinlan, A. R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics*, **47**, 1–34.

Reese, M. G. *et al.* (2010). A standard variation file format for human genome sequences. *Genome Biol.*, **11**(8), R88.

Wang, L. *et al.* (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.*, **41**(6), e74.

Zerbino, D. R. *et al.* (2018). Ensembl 2018. *Nucleic Acids Res.*, **46**(D1), D754–D761.

3.3 ologram : Un outil pour modéliser la distribution des chevauchements entre des jeux de régions génomiques

La plupart des approches basées sur le séquençage à haut débit permettent la définition de régions génomiques possédant des propriétés particulières : fixation d'une protéine, accessibilité, combinaisons de modifications épigénétiques, éléments géniques ou non-géniques, *etc*. Une question récurrente et essentielle en bioinformatique génomique est de déterminer s'il existe des co-localisations significatives sur le génome entre certaines de ces propriétés qui pourraient suggérer une relation fonctionnelle entre elles.

Plusieurs approches et implémentations sont disponibles (SIMOVSKI, KANDURI, GUNDERSEN et al. 2018), et diffèrent principalement sur la façon de modéliser l'hypothèse nulle et sur le test statistique qui en découle : binomial (CEAS, *etc.* ; SHIN, LIU, MANRAI et al. 2009), hypergéométrique (Bedtools Fisher, LOLA, *etc.* ; QUINLAN 2014 ; SHEFFIELD et BOCK 2016) ou empirique (HyperBrowser, MULTOVL, *etc.* ; SANDVE, GUNDERSEN, RYDBECK et al. 2010 ; ASZÓDI 2012). Dans ce dernier cas, les régions des jeux de données à comparer sont mélangées de manière uniforme sur le génome un nombre de fois suffisant pour obtenir une distribution des chevauchements qu'il est possible d'obtenir aléatoirement. La comparaison des chevauchements obtenus avec cette distribution permet d'obtenir une P-valeur permettant de rejeter l'hypothèse nulle. Une information perdue dans le processus est la distribution des distances entres régions pour chaque jeu de données. On peut pourtant suspecter que, de par l'organisation du génome, certaines propriétés ne se répartissent pas de manière uniforme. 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 déployer et appliquer au travers de la méthodologie de recherche présentée en SECTION 3.1.

Cette problématique est particulièrement à-propos dans le cadre de la manipulation de données au format GTF, et donc de la boîte à outils pygtftk. C'est pourquoi une telle méthode conservant la distribution des distances entre régions y a été implémentée, dans la majeure partie par Quentin Ferré. Cette méthode, proposant en sus un nouveau modèle de fitting plus approprié, fait l'objet d'un article en révision (ARTICLE 6 ; FERRÉ, CHARBONNIER, SADOUNI et al. 2019), et a pu être appliquée à plusieurs cas d'études pour cette thèse.

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Application Note

OXFORD

Genome analysis

OLOGRAM : Modeling the distribution of overlap length between genomic region sets

Q. Ferré [1,2,3,†], G. Charbonnier [1,3,†], N. Sadouni [1,3], F. Lopez [1,3,4], Y. Kermezli [1,3,5], S. Spicuglia [1,3], C. Capponi [2], B. Ghattas [6], 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]Plate-forme BCF Calanques BioInformatics Core Facility, [5]Tlemcen University, The Laboratory of Applied Molecular Biology and Immunology, Algeria, [6]Aix Marseille Univ, CNRS, UMR 7373, IMM, Marseille, France.

*To whom correspondence should be addressed. †These authors contributed equally.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Various bioinformatics analyses can 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, but the scarce existing methods only assess the number of intersections without considering their lengths.
Results: Here, we introduce *OLOGRAM*, which performs overlap statistics between sets of genomic regions – described in BED or GTF. It performs 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 genomic areas during the shuffling is supported.
Availability: 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 using a statistical test to determine whether the amount of overlaps between them is attributable to chance (Haiminen *et al.*, 2008).

A comprehensive review of such methods and their implementations in a web interface are provided by the *Coloc-stats* tool (Simovski *et al.*, 2018), showing the biggest difference between them to be their choice of null model. Many approaches use a Binomial or hypergeometric test in the same manner as *GREAT* (McLean *et al.*, 2010) *CEAS* (Ji *et al.*, 2006), or *BEDTOOLS fisher*. However, this entails considering only the peak centers, not the entire peaks. Shuffling has also been used to evaluate the overlap counts, through the use of an empirical null distribution generated by random shuffling of the regions within the sets. For example, *Pybedtools* incorporates a wrapper for such an analysis (Dale *et al.*, 2011), and it was also used to tackle the N-fold overlap problem (Aszódi, 2012).

However, current approaches tend to only consider the number of overlaps, not their length. They also discard the distribution of inter-features distances in their shuffling. In this study, we propose a new method – implemented in a tool named *OLOGRAM (OverLap Of Genomic Regions Analysis using Monte Carlo)* – to conveniently assess region overlap statistics via Monte Carlo simulation.

2 Methods

2.1 Permutation and intersection computation

Let A and B be two sets of genomic intervals (*a.k.a.* regions) where no regions overlap within A or 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 (distance from last region's end to next region's beginning).

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* (Quinlan and Hall, 2010) which sets regions at random positions, disregarding the distribution of inter-region distances, and unlike our method has a limit on the number of tries to find a suitable position which can result in discarded regions.

Our approach can also exclude regions from the shuffles, for example using only regions with good mappability, or using only accessible chromatin regions. This is done by shuffling across a shorter, concatenated "sub-genome" generated by removing the excluded regions from both sets using a custom algorithm.

The algorithm then computes the regions' intersections between the i^{th} shuffle of A and the i^{th} 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

Here, we discuss a potential statistical framework to model this problem. Under (H_0) of no association between A and B, consider the Bernouilli random variables $I_{i,j} = \mathbb{1}_{A_i \cap B_j \neq \varnothing}$ which *(i)* have different but very small probabilities $p_{i,j}$ (as region size is typically small relative to chromosome size) and *(ii)* are dependent (as the regions are not independently placed, but permuted). As such, they are exchangeable within a shuffle.

Let N be the number of intesections and S the total number of overlapping nucleotides. Then $N = \sum_{i,j} I_{i,j}$ can be modeled with a Beta-Binomial (Yu and Zelterman, 2008), itself modeled with a Negative Binomial. Unlike in *BEDTOOLS shuffle*, their dependency 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 Λ_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 when the number of shuffles is large enough. Consequently, we reckon our model is plausible, and assume N and S follow Negative Binomial distributions of unknown parameters, which are approximated via this Monte Carlo approach.

3 Implementation

Our method is implemented as a plugin to *pygtftk* (Lopez *et al.*, 2019). As such, *ologram* can be passed a GTF file treated by it. Please see the documentation and Supplementary Figures for some usage examples. The relevant files are *pygtftk/pygtftk/ologram.py* and all files in *pygtftk/stats/intersect*. The package is available through Bioconda.

Most of the code is written in Python 3, with performance-critical operations multi-threaded and/or written in C++ and/or Cython (Behnel *et al.*, 2011). We use the *numpy* (Oliphant, 2006), *pandas* (McKinney, 2011) and *pybedtools* (Dale *et al.*, 2011) packages, among others. 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 "peak" 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.

With the human Ensembl GTF (release 92) and using H3K4me3 peaks from the ENCFF112BHN experiment (35358 lines) as testing data, computation time was 28 minutes on an Intel i7-7820HQ processor with 8 threads. As the computing cost scales with the number of lines in the files, we recommend they be kept of reasonable size (tens of thousands of peaks).

3.1 Results

Using our testing data, we found both S and N indeed follow a Negative Binomial distribution. 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. This principle has been used in other works (Aszódi, 2012).

Fitting a true distribution as opposed to an empirical p-value allows for better assessment of extreme overlaps likely not encountered during the shuffling. To confirm the goodness of this fit, a fitting quality is given as $1 - V$ where V is Cramér's V score (Cramér, 1946) for the contingency table of observed vs. expected histogram bins.

When compared to *BEDTOOLS shuffle*, our method results in a higher variance and slightly higher mean for both statistics of interest. We assume this is due to taking into account the distribution of inter-features distances, resulting in a wider variability for the shuffles.

An example of *OLOGRAM* output is available in *Suppl. Fig. 1* showing its pertinence. We showcase interactions with *pygtftk* in *Suppl. Fig. 2*, and the importance of considering both S and N in *Suppl. Fig. 3*.

4 Conclusion

We have implemented a method which allows to consider the information found in the number of overlapping base pairs, with a new shuffling paradigm that conserves inter-region length, used to fit a Negative Binomial model, unlike previous approaches. A number of improvements to the tool are considered for the future – detailed in code comments – most notably the use of a Markov model to produce "shuffles" under (H_0), currently offered in beta.

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)

References

Aszódi, A. (2012). MULTOVL: fast multiple overlaps of genomic regions. *Bioinformatics*, **28**(24), 3318–3319.

Behnel, S. *et al.* (2011). Cython: The best of both worlds. *Computing in Science Engineering*, **13**(2), 31 –39.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Dale, R. K. *et al.* (2011). Pybedtools: a flexible python library for manipulating genomic datasets and annotations. *Bioinformatics*.

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 Research*, **34**, W551–W554.

Lopez, F. *et al.* (2019). Explore, edit and leverage genomic annotations using python GTF toolkit. *Bioinformatics*.

McKinney, W. (2011). pandas : powerful python data analysis toolkit.

McLean, C. Y. *et al.* (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, **28**(5), 495–501.

Oliphant, T. (2006). *Guide to NumPy*.

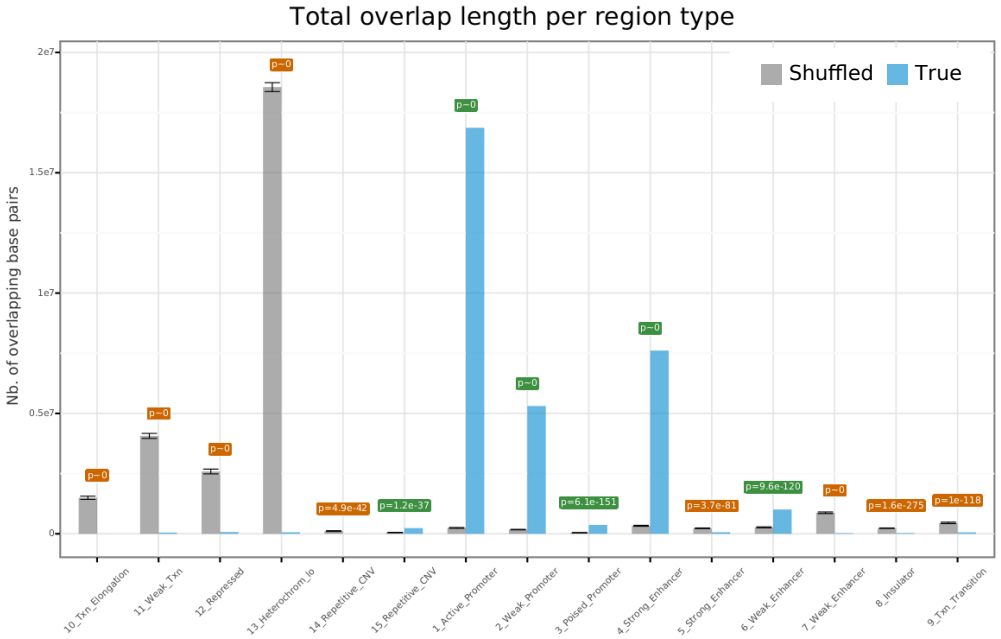
Omair, M. A. *et al.* (2018). A bivariate model based on compound negative binomial distribution. *Revista Colombiana de Estadística*, **41**(1), 87–108.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6), 841–842.

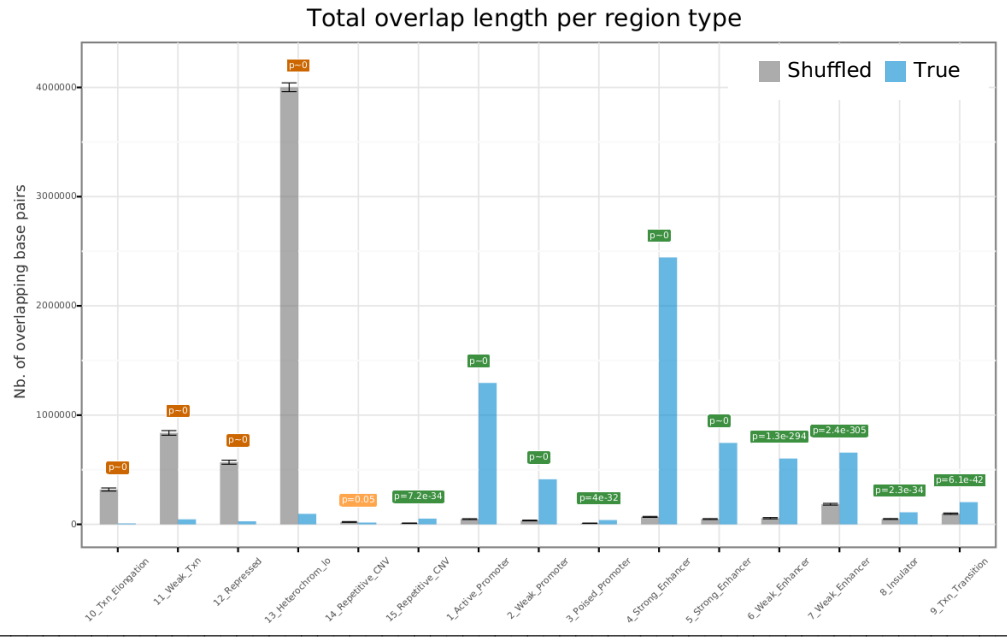
Shamos, M. I. and Hoey, D. (1976). Geometric intersection problems. In *17th Annual Symposium on Foundations of Computer Science (sfcs 1976)*, pages 208–215.

Simovski, B. *et al.* (2018). Coloc-stats: a unified web interface to perform colocalization analysis of genomic features. *Nucleic Acids Research*, **46**, W186–W193.

Yu, C. and Zelterman, D. (2008). Sums of exchangeable bernoulli random variables for family and litter frequency data. *Computational Statistics & Data Analysis*, **52**(3), 1636–1649.



A



B

Supplementary Figure 1

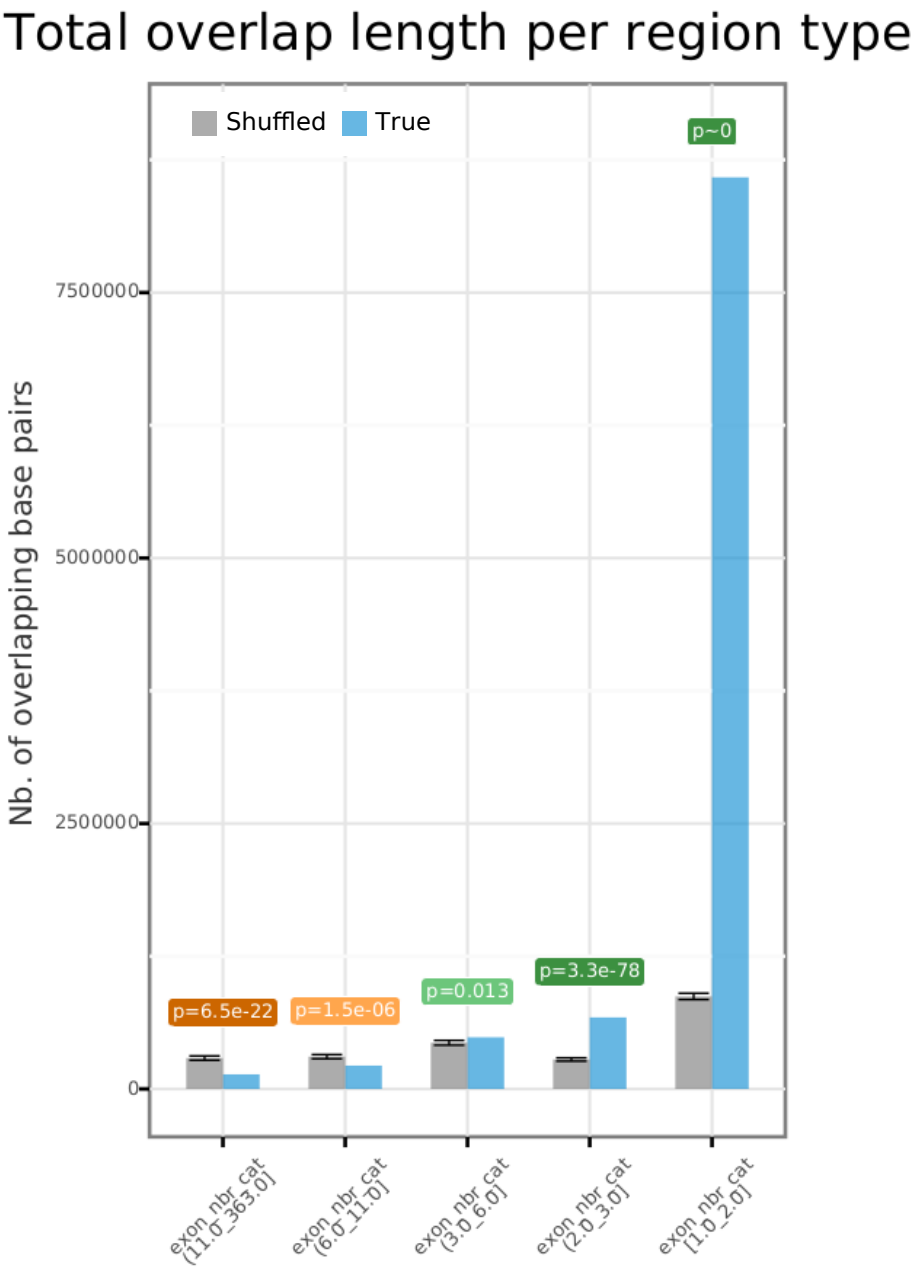
Example of OLOGRAM results, calculating the significance of intersections between :

A – H3K4me3 vs ChromHMM states

B – EP300 vs ChromHMM states

As expected, EP300 is mostly enriched in enhancer-associated states, and H3K4me3 in promoter-associated ones.

The EP300 peaks and H3K4me3 peaks come from ENCODE datasets, respectively ENCFF433PKW and ENCFF616DLO, in the K562 cell line. The ChromHMM states used are available as the wgEncodeEH000790 dataset, lifted over from hg19 to hg38.

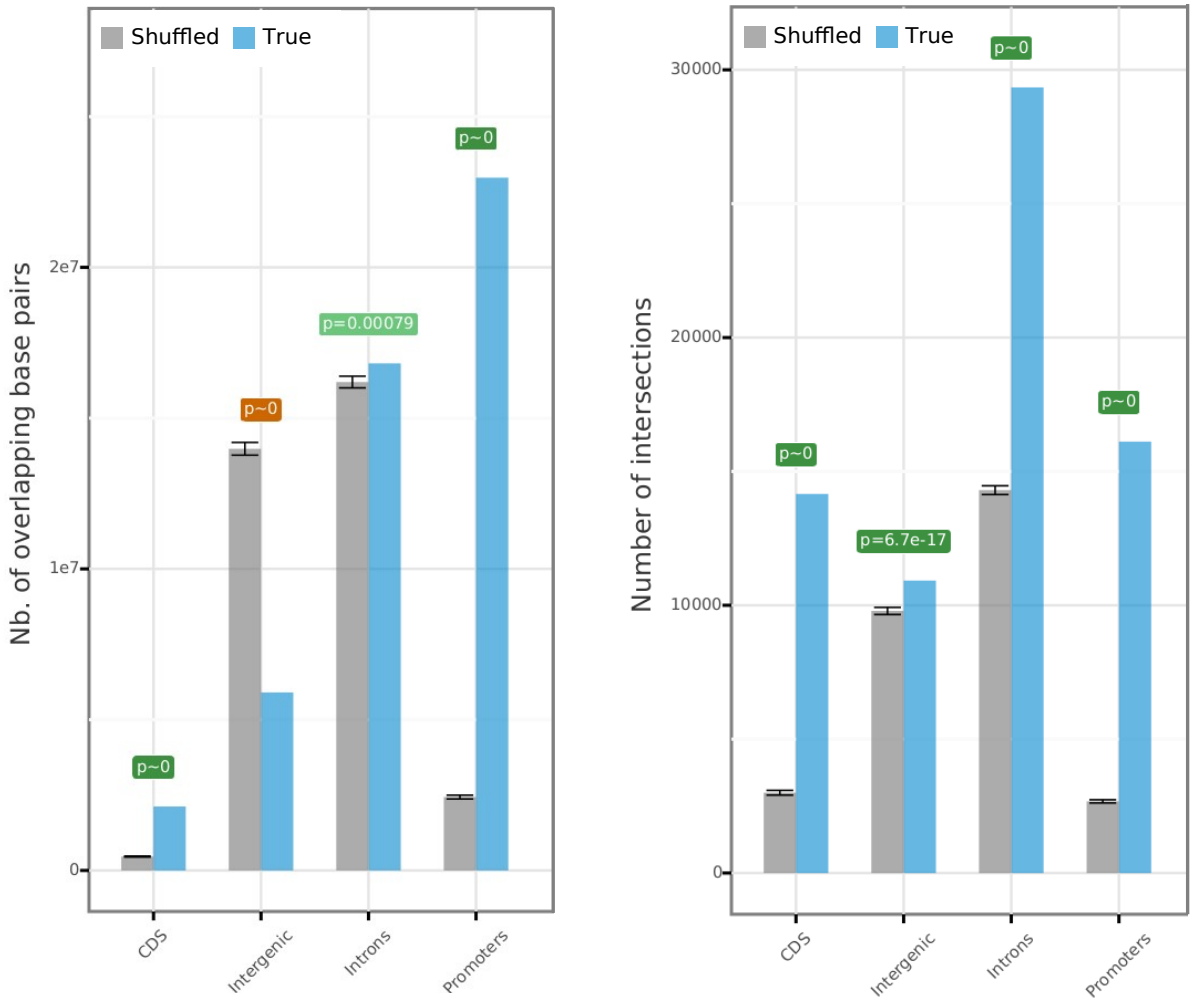


Supplementary Figure 2

Example of analysis result using OLOGRAM and providing as input a GTF treated by *pygtftk.* Here, exons have been numbered for each gene from 5' to 3'.

We calculate the significance of intersections between H3K4me3 peaks and GTF-defined numbered exons. The peaks are much more present in the first exons, likely due to the broadness of H3K4me3 peaks.

This uses the same datasets as in Supplementary Figure 1. The GTF used is the Ensembl human GTF (hg38, release 92).



Supplementary Figure 3

Further example of OLOGRAM result. Computing the significance of intersections between H3K4me3 peaks and the regions defined in the hg38 Ensembl GTF. The H3K4me3 peaks used are the same as in Suppl. Fig. 1, and the GTF the same as in Suppl. Fig. 2.

We compare the results for N (number of intersections) and S (total number of overlapping nucleotides) for a subset of GTF elements. For example, the peaks appear to be significantly enriched in introns based on N, but for S that it is not the case ; and vice-versa for intergenic regions.

Hence S is an important statistic to consider : in this particular example it may mean that the overlaps of peaks and introns are frequent but short. Note that here, an "intersection" means having at least one nucleotide in common.

3.4 greatr : Un outil pour effectuer des analyses d'enrichissement fonctionnel de régions cis-régulatrices sur plusieurs échantillons simultanément

Une application spécialisée de la problématique présentée dans la SECTION 3.3 consiste à réaliser une interprétation fonctionnelle de localisation de régions génomiques par association aux ontologies des gènes à proximité. Une façon de procéder est de les associer aux gènes les plus proches, puis d'utiliser un des nombreux outils développés initialement pour l'analyse de données de biopuces. Dans une analyse standard, la fraction de gènes annotés pour un terme d'ontologie parmi tous les gènes, et la fraction de gènes associés à ce terme parmi tous les gènes associés à l'événement étudié, sont comparés pour obtenir une P-valeur d'enrichissement.

Cette procédure possède un inconvénient majeur lorsque les événements étudiés sont distaux, *i.e.* à plus de quelques kilobases, par rapport au gène le plus proche. Elle introduit un biais de modélisation pour les gènes isolés dans le génome, tel que certains gènes développementaux, et résulte en de nombreux faux positifs (TAHER et OVCHARENKO 2009). En utilisant une approche basée sur la loi binomiale appliquée aux régions génomiques plutôt que la loi hypergéométrique appliquée aux gènes, l'outil GREAT propose un modèle statistique plus précis pour l'étude des régions cis-régulatrices (MCLEAN, BRISTOR, HILLER et al. 2010), qui en fait encore aujourd'hui sa spécificité parmi les outils disponibles dédiés à cette tâche (TABLE 3.4).

GREAT repose sur un outil accessible *via* une page internet ce qui le rend facile d'accès mais aussi fastidieux pour des applications à de nombreux échantillons. Un client sous forme de paquet R a été récemment développé pour permettre d'effectuer des requêtes *via* le terminal (GU 2019). En s'appuyant sur ce dernier, j'ai développé l'outil en ligne de commande greatr pour automatiser les requêtes pour des échantillons multiples et permettre la génération de heatmaps synthétiques pour la comparaison de plusieurs échantillons.

A partir d'une liste de fichiers de coordonnées génomiques au format BED et de l'identifiant du génome de référence de ces coordonnées (hg19, mm9, mm10, danRer7), l'outil soumet les requêtes pour chaque échantillon au serveur GREAT et récupère les tables de données et figures produites par GREAT. A celles-ci s'ajoutent les tables fusionnées pour l'ensemble des échantillons, ontologies et métriques disponibles et les heatmaps correspondantes pour les termes significatifs dans au moins un échantillon.

Cette approche génère potentiellement de larges heatmaps (FIGURE 3.20a) en fonction du nombre d'échantillons. Du fait de la redondance de nombreux termes dans les ontologies GO (FIGURE 3.20b), les termes significatifs peuvent être filtrés selon leur similarité sémantique (YU, LI, QIN et al. 2010) afin d'obtenir des

tool	DAVID & others	Enrichr	GSEA	g :Profiler	Metascape	HOMER	GeneSCF	clusterProfiler	ChIPseeker	ChIPpeakAnno	GREAT
Gene-based	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Region-based	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
Accept regions	✗	✓	✗	✓	✗	✗	✗	✗	✓	✓	✓
Web-app	✓	✓	✗	✓	✓	✗	✗	✗	✓	✗	✓
API	✓	✗	✗	✓	✗	✗	✗	✗	✓	✗	✓
Local run	✗	✗	✓	✗	✗	✓	✓	✓	✓	✓	✗
Multiple samples	✓ limited	✗	✗	✓ limited	✓ limited	✗	✗	✓ slow	✓ slow	✗	✗

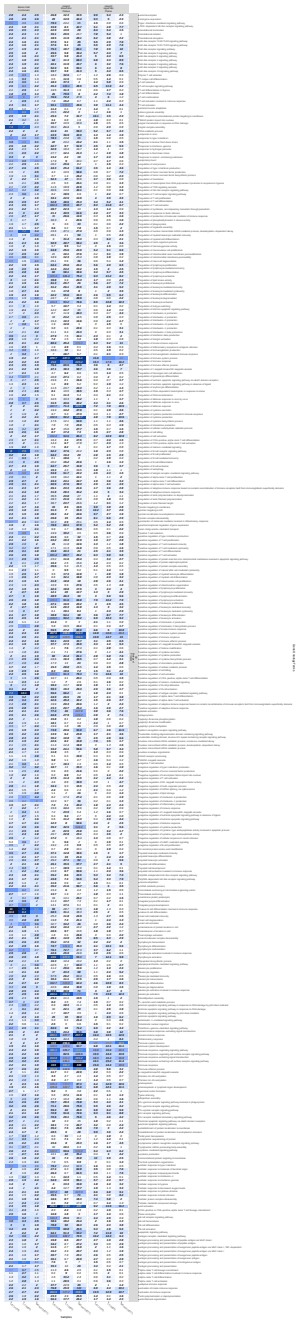
TABLE 3.4 – Propriétés des outils d'analyse d'enrichissement fonctionnel. D'autres outils tels que GORILLA, PANTHERDB, REACTOME, possèdent des propriétés similaires à DAVID ou Enrichr.

résultats plus synthétiques (FIGURES 3.20c et 3.20d). Une métrique de similarité entre termes (WANG, DU, PAYATTAKOOL et al. 2007) permet de regrouper tous les termes significatifs en groupes où seul le terme possédant le meilleur enrichissement ou la meilleure P-valeur du test binomial est retenu, ce qui permet d'obtenir respectivement des termes plus spécifiques ou génériques.

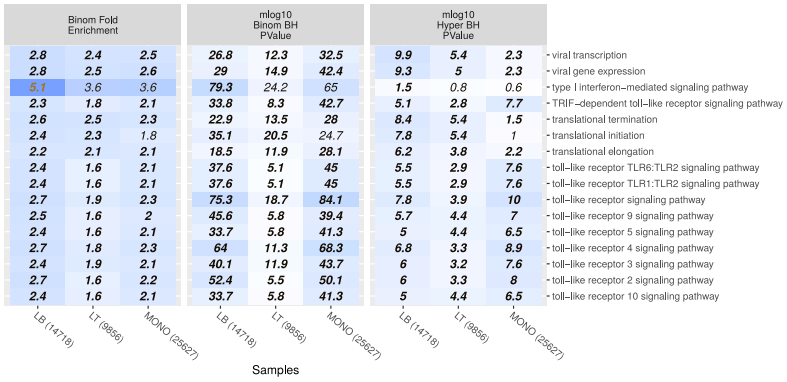
Les différents paramètres sont réglables en ligne de commande ou par un fichier de configuration yaml. Les réglages par défaut répliquent ceux du serveur GREAT (FIGURE 3.21). L'outil n'est pour l'instant disponible que *via* conda avec les dépôts bioconda et conda-forge configurés :

```
conda install -c guillaumecharbonnier r-greatr
```

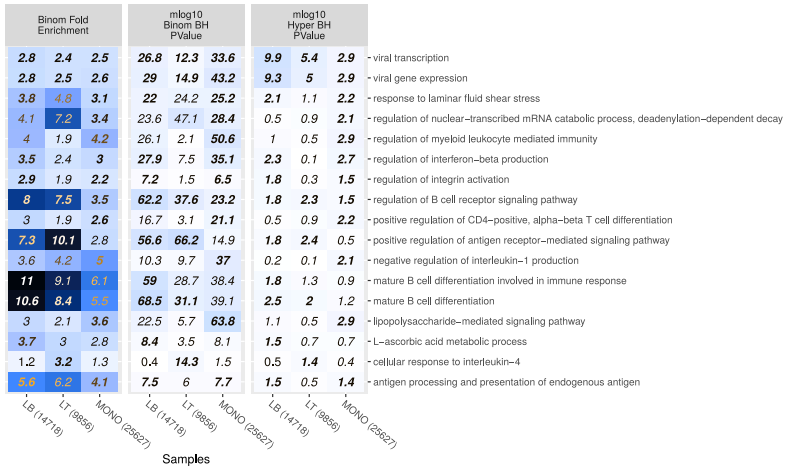
Si les génomes de référence humain et murin disponibles dans GREAT ont permis de réaliser les analyses bioinformatiques durant cette thèse, les possibilités pour les autres espèces sont limitées. Afin de disposer de l'algorithme pour des applications aux autres espèces, ainsi que pour s'affranchir du risque d'indisponibilité du serveur GREAT et de la limitation du nombre de requêtes consécutives autorisées ($\simeq$ 100 requêtes quotidiennes), l'intégration des bases de données AnnotationDbi et OrgDb, ainsi que des algorithmes utilisés par GREAT (greatTools), est envisagée dans l'outil. OLOGRAM (ARTICLE 6) est une approche qui pourrait également répondre aux mêmes problématiques que GREAT si une fonctionnalité clé en main intégrant un tel pipeline était développée. Les deux outils proposent



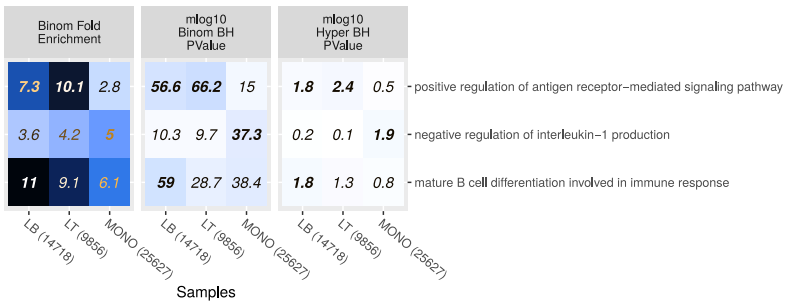
(a) Termes significatifs



(b) Premières lignes des termes significatifs



(c) Filtrage des termes significatifs par similarité sémantique



(d) Filtrage maximal par similarité sémantique

FIGURE 3.20 – Exemple de heatmaps produites par greatr pour l'ontologie GO Biological Process. Les trois échantillons sont extraits du projet présenté en SECTION 2.1 et correspondent aux régions classifiées comme *enhancers actifs* par segmentation (FIGURE S1C de l'ARTICLE 1). LB : Lymphocyte B. LT : Lymphocyte T. MONO : Monocyte. BH : Correction de Benjamini-Hochberg. La valeur entre parenthèses après le nom d'échantillon correspond au nombre de régions dans l'échantillon. Les valeurs des métriques sont indiquées en gras si le terme passe le seul de p-valeur pour le test binomial, et en italique s'il passe le test hypergéométrique. Le filtrage sémantique est rélalisé ici en retenant le terme de chaque groupe avec le meilleur enrichissement binomial.

des modélisations différentes de la problématique. Celle d'OLOGRAM est intrinsèquement plus demandeuse en ressources de calculs mais pourrait s'avérer plus juste pour certaines problématiques où l'hypothèse d'équirépartion des pics dans le génome s'avère inappropriée. Des analyses de comparaison des performances des deux outils doivent encore être réalisées. En fonction de l'issue et des optimisations qu'il est possible d'appliquer à chaque outil, greatr pourrait être intégré ou remplacé par une sous-fonction d'OLOGRAM proposant le même type d'analyses et de figures.

```
A tool to produce genomic region-based enrichment analysis for multiple samples
    relying on automated queries to GREAT.

Usage:
  greatr [-i <indir>] [-f <files>] [-b background] [-o <outdir>] [-a <assembly>] [-l <
    slimList] [-r <subsampleReplicates>] [-c collapseSamples] [-m <filterMetrics>] [-
    s <filterGreaterLowerThans>] [-t <filterThresholds>] [-l <slimList] [-y <yaml>]

Options:
  -i <indir>              Input directory. [default: .]
  -f <files>              Input sample bed files as comma-separated list. Combined
    with <indir>. If not provided, all bed files inside <indir> will be taken.
  -b <background>           An optional additional bed file containing regions not
    found in <files> to add to the background model.
  -o <outdir>             Output directory. If enrichment_tables.Rdata is present
    from a previous or interrupted run, new input files will be appended and the
    processing of existing ones will be skipped. All plots will be redrawn. [default:
    .]
  -a <assembly>           Assembly (hg19,mm9,mm10, danRer7) [default: hg19]
  -c <collapseSamples>        An optional commma-separated list of group of samples
    to merge for additional analysis. Samples from the same group should be
    semicolon-separeted, e.g. Group1Sample1;Group1Sample2,Group2Sample1.
  -r <subsampleReplicates>     Number of replicates for subsampling, no subsampling
    if unspecified.
  -m <filterMetrics>         Comma-separated list of metrics to use for filtering.
    Defaults try to match GREAT default settings. See together <
    filterGreaterLowerThans> and <filterThresholds>. [default:
    Binom_Fold_Enrichment,Binom_Adjp_BH,Hyper_Adjp_BH]
  -s <filterGreaterLowerThans>  Vector of 'lower' or 'greater' to apply
    fiterThresholds values to filterMetrics. [default: greater,lower,lower]
  -t <filterThresholds>       Vector of thresholds to apply to filterMetrics. Values
    matching threshold are kept, i.e. '<=' and '>=' are used for comparison. [default:
    2,0.05,0.05]
  -l <slimList>             An optional yaml file containing IDs to limit analysis
    for each ontology in order to keep output readable and non-redundant. Defaults
    use GO Slim Generic.
  -y <yaml>               An optional yaml file that could store all settings above.
```

FIGURE 3.21 – Documentation de greatr

Discussion et perspectives

Dynamiques épigénétiques Le projet d'étude de la thymopoïèse a tout d'abord permis de vérifier comment différentes caractéristiques épigénétiques permettent de discriminer des sous-populations d'une voie de différentiation. Sur des projections bidimensionnelles du paysage épigénétique, les sous-populations thymiques se positionnent sur une trajectoire cohérente relativement aux étapes de différentiation, entre les cellules souches hématopoïétiques et les lymphocytes T périphériques. Ce résultat corrobore des résultats similaires obtenus sur d'autres voies de différentiation sur la base du méthylome (KULIS, MERKEL, HEATH et al. 2015 ; SCHUYLER, MERKEL, RAINERI et al. 2016) et de combinaison de marques d'histones (SANTA-PAU, JUAN, PANCALDI et al. 2017).

La différentiation des cellules T est accompagnée d'un remodelage important des modifications d'histones impliquant une fermeture progressive de la majorité de la chromatine, reflétant le modèle classique de la différentiation cellulaire (WADDINGTON 1957). Les modalités de fermeture de la chromatine n'ont pas été le centre d'intérêt de l'analyse réalisée mais pourraient être explorées davantage à partir des données disponibles, notamment en analysant la part d'hétérochromatine formée que l'on peut associer à un mécanisme de répression du groupe Polycomb en utilisant H3K27me3 comme marqueur (SANULLI, JUSTIN, TEISSANDIER et al. 2015).

L'étude des régions régulatrices distales a révélé qu'une majorité d'entre elles sont constitutivement hypométhylées indépendamment de leur statut d'activation, ce qui suggère que l'hypométhylation de l'ADN pourrait être utilisée comme marqueur d'amplificateurs putatifs avant ou après activation. Parmi eux, l'amplificateur du locus du TCRA est trouvé dans un état hypométhylé et accessible avant l'activation du locus du *TCRA*. La démonstration par une combinaison d'approches bioinformatiques et biochimiques du rôle des protéines HOXA en tant que répresseur de cet amplificateur révèle leur contribution à l'engagement dans une voie de différentiation $\gamma\delta$ plutôt que $\alpha\beta$ (ROBEY et FOWLKES 1998).

Le projet d'étude de la spermatogénèse a tout d'abord mis en évidence la coexistence de l'acétylation et de la butyrylation sur les lysines 5 et 8 des TSS des gènes actifs. Bien que ces modifications soient mutuellement exclusives sur chaque résidu, le déroulement attendu du protocole de ChIP-seq permet d'expliquer ce résultat. Puisque les immunoprécipitations de chromatine sont réalisées sur des groupes de cellules, les enrichissements localisés correspondent à la moyenne de combinaisons de modifications différentes sur les cellules traitées. Le modèle d'échange dynamique entre ces modifications, suggéré par les approches biochimiques, et déjà observé pour d'autres modifications d'histones (CLAYTON, HAZZALIN et MAHADEVAN 2006 ; CRUMP, HAZZALIN, BOWERS et al. 2011), est alors compatible avec les résultats bioinformatiques.

Un point non étudié au moment de la réalisation de cette analyse est l'étude de la répartition de ces modifications sur des éléments régulateurs distaux. La quan-

tification de modifications épigénétiques au niveau de régions géniques discrètes (TSS, 5'UTR, *etc.*) est une analyse bioinformatique triviale réalisable à partir de la manipulation d'un échantillon et d'un fichier d'annotations. Les éléments régulateurs distaux sont dynamiques et souvent spécifiques d'un type cellulaire. Leur prédiction nécessite l'intégration de différentes sources de données issues de plusieurs approches de séquençage à haut débit ce qui rend plus complexe leur association avec une nouvelle modification épigénétique. Les approches appliquées au projet d'étude de la thymopoïèse pourraient être réutilisées pour vérifier comment ces marques sont distribuées dans les régions distales sachant qu'environ trois quarts d'entre elles se retrouvent à plus de 5 kb du TSS le plus proche.

Cette section contient des résultats confidentiels au moment de la soutenance.

Méthodologie de recherche reproductible Les différentes analyses de données et développements logiciels effectués au cours de cette thèse ont permis de définir une méthodologie de recherche permettant d'assurer la reproductibilité, l'automatisation, le partage, l'archivage, le développement simplifié, l'optimisation et la réutilisation d'analyses complexes de données volumineuses et hétérogènes. Lorsque certaines approches analytiques n'étaient pas disponibles pour un traitement reproductible automatisé en ligne de commande, des outils ont été développés pour combler ces limitations.

Pour chacune des thématiques biologiques explorées, des milliers de fichiers d'entrées de données brutes ou pré-traitées ont été intégrés à partir de sources diverses afin de produire les figures ayant permis les interprétations biologiques présentées dans ce manuscrit. Plus de 150 outils en ligne de commande ont été utilisés à au moins une étape du processus d'analyse et ont généré des dizaines de milliers de fichiers. Grâce à la méthodologie mise au point et dérivée de Snakemake, la manière exacte de générer chacun de ces fichiers est définie par une bijection entre son chemin de fichier et l'enchaînement des outils, paramètres et données nécessaires à sa production de manière automatisée. De ce principe découle l'idée que le chemin de fichier agit comme un identifiant unique explicite qui peut être conservé pour l'archivage et le partage de traitement de données à la place du fichier traité lui-même potentiellement volumineux. Pour cette raison, l'ensemble des fichiers produits, ou à produire, par virtuellement n'importe quelle analyse possible sont déjà rangées et organisées de manière non ambigüe dans l'arborescence des fichiers de sortie du flux de travail. Pour les outils d'analyses qui n'ont pas encore été intégrés dans le flux de travail, l'emplacement de leurs produits est fixé par les conventions quasi-algorithmiques de la méthodologie, ce qui limite les errements de nommage dans la phase d'écriture des règles Snakemake. Ceci simplifie grandement et optimise le travail de développement, comparaison et optimisation de traitements alternatifs de données pour le bioinformaticien. Cela fournit également une mécanique pour assurer naturellement la traçabilité de tous les fichiers intégrés dans le flux de travail, puisque les étapes de récupération, de prétraitement et d'homogénéisation d'échantillons

de sources diverses peuvent être directement incluses comme règles dans le flux de travail automatisé.

Cette approche bijective diverge de la méthodologie de référence prônée par les contributeurs principaux de Snakemake. Dans celle-ci, un fichier de sortie du flux de travail peut avoir été produit de plusieurs manières en fonction des ajustements d'un fichier de configuration yaml du flux de travail. L'intérêt d'une telle approche est de rendre plus accessible le flux de travail à un utilisateur néophyte, grâce à une arborescence de fichiers de sortie simplifiée dont les dossiers possèdent des noms de propriétés usuelles connues et partagées par le bioinformaticien et le biologiste. Par exemple, les fichiers de fragments alignés sur le génome de référence peuvent se retrouver dans un dossier `aligned-reads` plutôt que nécessairement disséminés dans des dossiers `bowtie`, `star`, `bwa` ou autre nom d'aligneur. En échange, de son accessibilité pour le biologiste, cette approche limite la simplicité pour le développeur de flux de travail. Puisque ces deux approches ne s'adressent pas aux mêmes utilisateurs, elles possèdent des champs d'applications différents, bien que parfois chevauchants, et sont donc complémentaires. À ce titre, il est relativement aisé d'intégrer des flux de travail analyse-spécifique dans la philosophie de metaworkflow puisque cela consiste surtout à ajuster les patterns de fichiers d'entrées et de sorties de chaque règle. Plusieurs flux de travail développés par d'autres bioinformaticiens (analyse Capstarr-seq d'Aurélien Griffon et RNA-seq de Denis Puthier) ont ainsi été intégrés dans metaworkflow. Le processus inverse devrait permettre à des bioinformaticiens d'extraire des portions de metaworkflow afin de produire des flux de travail analyse-spécifique en Snakemake, ou basés sur d'autres gestionnaires de flux de travail selon les préférences personnelles des membres de chaque projet.

Une bonne connaissance des modèles biologiques d'études et des approches analytiques existantes sont des prérequis pour l'étude efficace de mécanismes épigénétiques car cela permet d'une part de poser des hypothèses pertinentes, d'autre part de pouvoir les vérifier rigoureusement. Les données traitées ont permis de supporter ou contrarier de nombreuses hypothèses mais pourraient aussi permettre de répondre à d'autres questions si le centre d'intérêt des projets de recherche était différent ou venait à évoluer, par exemple suite à la caractérisation d'un nouvel acteur ou mécanisme cellulaire. Des approches et algorithmes continuent d'être développées et pourraient fournir une vision différente des mécanismes étudiés en apportant une amélioration d'une ou plusieurs étapes de traitement. Reprendre une analyse bioinformatique pour intégrer de nouveaux outils peut être aussi lourd que de refaire l'ensemble des étapes la composant. Cette lourdeur incite souvent à ne pas remettre en cause ses propres résultats et approches. Une méthodologie de recherche entièrement reproductible et automatisée brille dans ce cas de figure en allégeant grandement ce processus. Il existe cependant deux situations dans lesquelles cette méthodologie ne peut s'appliquer. Premièrement pour l'édition finale des sous-figures d'articles qui doivent être assemblés et arrangés en panels homogènes. A défaut de mieux, il est important dans ce cas de conserver scrupuleusement les chemins de fichiers des

sous-figures afin de préserver leur traçabilité. Deuxièmement si des collaborateurs n'ont pas connaissance de la méthodologie ou la volonté de l'utiliser. En poursuivant la documentation de l'approche metaworkflow initiée dans ce manuscrit, j'espère pouvoir sensibiliser et rallier de nouveaux collaborateurs à ces préceptes ou au moins avoir fourni au lecteur les bases nécessaires à la compréhension, la réplication, la réutilisation mais aussi la critique et l'amélioration de la méthodologie et des approches analytiques appliquées lors de mon travail de recherche.

Bibliographie

[ANS+13] Nicholas L. Adkins, Hengyao Niu, Patrick Sung et al. « Nucleosome dynamics regulates DNA processing ». In : *Nature Structural and Molecular Biology* 20.7 (2013), p. 836-842. ISSN : 15459993. DOI : 10.1038/nsmb.2585 (cf. p. 30).

[ABC+07] C. David Allis, Shelley L. Berger, Jacques Cote et al. « New Nomenclature for Chromatin-Modifying Enzymes ». In : *Cell* 131.4 (2007), p. 633-636. ISSN : 00928674. DOI : 10.1016/j.cell.2007.10.039 (cf. p. 34).

[AAC+09] Mohammed Altaf, Andréanne Auger, Marcela Covic et al. « Connection between histone H2A variants and chromatin remodeling complexes ». In : *Biochemistry and Cell Biology* 87.1 (2009), p. 35-50. ISSN : 0829-8211. DOI : 10.1139/O08-140 (cf. p. 30).

[ACM+18] Adam Ameur, Huiwen Che, Marcel Martin et al. « De novo assembly of two swedish genomes reveals missing segments from the human GRCh38 reference and improves variant calling of population-scale sequencing data ». In : *Genes* 9.10 (2018), p. 1-16. ISSN : 20734425. DOI : 10.3390/genes9100486 (cf. p. 66).

[AGME+14] Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada et al. « An atlas of active enhancers across human cell types and tissues ». In : *Nature* 507.7493 (2014), p. 455-461. ISSN : 14764687. DOI : 10.1038/nature12787 (cf. p. 66).

[AGS04] Oscar Aparicio, Joseph V. Geisberg et Kevin Struhl. « Chromatin Immunoprecipitation for Determining the Association of Proteins with Specific Genomic Sequences In Vivo ». In : *Current Protocols in Molecular Biology* 65.1 (2004), p. 21.3.1-21.3.23. ISSN : 19343639. DOI : 10.1002/0471142727.mb2103s65 (cf. p. 59).

[ABA+09] L. Aravind, H. Bandukwala, S. Agarwal et al. « Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1 ». In : *Science* 324.5929 (2009), p. 930-935. ISSN : 0036-8075. DOI : 10.1126/science.1170116 (cf. p. 28).

[Asz12] András Aszódi. « MULTOVL : Fast multiple overlaps of genomic regions ». In : *Bioinformatics* 28.24 (2012), p. 3318-3319. ISSN : 13674803. DOI : 10.1093/bioinformatics/bts607 (cf. p. 214).

[AAA+15] Adam Auton, Gonçalo R. Abecasis, David M. Altshuler et al. « A global reference for human genetic variation ». In : *Nature* 526.7571 (2015), p. 68-74. ISSN : 0028-0836. DOI : 10.1038/nature15393 (cf. p. 65).

[BHR+99] Willy M Baarends, Jos W Hoogerbrugge, Henk P Roest et al. « Histone Ubiquitination and Chromatin Remodeling in Mouse Spermatogenesis ». In : *Developmental Biology* 207.2 (1999), p. 322-333. ISSN : 00121606. DOI : 10.1006/dbio.1998.9155 (cf. p. 54, 55).

[Bae13] Jae-Bum Bae. « Perspectives of International Human Epigenome Consortium ». In : *Genomics & Informatics* 11.1 (2013), p. 7. ISSN : 1598-866X. DOI : 10.5808/gi.2013.11.1.7 (cf. p. 69).

[BP16] Monya Baker et Dan Penny. « Is there a reproducibility crisis ? » In : *Nature* 533.7604 (2016), p. 452-454. ISSN : 14764687. DOI : 10.1038/533452A (cf. p. 72).

[BB16] Jianqiang Bao et Mark T Bedford. « Epigenetic regulation of the histone-to-protamine transition during spermiogenesis ». In : *Reproduction* 151.5 (2016), R55-R70. ISSN : 1470-1626. DOI : 10.1530/REP-15-0562 (cf. p. 55).

[BF18] V. Barra et D. Fachinetti. « The dark side of centromeres : types, causes and consequences of structural abnormalities implicating centromeric DNA ». In : *Nature Communications* 9.1 (2018). ISSN : 20411723. DOI : 10.1038/s41467-018-06545-y (cf. p. 26).

[BMH+17] Sophie Barral, Yuichi Morozumi, Naghmeh Hoghoughi et al. « Le mystère de la disparition des histones au cours de la spermatogenèse ». In : *médecine/sciences* 33.67 (2017), p. 588-590. ISSN : 0767-0974. DOI : 10.1051/medsci/20173306010 (cf. p. 54).

[BMT+17] Sophie Barral, Yuichi Morozumi, Hiroki Tanaka et al. « Histone Variant H2A.L.2 Guides Transition Protein-Dependent Protamine Assembly in Male Germ Cells ». In : *Molecular Cell* 66.1 (2017), 89-101.e8. ISSN : 10974164. DOI : 10.1016/j.molcel.2017.02.025 (cf. p. 60, 155).

[BCC+07] Artem Barski, Suresh Cuddapah, Kairong Cui et al. « High-Resolution Profiling of Histone Methylations in the Human Genome ». In : *Cell* 129.4 (2007), p. 823-837. ISSN : 00928674. DOI : 10.1016/j.cell.2007.05.009 (cf. p. 58, 59).

[BTW+03] C. H. Bassing, R. E. Tillman, B. B. Woodman et al. « T cell receptor (TCR) / locus enhancer identity and position are critical for the assembly of TCR and variable region genes ». In : *Proceedings of the National Academy of Sciences* 100.5 (2003), p. 2598-2603. ISSN : 0027-8424. DOI : 10.1073/pnas.0437943100 (cf. p. 45).

[BS14] Tuncay Baubec et Dirk Schübeler. « Genomic patterns and context specific interpretation of DNA methylation ». In : *Current Opinion in Genetics and Development* 25.1 (2014), p. 85-92. ISSN : 18790380. DOI : 10.1016/j.gde.2013.11.015 (cf. p. 28).

[BBD+13] Sean Bechhofer, Iain Buchan, David De Roure et al. « Why linked data is not enough for scientists ». In : *Future Generation Computer Systems* 29.2 (2013), p. 599-611. ISSN : 0167739X. DOI : 10.1016/j.future.2011.08.004 (cf. p. 83).

[BOW99] Stephan Beck, Alexander Olek et Jörn Walter. *From genomics to epigenomics : A loftier view of life*. 1999. DOI : 10.1038/70651 (cf. p. 69).

[Bel16] Christopher G. Bell. « Insights in human epigenomic dynamics through comparative primate analysis ». In : *Genomics* 108.3-4 (2016), p. 115-125. ISSN : 08887543. DOI : 10.1016/j.ygeno.2016.09.003 (cf. p. 37).

[BS12] Yuval Benjamini et Terence P. Speed. « Summarizing and correcting the GC content bias in high-throughput sequencing ». In : *Nucleic Acids Research* 40.10 (2012), e72-e72. ISSN : 1362-4962. DOI : 10.1093/nar/gks001 (cf. p. 59).

[BCC+97] A Bessis, N Champtiaux, L Chatelin et al. « The neuron-restrictive silencer element : a dual enhancer/silencer crucial for patterned expression of a nicotinic receptor gene in the brain. » In : *Proceedings of the National Academy of Sciences of the United States of America* 94.11 (1997), p. 5906-11. ISSN : 0027-8424 (cf. p. 24).

[BOB+13] Michael J. Booth, Tobias W B Ost, Dario Beraldi et al. « Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine ». In : *Nature Protocols* 8.10 (2013), p. 1841-1851. ISSN : 1754-2189. DOI : 10.1038/nprot.2013.115 (cf. p. 62).

[BBG+18] Guillaume Bourque, Kathleen H. Burns, Mary Gehring et al. « Ten things you should know about transposable elements ». In : *Genome biology* 19.1 (2018), p. 199. ISSN : 1474760X. DOI : 10.1186/s13059-018-1577-z (cf. p. 26).

[BB86] P Boursot et F Bonhomme. « Génétique et évolution du génome mitochondrial des Métazoaires ». In : *Genetics Selection Evolution* 18.1 (1986), p. 73. ISSN : 1297-9686. DOI : 10.1186/1297-9686-18-1-73 (cf. p. 14).

[BDS+08] Alan P. Boyle, Sean Davis, Hennady P. Shulha et al. « High-Resolution Mapping and Characterization of Open Chromatin across the Genome ». In : *Cell* 132.2 (2008), p. 311-322. ISSN : 00928674. DOI : 10.1016/j.cell.2007.12.014 (cf. p. 61).

[BPM+16] Nicolas L Bray, Harold Pimentel, Páll Melsted et al. « Near-optimal probabilistic RNA-seq quantification ». In : *Nature Biotechnology* 34.5 (2016), p. 525-527. ISSN : 1087-0156. DOI : 10.1038/nbt.3519 (cf. p. 58).

[BL15] Achim Breiling et Frank Lyko. « Epigenetic regulatory functions of DNA modifications : 5-methylcytosine and beyond ». In : *Epigenetics and Chromatin* 8.1 (2015), p. 1-9. ISSN : 17568935. DOI : 10.1186/s13072-015-0016-6 (cf. p. 28, 29).

[BGZ+13] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba et al. « Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position ». In : *Nature Methods* 10.12 (2013), p. 1213-1218. ISSN : 1548-7091. DOI : 10.1038/nmeth.2688 (cf. p. 61-63).

[BZ13] Rebecca J. Burgess et Zhiguo Zhang. « Histone chaperones in nucleosome assembly and human disease ». In : *Nature Structural and Molecular Biology* 20.1 (2013), p. 14-22. ISSN : 15459993. DOI : 10.1038/nsmb.2461 (cf. p. 30).

[But02] Jennifer E.F. Butler. « The RNA polymerase II core promoter : a key component in the regulation of gene expression ». In : *Genes & Development* 16.20 (2002), p. 2583-2592. ISSN : 08909369. DOI : 10.1101/gad.1026202 (cf. p. 21, 23).

[CDD+19] S. Carbon, E. Douglass, N. Dunn et al. « The Gene Ontology Resource : 20 years and still GOing strong ». In : *Nucleic Acids Research* 47.D1 (2019), p. D330-D338. ISSN : 0305-1048. DOI : 10.1093/nar/gky1055 (cf. p. 67).

[CBP+15] Christelle Cayrou, Benoit Ballester, Isabelle Peiffer et al. « The chromatin environment shapes DNA replication origin organization and defines origin classes ». In : *Genome Research* 25.12 (2015), p. 1873-1885. ISSN : 15495469. DOI : 10.1101/gr.192799.115 (cf. p. 26).

[Cha01] B. P. Chadwick. « Histone variant macroH2A contains two distinct macrochromatin domains capable of directing macroH2A to the inactive X chromosome ». In : *Nucleic Acids Research* 29.13 (2001), p. 2699-2705. ISSN : 13624962. DOI : 10.1093/nar/29.13.2699 (cf. p. 30).

[CMD+92] Muriel Chauvière, Arlette Martinage, Michel Debarle et al. « Molecular characterization of six intermediate proteins in the processing of mouse protamine P2 precursor ». In : *European Journal of Biochemistry* 204.2 (1992), p. 759-765. ISSN : 0014-2956. DOI : 10.1111/j.1432-1033.1992.tb16691.x (cf. p. 54).

[CXP+13] Kaifu Chen, Yuanxin Xi, Xuewen Pan et al. « DANPOS : Dynamic analysis of nucleosome position and occupancy by sequencing ». In : *Genome Research* 23.2 (2013), p. 341-351. ISSN : 1088-9051. DOI : 10.1101/gr.142067.112 (cf. p. 60).

[CTC+18] Fabio Ciccarone, Stefano Tagliatesta, Paola Caiafa et al. *DNA methylation dynamics in aging : how far are we from understanding the mechanisms ?* 2018. DOI : 10.1016/j.mad.2017.12.002 (cf. p. 36).

[CCT+19] Agata Cieslak, Guillaume Charbonnier, Melania Tesio et al. « Epigenomic dynamics of human T-cell precursors reveal TCRA enhancer silencing by HOXA ». In : *In preparation* (2019), p. 10-12 (cf. p. 87).

[CK92] Jon F. Claerbout et Martin Karrenbach. « Electronic documents give reproducible research a new meaning ». In : *SEG Technical Program Expanded Abstracts 1992*. Society of Exploration Geophysicists, 1992, p. 601-604. DOI : 10.1190/1.1822162 (cf. p. 72).

[CC09] Cedric R. Clapier et Bradley R. Cairns. « The Biology of Chromatin Remodeling Complexes ». In : *Annual Review of Biochemistry* 78.1 (2009), p. 273-304. ISSN : 0066-4154. DOI : 10.1146/annurev.biochem.77.062706.153223 (cf. p. 30).

[CHM06] Alison L. Clayton, Catherine A. Hazzalin et Louis C. Mahadevan. *Enhanced Histone Acetylation and Transcription : A Dynamic Perspective*. 2006. DOI : 10.1016/j.molcel.2006.06.017 (cf. p. 226).

[CEG+18] Levin Clément, Dynomant Emeric, Bruno Gonzalez et al. « A data-supported history of bioinformatics tools ». In : (2018), p. 1-14. arXiv : 1807.06808 (cf. p. 71).

[CD91] Robert Mullan Cook-Deegan. *The human genome project : the formation of federal policies in the United States, 19861990*. 1991, p. 99-167. ISBN : 0-309-59757-9 (cf. p. 13).

[CP98] Carl Costanzi et John R Pehrson. « Histone macroH2A1 is concentrated in the inactive X chromosome of female mammals ». In : *Nature* 393.6685 (1998), p. 599-601. ISSN : 00280836. DOI : 10.1038/31275 (cf. p. 30).

[CHB+11] N. T. Crump, C. A. Hazzalin, E. M. Bowers et al. « Dynamic acetylation of all lysine-4 trimethylated histone H3 is evolutionarily conserved and mediated by p300/CBP ». In : *Proceedings of the National Academy of Sciences* 108.19 (2011), p. 7814-7819. ISSN : 0027-8424. DOI : 10.1073/pnas.1100099108 (cf. p. 226).

[DPM+14] Lunzhi Dai, Chao Peng, Emilie Montellier et al. « Lysine 2-hydroxyisobutyrylation is a widely distributed active histone mark ». In : *Nature Chemical Biology* 10.5 (2014), p. 365-370. ISSN : 15524469. DOI : 10.1038/nchembio.1497 (cf. p. 141).

[DGS+18] Ryan Dale, Björn Grüning, Andreas Sjödin et al. « Bioconda : Sustainable and comprehensive software distribution for the life sciences ». In : *Nature Methods* 15.7 (2018), p. 475-476. ISSN : 15487105. DOI : 10.1038/s41592-018-0046-7 (cf. p. 80).

[DGACM+17] L.T.M. Dao, A.O. Galindo-Albarrán, J.A. Castro-Mondragon et al. « Genome-wide characterization of mammalian promoters with distal enhancer functions ». In : *Nature Genetics* 49.7 (2017). ISSN : 15461718. DOI : 10.1038/ng.3884 (cf. p. 24, 256).

[DNB+11] Alberto Darszon, Takuya Nishigaki, Carmen Beltran et al. « Calcium Channels in the Development, Maturation, and Function of Spermatozoa ». In : *Physiological Reviews* 91.4 (2011), p. 1305-1355. ISSN : 0031-9333. DOI : 10.1152/physrev.00028.2010 (cf. p. 47).

[DTC10] Chandrima Das, Jessica K. Tyler et Mair E A Churchill. « The histone shuffle : Histone chaperones in an energetic dance ». In : *Trends in Biochemical Sciences* 35.9 (2010), p. 476-489. ISSN : 09680004. DOI : 10.1016/j.tibs.2010.04.001 (cf. p. 30).

[DB10] John L. Davey et Mark W. Blaxter. « RADseq : Next-generation population genetics ». In : *Briefings in Functional Genomics* 9.5-6 (2010), p. 416-423. ISSN : 20412649. DOI : 10.1093/bfgp/elq031 (cf. p. 57).

[DSS+19] Daniel P. Depledge, Kalanghad Puthankalam Srinivas, Tomohiko Sadaoka et al. « Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen ». In : *Nature Communications* 10.1 (2019), p. 754. ISSN : 2041-1723. DOI : 10.1038/s41467-019-08734-9 (cf. p. 58).

[DMK+15] Brian G. Dias, Stephanie A. Maddox, Torsten Klengel et al. « Epigenetic mechanisms underlying learning and the inheritance of learned behaviors ». In : *Trends in Neurosciences* 38.2 (2015), p. 96-107. ISSN : 1878108X. DOI : 10.1016/j.tins.2014.12.003 (cf. p. 36).

[DLL+15] Guo-Lian Ding, Ye Liu, Miao-E Liu et al. « The effects of diabetes on male fertility and epigenetic regulation during spermatogenesis ». In : *Asian Journal of Andrology* 17.6 (2015), p. 948. ISSN : 1008-682X. DOI : 10.4103/1008-682x.150844 (cf. p. 55).

[DDS+13] Alexander Dobin, Carrie A. Davis, Felix Schlesinger et al. « STAR : ultrafast universal RNA-seq aligner ». In : *Bioinformatics* 29.1 (2013), p. 15-21. ISSN : 1460-2059. DOI : 10.1093/bioinformatics/bts635 (cf. p. 58).

[Dul10] Catherine Dulac. « Brain function and chromatin plasticity ». In : *Nature* 465.7299 (2010), p. 728-735. ISSN : 00280836. DOI : 10.1038/nature09231 (cf. p. 27).

[DKA+12] Ian Dunham, Anshul Kundaje, Shelley F. Aldred et al. « An integrated encyclopedia of DNA elements in the human genome ». In : *Nature* 489.7414 (2012), p. 57-74. ISSN : 0028-0836. DOI : 10.1038/nature11247 (cf. p. 66).

[DCJ+15] Stephanie O.M. Dyke, Warren A. Cheung, Yann Joly et al. « Epigenome data release : A participant-centered approach to privacy protection ». In : *Genome Biology* 16.1 (2015), p. 1-12. ISSN : 1474760X. DOI : 10.1186/s13059-015-0723-0 (cf. p. 69).

[EAS+17] Sara El Kennani, Annie Adrait, Alexey K. Shaytan et al. « MS-HistoneDB, a manually curated resource for proteomic analysis of human and mouse histones ». In : *Epigenetics and Chromatin* 10.1 (2017), p. 1-18. ISSN : 17568935. DOI : 10.1186/s13072-016-0109-x (cf. p. 30, 33).

[FAN01] FANTOM Consortium. « Functional annotation of a full-length mouse cDNA collection ». In : *Nature* 13.February (2001), p. 685-690 (cf. p. 66, 67).

[FCK+12] Romain Fenouil, Pierre Cauchy, Frederic Koch et al. « CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. » In : *Genome Research* 22.12 (2012), p. 2399-408. ISSN : 1549-5469. DOI : 10.1101/gr.138776.112 (cf. p. 23).

[FCS+19] Q Ferré, G Charbonnier, N Sadouni et al. « OLOGRAM : Modeling the distribution of overlap length between genomic region sets ». In : (2019), p. 12-14 (cf. p. 214).

[FSA16] Dan Filipescu, Emmanuelle Szenker et Geneviève Almouzni. « Developmental roles of histone H3 variants and their chaperones ». In : *The Functional Nucleus*. T. 29. 11. Cham : Springer International Publishing, 2016, p. 385-419. ISBN : 9783319388823. DOI : 10.1007/978-3-319-38882-3_17 (cf. p. 30).

[FKR+14] Alistair R.R. Forrest, Hideya Kawaji, Michael Rehli et al. « A promoter-level mammalian expression atlas ». In : *Nature* 507.7493 (2014), p. 462-470. ISSN : 14764687. DOI : 10.1038/nature13182 (cf. p. 66).

[FDF+19] Adam Frankish, Mark Diekhans, Anne Maud Ferreira et al. « GENCODE reference annotation for the human and mouse genomes ». In : *Nucleic Acids Research* 47.D1 (2019), p. D766-D773. ISSN : 13624962. DOI : 10.1093/nar/gky955 (cf. p. 66).

[FMK+03] Christopher A French, Isao Miyoshi, Ichiro Kubonishi et al. « BRD4-NUT fusion oncogene : A novel mechanism in aggressive carcinoma ». In : *Cancer Research* 63.2 (2003), p. 304-307. ISSN : 00085472 (cf. p. 170).

[FLV17] Daniel Frías-Lasserre et Cristian A. Villagra. « The importance of ncRNAs as epigenetic mechanisms in phenotypic variation and organic evolution ». In : *Frontiers in Microbiology* 8.DEC (2017), p. 1-13. ISSN : 1664302X. DOI : 10.3389/fmicb.2017.02483 (cf. p. 27).

[FMM+92] Marianne Frommer, L. E. McDonald, Douglas S Millar et al. « A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. » In : *Proceedings of the National Academy of Sciences* 89.5 (1992), p. 1827-1831. ISSN : 0027-8424. DOI : 10.1073/pnas.89.5.1827 (cf. p. 62).

[GF06] Miklos Gaszner et Gary Felsenfeld. « Insulators : Exploiting transcriptional and epigenetic mechanisms ». In : *Nature Reviews Genetics* 7.9 (2006), p. 703-713. ISSN : 14710056. DOI : 10.1038/nrg1925 (cf. p. 24, 25).

[GBM+12] Jonathan Gaucher, Fayçal Boussouar, Emilie Montellier et al. « Bromodomain-dependent stage-specific male genome programming by Brdt ». In : *The EMBO Journal* 31.19 (2012), p. 3809-3820. ISSN : 0261-4189. DOI : 10.1038/emboj.2012.233 (cf. p. 52, 141).

[GAM+04] Thierry Gautier, D. Wade Abbott, Annie Molla et al. « Histone variant H2ABbd confers lower stability to the nucleosome ». In : *EMBO reports* 5.7 (2004), p. 715-720. ISSN : 1469-221X. DOI : 10.1038/sj.embor.7400182 (cf. p. 30).

[GLZ+17] Shao-Qin Ge, Sheng-Li Lin, Zheng-Hui Zhao et al. « Epigenetic dynamics and interplay during spermatogenesis and embryogenesis : implications for male fertility and offspring health ». In : *Oncotarget* 8.32 (2017), p. 53804-53818. ISSN : 1949-2553. DOI : 10.18632/oncotarget.17479 (cf. p. 55).

[Gen00] Gene Ontology Consortium. « Gene Ontology : tool for the unification of biology ». In : 25.may (2000), p. 25-29 (cf. p. 67).

[GKM+07] Paul G Giresi, Jonghwan Kim, Ryan M McDaniell et al. « FAIRE ((F)under-barormaldehyde-(A)under-barssisted (I)under-barsolation of (R)under-baregulatory (E)under-barlements) isolates active regulatory elements from human chromatin ». In : *Genome Research* 17 (2007), 877-885 ST -FAIRE ((F)under-barormaldehyde-(A)un. ISSN : 1088-9051. DOI : 10.1101/gr.5533506 (cf. p. 61).

[GAF+12] Daniel E. Gomez, Romina G. Armando, Hernán G. Farina et al. « Telomere structure and telomerase in health and disease (Review) ». In : *International Journal of Oncology* 41.5 (2012), p. 1561-1569. ISSN : 10196439. DOI : 10.3892/ijo.2012.1611 (cf. p. 26).

[GFI16] Steven N. Goodman, Daniele Fanelli et John P. A. Ioannidis. « What does research reproducibility mean ? » In : 8.341 (2016). DOI : 10.1108/CG-10-2012-0073 (cf. p. 72).

[GZH+16] Afsaneh Goudarzi, Di Zhang, He Huang et al. « Dynamic Competing Histone H4 K5K8 Acetylation and Butyrylation Are Hallmarks of Highly Active Gene Promoters ». In : *Molecular Cell* 62.2 (2016), p. 169-180. ISSN : 10972765. DOI : 10.1016/j.molcel.2016.03.014 (cf. p. 52, 55, 141, 170).

[GER+07] Jérôme Govin, Emmanuelle Escoffier, Sophie Rousseaux et al. « Pericentric heterochromatin reprogramming by new histone variants during mouse spermiogenesis ». In : *The Journal of Cell Biology* 176.3 (2007), p. 283-294. ISSN : 0021-9525. DOI : 10.1083/jcb.200604141 (cf. p. 30, 155).

[GDM+19] Kevin Grosselin, Adeline Durand, Justine Marsolier et al. « High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer ». In : *Nature Genetics* 51.6 (2019), p. 1060-1066. ISSN : 1061-4036. DOI : 10.1038/s41588-019-0424-9 (cf. p. 37).

[Gu19] Zuguang Gu. *rGREAT : Client for GREAT Analysis*. 2019. DOI : https://doi.org/doi:10.18129/B9.bioc.rGREAT (cf. p. 221).

[HFG+06] Jennifer Harrow, Adam Frankish, Jose M Gonzalez et al. « GENCODE : producing a reference annotation for ENCODE ». In : *Genome Biology* 7.Suppl 1 (2006), p. 1760-1774. ISSN : 14656906. DOI : 10.1186/gb-2006-7-s1-s4 (cf. p. 66).

[HFG+12] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez et al. « GENCODE : The reference human genome annotation for the ENCODE project ». In : *Genome Research* 22.9 (2012), p. 1760-1774. ISSN : 10889051. DOI : 10.1101/gr.135350.111 (cf. p. 66, 67).

[HM06] Katsuhiko Hayashi et Yasuhisa Matsui. « Meisetz, A Novel Histone Tri-Methyltransferase, Regulates Meiosis-Specific Epigenesis ». In : *Cell Cycle* 5.6 (2006), p. 615-620. issn : 1538-4101. doi : 10.4161/cc.5.6.2572 (cf. p. 54).

[HPPF+00] Mira Hazzouri, Christophe Pivot-Pajot, Anne-Karen Faure et al. « Regulated hyperacetylation of core histones during mouse spermatogenesis : involvement of histone-deacetylases ». In : *European Journal of Cell Biology* 79.12 (2000), p. 950-960. issn : 01719335. doi : 10.1078/0171-9335-00123 (cf. p. 52, 55).

[HM14] Edith Heard et Robert A. Martienssen. « Transgenerational epigenetic inheritance : Myths and mechanisms ». In : *Cell* 157.1 (2014), p. 95-109. issn : 10974172. doi : 10.1016/j.cell.2014.02.045 (cf. p. 36).

[HS15] Steven Henikoff et Mitchell Mitchell Smith. « Histone Variants and Epigenetics ». In : *Cold Spring Harbor Perspectives in Biology* 7.1 (2015), a019364. issn : 1943-0264. doi : 10.1101/cshperspect.a019364 (cf. p. 32).

[HBP+14] Vincent J. Henry, Anita E. Bandrowski, A.-S. Pepin et al. « OMICtools : an informative directory for multi-omic data analysis ». In : *Database* 2014.13 (2014), bau069-bau069. issn : 1758-0463. doi : 10.1093/database/bau069 (cf. p. 71).

[HWH+12] Brian R. Herb, Florian Wolschin, Kasper D. Hansen et al. « Reversible switching between epigenetic states in honeybee behavioral subcastes ». In : *Nature Neuroscience* 15.10 (2012), p. 1371-1373. issn : 10976256. doi : 10.1038/nn.3218 (cf. p. 27).

[Her19] Matthew Herper. *Illumina Promises To Sequence Human Genome For $100 – But Not Quite Yet*. 2019 (cf. p. 13).

[HDW+19] Erin E. Heyer, Ira W. Deveson, Danson Wooi et al. « Diagnosis of fusion genes using targeted RNA sequencing ». In : *Nature Communications* 10.1 (2019), p. 1-12. issn : 20411723. doi : 10.1038/s41467-019-09374-9 (cf. p. 58).

[HG10] Cathryn A Hogarth et Michael D Griswold. « Review series The key role of vitamin A in spermatogenesis ». In : *J Clin Invest* 120.4 (2010), p. 956. issn : 00219738. doi : 10.1172/JCI41303.956 (cf. p. 50).

[HSD+14] Gary C. Hon, Chun Xiao Song, Tingting Du et al. « 5mC oxidation by Tet2 modulates enhancer activity and timing of transcriptome reprogramming during differentiation ». In : *Molecular Cell* 56.2 (2014), p. 286-297. issn : 10974164. doi : 10.1016/j.molcel.2014.08.026 (cf. p. 28).

[HCF+18] Gangqing Hu, Kairong Cui, Difeng Fang et al. « Transformation of Accessible Chromatin and 3D Nucleome Underlies Lineage Commitment of Early T Cells ». In : *Immunity* 48.2 (2018), 227-242.e8. ISSN : 10974180. DOI : 10.1016/j.immuni.2018.01.013 (cf. p. 45, 46).

[HSG+14] He Huang, Benjamin R. Sabari, Benjamin A. Garcia et al. « SnapShot : Histone modifications ». In : *Cell* 159.2 (2014), 458-458.e1. ISSN : 10974172. DOI : 10.1016/j.cell.2014.09.037 (cf. p. 34, 35).

[HPS+10] Yun Huang, William A. Pastor, Yinghua Shen et al. « The Behaviour of 5-Hydroxymethylcytosine in Bisulfite Sequencing ». In : *PLoS ONE* 5.1 (2010). Sous la dir. de Jun Liu, e8888. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0008888 (cf. p. 62).

[IDT+10] Shinsuke Ito, Ana C. Dalessio, Olena V. Taranova et al. « Role of tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification ». In : *Nature* 466.7310 (2010), p. 1129-1133. ISSN : 00280836. DOI : 10.1038/nature09303 (cf. p. 28).

[ISD+11] Shinsuke Ito, Li Shen, Qing Dai et al. « Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. » In : *Science* 333.6047 (2011), p. 1300-3 (cf. p. 29).

[JG04] Jessica M. Jones et Martin Gellert. « The taming of a transposon : V(D)J recombination and the immune system ». In : *Immunological Reviews* 200.1 (2004), p. 233-248. ISSN : 0105-2896. DOI : 10.1111/j.0105-2896.2004.00168.x (cf. p. 26).

[JBB+08] Peter A. Jones, Stephen B. Baylin, Bradley E. Bernstein et al. « Moving AHEAD with an international human epigenome project. » In : *Nature* 454.7205 (2008), p. 711-715. ISSN : 1476-4687. DOI : 10.1038/454711a (cf. p. 69).

[JM05] Peter A. Jones et Robert Martienssen. « A blueprint for a human epigenome project : The AACR Human Epigenome Workshop ». In : *Cancer Research* 65.24 (2005), p. 11241-11246. ISSN : 00085472. DOI : 10.1158/0008-5472.CAN-05-3865 (cf. p. 69).

[KMTC17] Ilene Karsch-Mizrachi, Toshihisa Takagi et Guy Cochrane. « The international nucleotide sequence database collaboration ». In : *Nucleic Acids Research* 46.D1 (2017), p. D48-D51. ISSN : 0305-1048. DOI : 10.1093/nar/gkx1097 (cf. p. 71).

[KWS+14] M. Kellis, B. Wold, M. P. Snyder et al. « Defining functional DNA elements in the human genome ». In : *Proceedings of the National Academy of Sciences* 111.17 (2014), p. 6131-6138. ISSN : 0027-8424. DOI : 10.1073/pnas.1318948111 (cf. p. 21, 24).

[KB06] Robert J. Klose et Adrian P. Bird. « Genomic DNA methylation : The mark and its mediators ». In : *Trends in Biochemical Sciences* 31.2 (2006), p. 89-97. issn : 09680004. doi : 10.1016/j.tibs.2005.12.008 (cf. p. 28).

[Knu84] Donale E. Knuth. « Literate Programming ». In : *The Computer Journal* 27.2 (1984), p. 97-111 (cf. p. 80).

[KKG+12] Petros Kolovos, Tobias A. Knoch, Frank G. Grosveld et al. « Enhancers and silencers : An integrated and simple model for their function ». In : *Epigenetics and Chromatin* 5.1 (2012), p. 1-8. issn : 17568935. doi : 10.1186/1756-8935-5-1 (cf. p. 24).

[KMH+15] Marta Kulis, Angelika Merkel, Simon Heath et al. « Whole-genome fingerprint of the DNA methylome during human B cell differentiation ». In : *Nature Genetics* 47.7 (2015), p. 746-756. issn : 15461718. doi : 10.1038/ng.3291 (cf. p. 226).

[KSB17] Gregory M. Kurtzer, Vanessa Sochat et Michael W. Bauer. « Singularity : Scientific containers for mobility of compute ». In : *PLoS ONE* 12.5 (2017), p. 1-20. issn : 19326203. doi : 10.1371/journal.pone.0177459 (cf. p. 78).

[LAKK+15] Ilkka Lappalainen, Jeff Almeida-King, Vasudev Kumanduri et al. « The European Genome-phenome Archive of human data consented for biomedical research ». In : *Nature Genetics* 47.7 (2015), p. 692-695. issn : 15461718. doi : 10.1038/ng.3312 (cf. p. 65).

[LHL+17] Peter A. Larsen, R. Alan Harris, Yue Liu et al. « Hybrid de novo genome assembly and centromere characterization of the gray mouse lemur (Microcebus murinus) ». In : *BMC Biology* 15.1 (2017), p. 1-17. issn : 17417007. doi : 10.1186/s12915-017-0439-6 (cf. p. 66).

[LGC09] Michael Lawrence, Robert Gentleman et Vincent Carey. « rtracklayer : An R package for interfacing with genome browsers ». In : *Bioinformatics* 25.14 (2009), p. 1841-1842. issn : 13674803. doi : 10.1093/bioinformatics/btp328 (cf. p. 211).

[LDS16] Moyra Lawrence, Sylvain Daujat et Robert Schneider. *Lateral Thinking : How Histone Modifications Regulate Gene Expression*. 2016. doi : 10.1016/j.tig.2015.10.007 (cf. p. 34).

[LM13] Alan C Leonard et Marcel Méchali. « DNA Replication Origins ». In : (2013), p. 1-17. issn : 1943-0264 (Electronic). doi : 10.1101/cshperspect.a010116 (cf. p. 24).

[Li05] H. Li. « TreeFam : a curated database of phylogenetic trees of animal gene families ». In : *Nucleic Acids Research* 34.90001 (2005), p. D572-D580. ISSN : 0305-1048. DOI : 10.1093/nar/gkj118 (cf. p. 67).

[LDG05] Ling-Hong Li, James M. Donald et Mari S. Golub. « Review on testicular development, structure, function, and regulation in common marmoset ». In : *Birth Defects Research Part B : Developmental and Reproductive Toxicology* 74.5 (2005), p. 450-469. ISSN : 1542-9733. DOI : 10.1002/bdrb.20057 (cf. p. 30).

[LBT+15] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir et al. « The Molecular Signatures Database Hallmark Gene Set Collection ». In : *Cell Systems* 1.6 (2015), p. 417-425. ISSN : 24054712. DOI : 10.1016/j.cels.2015.12.004 (cf. p. 67).

[LSP+11] Arthur Liberzon, Aravind Subramanian, Reid Pinchback et al. « Molecular signatures database (MSigDB) 3.0 ». In : *Bioinformatics* 27.12 (2011), p. 1739-1740. ISSN : 13674803. DOI : 10.1093/bioinformatics/btr260 (cf. p. 67).

[LCK+19] F Lopez, G Charbonnier, Y Kermezli et al. « Explore, edit and leverage genomic annotations using Python GTF toolkit ». In : *Bioinformatics* Cli (2019). Sous la dir. de John Hancock, p. 2-3. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btz116 (cf. p. 211).

[LT05] Noel F. Lowndes et Geraldine W.-L. Toh. « DNA Repair : The Importance of Phosphorylating Histone H2AX ». In : *Current Biology* 15.3 (2005), p. 99-102. ISSN : 09609822. DOI : 10.1016/j.cub.2005.01.028 (cf. p. 30).

[LLJ+14] Falong Lu, Yuting Liu, Lan Jiang et al. « Role of Tet proteins in enhancer activity and telomere elongation ». In : *Genes and Development* 28.19 (2014), p. 2103-2119. ISSN : 15495477. DOI : 10.1101/gad.248005.114 (cf. p. 28).

[MIT+14] Kazuhiro Maeshima, Ryosuke Imai, Sachiko Tamura et al. « Chromatin as dynamic 10-nm fibers ». In : *Chromosoma* 123.3 (2014), p. 225-237. ISSN : 14320886. DOI : 10.1007/s00412-014-0460-2 (cf. p. 17).

[MLG+10] Maud Marques, Liette Laflamme, Alain L. Gervais et al. « Reconciling the positive and negative roles of histone H2A.Z in gene transcription ». In : *Epigenetics* 5.4 (2010), p. 267-272. ISSN : 1559-2294. DOI : 10.4161/epi.5.4.11520 (cf. p. 30).

[MD02] William F. Marzluff et Robert J. Duronio. « Histone mRNA expression : Multiple levels of cell cycle regulation and important developmental consequences ». In : *Current Opinion in Cell Biology* 14.6 (2002), p. 692-699. ISSN : 09550674. DOI : 10.1016/S0955-0674(02)00387-3 (cf. p. 29).

[MBH+10] Cory Y. McLean, Dave Bristor, Michael Hiller et al. « GREAT improves functional interpretation of cis-regulatory regions ». In : *Nature Biotechnology* 28.5 (2010), p. 495-501. ISSN : 10870156. DOI : 10.1038/nbt.1630. arXiv : 15334406 (cf. p. 221).

[ML14] Clifford A Meyer et X Shirley Liu. « Identifying and mitigating bias in next-generation sequencing methods for chromatin biology ». In : *Nature Reviews Genetics* 15.11 (2014), p. 709-721. ISSN : 1471-0056. DOI : 10.1038/nrg3788 (cf. p. 59).

[MEK15] Karen H Miga, Christopher Eisenhart et W James Kent. « Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments. » In : *Nucleic acids research* 43.20 (2015), e133. ISSN : 1362-4962. DOI : 10.1093/nar/gkv671 (cf. p. 59).

[MLO+18] Biswapriya B Misra, Carl D Langefeld, Michael Olivier et al. « Integrated Omics : Tools, Advances, and Future Approaches. » In : *Journal of molecular endocrinology* 3.1 (2018), p. 57-111. ISSN : 1479-6813. DOI : 10.1530/JME-18-0055 (cf. p. 65).

[MXW+07] Gaku Mizuguchi, Hua Xiao, Jan Wisniewski et al. « Nonhistone Scm3 and Histones CenH3-H4 Assemble the Core of Centromere-Specific Nucleosomes ». In : *Cell* 129.6 (2007), p. 1153-1164. ISSN : 00928674. DOI : 10.1016/j.cell.2007.04.026 (cf. p. 30).

[MBR+13] Emilie Montellier, Fayçal Boussouar, Sophie Rousseaux et al. « Chromatin-to-nucleoprotamine transition is controlled by the histone H2B variant TH2B ». In : *Genes & Development* 27.15 (2013), p. 1680-1692. ISSN : 0890-9369. DOI : 10.1101/gad.220095.113 (cf. p. 141).

[MUS12] Andrew Morin, Jennifer Urban et Piotr Sliz. « A Quick Guide to Software Licensing for the Scientist-Programmer ». In : *PLoS Computational Biology* 8.7 (2012), e1002598. DOI : 10.1371/journal.pcbi.1002598 (cf. p. 83, 84).

[MBF+08] Ryan D Morin, Matthew Bainbridge, Anthony Fejes et al. « Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing ». In : *BioTechniques* 45.1 (2008), p. 81-94. ISSN : 0736-6205. DOI : 10.2144/000112900 (cf. p. 57).

[MRS+09] Jeanne Morinière, Sophie Rousseaux, Ulrich Steuerwald et al. « Cooperative binding of two acetylation marks on a histone tail by a single bromodomain ». en. In : *Nature* 461.7264 (2009), p. 664-668. ISSN : 0028-0836. DOI : 10.1038/nature08397 (cf. p. 141).

[MLL+16] Paul Muir, Shantao Li, Shaoke Lou et al. « The real cost of sequencing : scaling computation to keep pace with data generation ». In : *Genome Biology* 17.1 (2016), p. 53. ISSN : 1474-760X. DOI : 10.1186/s13059-016-0917-0 (cf. p. 71).

[Nat18] Nature Methods Editorial. « Sharing epigenomes globally ». In : *Nature Methods* 15.3 (2018), p. 151-151. ISSN : 1548-7091. DOI : 10.1038/nmeth.4630 (cf. p. 69).

[OSR+07] Yuki Okada, Greg Scott, Manas K. Ray et al. « Histone demethylase JHDM2A is critical for Tnp1 and Prm1 transcription and spermatogenesis ». In : *Nature* 450.7166 (2007), p. 119-123. ISSN : 0028-0836. DOI : 10.1038/nature06236 (cf. p. 54).

[OCB+14] David A. Orlando, Mei Wei Chen, Victoria E. Brown et al. « Quantitative ChIP-Seq Normalization Reveals Global Modulation of the Epigenome ». In : *Cell Reports* 9.3 (2014), p. 1163-1170. ISSN : 22111247. DOI : 10.1016/j.celrep.2014.10.018 (cf. p. 170).

[OPD+17] Horng D. Ou, Sébastien Phan, Thomas J. Deerinck et al. « ChromEMT : Visualizing 3D chromatin structure and compaction in interphase and mitotic cells ». In : *Science* 357.6349 (2017). ISSN : 10959203. DOI : 10.1126/science.aag0025 (cf. p. 18).

[Par09] Peter J. Park. « ChIPseq : advantages and challenges of a maturing technology ». In : *Nature Reviews Genetics* 10.10 (2009), p. 669-680. ISSN : 1471-0056. DOI : 10.1038/nrg2641 (cf. p. 59).

[PJA+03] E. Passegue, C. H. M. Jamieson, L. E. Ailles et al. « Normal and leukemic hematopoiesis : Are leukemias a stem cell disorder or a reacquisition of stem cell characteristics ? » In : *Proceedings of the National Academy of Sciences* 100.Supplement 1 (2003), p. 11842-11849. ISSN : 0027-8424. DOI : 10.1073/pnas.2034201100 (cf. p. 40).

[PBZC+11] Aleksandra Pekowska, Touati Benoukraf, Joaquin Zacarias-Cabeza et al. « H3K4 tri-methylation provides an epigenetic signature of active enhancers ». In : *EMBO Journal* 30.20 (2011), p. 4198-4210. ISSN : 02614189. DOI : 10.1038/emboj.2011.295 (cf. p. 45).

[PCA+] Aleksandra Pekowska, Guillaume Charbonnier, Jean Christophe Andrau et al. « H3K4 methylation at enhancers and broad promoters ». In : *Letter to Molecular Cell* () (cf. p. 256).

[PW14] Veronica J. Peschansky et Claes Wahlestedt. « Non-coding RNAs as direct and indirect modulators of epigenetic regulation ». In : *Epigenetics* 9.1 (2014), p. 3-12. issn : 15592308. doi : 10.4161/epi.27473 (cf. p. 27).

[POS+01] Antoine H.F.M. Peters, Dónal O'Carroll, Harry Scherthan et al. « Loss of the Suv39h Histone Methyltransferases Impairs Mammalian Heterochromatin and Genome Stability ». In : *Cell* 107.3 (2001), p. 323-337. issn : 00928674. doi : 10.1016/S0092-8674(01)00542-6 (cf. p. 54).

[PSW+14] Toni Pfaffeneder, Fabio Spada, Mirko Wagner et al. « Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA ». In : *Nature Chemical Biology* 10.7 (2014), p. 574-581. issn : 15524469. doi : 10.1038/nchembio.1532 (cf. p. 29).

[PBL+14] Anton Polishko, Evelien M Bunnik, Karine G Le Roch et al. « PuFFIN - a parameter-free method to build nucleosome maps from paired-end reads ». In : *BMC Bioinformatics* 15.Suppl 9 (2014), S11. issn : 1471-2105. doi : 10.1186/1471-2105-15-S9-S11 (cf. p. 60).

[PD15] Ana Pombo et Niall Dillon. « Three-dimensional genome architecture : players and mechanisms ». In : *Nature Reviews Molecular Cell Biology* 16.4 (2015), p. 245-257. issn : 1471-0072. doi : 10.1038/nrm3965 (cf. p. 18).

[QSC+12] Michael A. Quail, Miriam Smith, Paul Coupland et al. « A tale of three next generation sequencing platforms : comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers ». In : *BMC Genomics* 13.1 (2012), p. 1. issn : 14712164. doi : 10.1186/1471-2164-13-341 (cf. p. 56).

[Qui14] Aaron R. Quinlan. « BEDTools : The Swiss-Army Tool for Genome Feature Analysis ». In : *Current Protocols in Bioinformatics* 47.1 (2014), p. 11.12.1-11.12.34. issn : 19343396. doi : 10.1002/0471250953.bi1112s47 (cf. p. 214).

[RNA+16] Bahareh Rabbani, Hirofumi Nakaoka, Shahin Akhondzadeh et al. « Next generation sequencing : Implications in personalized medicine and pharmacogenomics ». In : *Molecular BioSystems* 12.6 (2016), p. 1818-1830. issn : 17422051. doi : 10.1039/c6mb00115g (cf. p. 56).

[RH18] Alejandro Reyes et Wolfgang Huber. « Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues ». In : *Nucleic Acids Research* 46.2 (2018), p. 582-592. issn : 13624962. doi : 10.1093/nar/gkx1165 (cf. p. 21).

[RSD+10] Nicolas Reynoird, Brian E Schwartz, Manuela Delvecchio et al. « Oncogenesis by sequestration of CBP/p300 in transcriptionally inactive hyperacetylated chromatin domains ». In : *The EMBO Journal* 29.17 (2010), p. 2943-2952. ISSN : 0261-4189. DOI : 10.1038/emboj.2010.176 (cf. p. 170).

[RP11] Ho Sung Rhee et B. Franklin Pugh. « Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution ». In : *Cell* 147.6 (2011), p. 1408-1419. ISSN : 00928674. DOI : 10.1016/j.cell.2011.11.013 (cf. p. 59).

[RF98] Ellen Robey et BJ Fowlkes. « The $\alpha\beta$ versus $\gamma\delta$ T-cell lineage choice ». In : *Current Opinion in Immunology* 10.2 (1998), p. 181-187. ISSN : 09527915. DOI : 10.1016/S0952-7915(98)80247-1 (cf. p. 226).

[RDL+18] Luis F. Rossi, Noé U. De La Sancha, Juan P. Luaces et al. « Morphological description and comparison of sperm from eighteen species of cricetid rodents ». In : *Journal of Mammalogy* 99.6 (2018), p. 1398-1404. ISSN : 15451542. DOI : 10.1093/jmammal/gyy146 (cf. p. 47).

[RMY08] Ellen V. Rothenberg, Jonathan E. Moore et Mary A. Yui. « Launching the T-cell-lineage developmental programme ». In : *Nature Reviews Immunology* 8.1 (2008), p. 9-21. ISSN : 1474-1733. DOI : 10.1038/nri2232 (cf. p. 40, 41, 45).

[SGR+10] Geir K. Sandve, Sveinung Gundersen, Halfdan Rydbeck et al. « The Genomic HyperBrowser : inferential genomics at the sequence level ». In : *Genome Biology* 11.12 (2010), R121. ISSN : 1465-6906. DOI : 10.1186/gb-2010-11-12-r121 (cf. p. 214).

[SLZ08] Chaitanya R. Sanna, Wen Hsiung Li et Liqing Zhang. « Overlapping genes in the human and mouse genomes ». In : *BMC Genomics* 9 (2008), p. 1-11. ISSN : 14712164. DOI : 10.1186/1471-2164-9-169 (cf. p. 21).

[SPJP+17] Enrique Carrillo-de Santa-Pau, David Juan, Vera Pancaldi et al. « Automatic identification of informative regions with epigenomic changes associated to hematopoiesis ». In : *Nucleic Acids Research* 45.16 (2017), p. 9244-9259. ISSN : 13624962. DOI : 10.1093/nar/gkx618. arXiv : 1611.06654 (cf. p. 37, 87, 226).

[SJT+15] Serena Sanulli, Neil Justin, Aurélie Teissandier et al. « Jarid2 Methylation via the PRC2 Complex Regulates H3K27me3 Deposition during Cell Differentiation ». In : *Molecular Cell* 57.5 (2015), p. 769-783. ISSN : 10972765. DOI : 10.1016/j.molcel.2014.12.020 (cf. p. 226).

[Sat07] Naoki Sato. « Origin and Evolution of Plastids : Genomic View on the Unification and Diversity of Plastids ». In : *The Structure and Function of Plastids*. Sous la dir. de Robert R Wise et J Kenneth Hoober. Dordrecht : Springer Netherlands, 2007, p. 75-102. isbn : 978-1-4020-4061-0. doi : 10.1007/978-1-4020-4061-0_4 (cf. p. 14).

[SPS12] Undraga Schagdarsurengin, Agnieszka Paradowska et Klaus Steger. « Analysing the sperm epigenome : roles in early embryogenesis and assisted reproduction ». In : *Nature Reviews Urology* 9.11 (2012), p. 609-619. issn : 1759-4812. doi : 10.1038/nrurol.2012.183 (cf. p. 53).

[SMR+16] Ronald P. Schuyler, Angelika Merkel, Emanuele Raineri et al. « Distinct Trends of DNA Methylation Patterning in the Innate and Adaptive Immune Systems ». In : *Cell Reports* 17.8 (2016), p. 2101-2111. issn : 22111247. doi : 10.1016/j.celrep.2016.10.054 (cf. p. 226).

[SB16] Nathan C. Sheffield et Christoph Bock. « LOLA : enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor ». In : *Bioinformatics* 32.4 (2016), p. 587-589. issn : 1367-4803. doi : 10.1093/bioinformatics/btv612 (cf. p. 214).

[SLM+09] Hyunjin Shin, Tao Liu, Arjun K. Manrai et al. « CEAS : Cis-regulatory element annotation system ». In : *Bioinformatics* 25.19 (2009), p. 2605-2606. issn : 13674803. doi : 10.1093/bioinformatics/btp479 (cf. p. 23, 214).

[SBB+18] H. Shiota, S. Barral, T. Buchou et al. « Nut Directs p300-Dependent, Genome-Wide H4 Hyperacetylation in Male Germ Cells ». In : *Cell Reports* 24.13 (2018). issn : 22111247. doi : 10.1016/j.celrep.2018.08.069 (cf. p. 170).

[SKG+18] Boris Simovski, Chakravarthi Kanduri, Sveinung Gundersen et al. « Coloc-stats : a unified web interface to perform colocalization analysis of genomic features ». In : *Nucleic Acids Research* 46.W1 (2018), W186-W193. issn : 0305-1048. doi : 10.1093/nar/gky474 (cf. p. 214).

[SRG+18] Roderick C. Slieker, Caroline L. Relton, Tom R. Gaunt et al. « Age-related DNA methylation changes are tissue-specific with ELOVL2 promoter methylation as exception ». In : *Epigenetics and Chromatin* 11.1 (2018), p. 1-11. issn : 17568935. doi : 10.1186/s13072-018-0191-3 (cf. p. 36).

[SM13] Zachary D. Smith et Alexander Meissner. « DNA methylation : Roles in mammalian development ». In : *Nature Reviews Genetics* 14.3 (2013), p. 204-220. issn : 14710056. doi : 10.1038/nrg3354 (cf. p. 27).

[SPN+17] Tatiana A. Soboleva, Brian J. Parker, Maxim Nekrasov et al. « A new link between transcriptional initiation and pre-mRNA splicing : The RNA binding histone variant H2A.B ». In : *PLOS Genetics* 13.2 (2017). Sous la dir. de Robert Schneider, e1006633. issn : 1553-7404. doi : 10.1371/journal.pgen.1006633 (cf. p. 30).

[Sto09] Victoria Stodden. « Enabling reproducible research : Licensing scientific innovation ». In : *International Journal of Communications Law and Policy* 13 (2009), p. 1-25 (cf. p. 83).

[SA00] Brian D. Strahl et C. David Allis. « The language of covalent histone modifications ». In : *Nature* 403.6765 (2000), p. 41-45. issn : 0028-0836. doi : 10.1038/47412 (cf. p. 34).

[SHA+16] Hendrik G. Stunnenberg, Martin Hirst, Sergio Abrignani et al. « The International Human Epigenome Consortium : A Blueprint for Scientific Collaboration and Discovery ». In : *Cell* 167.5 (2016), p. 1145-1149. issn : 00928674. doi : 10.1016/j.cell.2016.11.007 (cf. p. 69).

[STM+05] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha et al. « Gene set enrichment analysis : a knowledge-based approach for interpreting genome-wide expression profiles. » In : *Proceedings of the National Academy of Sciences of the United States of America* 102.43 (2005), p. 15545-50. issn : 0027-8424. doi : 10.1073/pnas.0506580102 (cf. p. 67).

[SRG+15] Peter H. Sudmant, Tobias Rausch, Eugene J. Gardner et al. « An integrated map of structural variation in 2,504 human genomes ». In : *Nature* 526.7571 (2015), p. 75-81. issn : 0028-0836. doi : 10.1038/nature15394 (cf. p. 65).

[SWW+11] Hao Sun, Jiejun Wu, Priyankara Wickramasinghe et al. « Genome-wide mapping of RNA Pol-II promoter usage in mouse tissues by ChIP-seq ». In : *Nucleic Acids Research* 39.1 (2011), p. 190-201. issn : 1362-4962. doi : 10.1093/nar/gkq775 (cf. p. 23).

[TKO+10] H. Tachiwana, W. Kagawa, A. Osakabe et al. « Structural basis of instability of the nucleosome containing a testis-specific histone variant, human H3T ». In : *Proceedings of the National Academy of Sciences* 107.23 (2010), p. 10454-10459. issn : 0027-8424. doi : 10.1073/pnas.1003064107 (cf. p. 30).

[TO09] Leila Taher et Ivan Ovcharenko. « Variable locus length in the human genome leads to ascertainment bias in functional inference for non-coding elements ». In : *Bioinformatics* 25.5 (2009), p. 578-584. issn : 1460-2059. doi : 10.1093/bioinformatics/btp043 (cf. p. 221).

[TLL+11] Minjia Tan, Hao Luo, Sangkyu Lee et al. « Identification of 67 Histone Marks and Histone Lysine Crotonylation as a New Type of Histone Modification ». In : *Cell* 146.6 (2011), p. 1016-1028. ISSN : 00928674. DOI : 10.1016/j.cell.2011.08.008 (cf. p. 52, 55, 141).

[TRS+15] Marco Trerotola, Valeria Relli, Pasquale Simeone et al. « Epigenetic inheritance and the missing heritability ». In : *Human genomics* 9 (2015), p. 17. ISSN : 14797364. DOI : 10.1186/s40246-015-0041-3 (cf. p. 36).

[TBL+14] Maria Tsompana, Michael J. MJ Buck, K Luger et al. « Chromatin accessibility : a window into the genome ». In : *Epigenetics & Chromatin* 7.1 (2014), p. 33. ISSN : 1756-8935. DOI : 10.1186/1756-8935-7-33 (cf. p. 64).

[TST+18] Clare Turnbull, Richard H. Scott, Ellen Thomas et al. « The 100 000 Genomes Project : bringing whole genome sequencing to the NHS ». In : *BMJ* 361.April (2018), k1687. ISSN : 0959-8138. DOI : 10.1136/bmj.k1687 (cf. p. 65).

[Tur05] Bryan M Turner. « Reading signals on the nucleosome with a new nomenclature for modified histones ». In : *Nature Structural & Molecular Biology* 12.2 (2005), p. 110-112. ISSN : 1545-9993. DOI : 10.1038/nsmb0205-110 (cf. p. 34, 35).

[Tur09] Bryan M. Turner. « Epigenetic responses to environmental change and their evolutionary implications ». In : *Philosophical Transactions of the Royal Society B : Biological Sciences* 364.1534 (2009), p. 3403-3418. ISSN : 14712970. DOI : 10.1098/rstb.2009.0125 (cf. p. 36).

[Tyl02] Jessica K. Tyler. « Chromatin assembly ». In : *European Journal of Biochemistry* 269.9 (2002), p. 2268-2274. DOI : 10.1046/j.1432-1033.2002.02890.x (cf. p. 30, 31).

[ULB+19] Jacob C. Ulirsch, Caleb A. Lareau, Erik L. Bao et al. « Interrogation of human hematopoiesis at single-cell and single-variant resolution ». In : *Nature Genetics* 51.April (2019). ISSN : 15461718. DOI : 10.1038/s41588-019-0362-6 (cf. p. 37, 41).

[VH06] Matthew K. Vickaryous et Brian K. Hall. « Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest ». In : *Biological Reviews of the Cambridge Philosophical Society* 81.3 (2006), p. 425-455. ISSN : 14647931. DOI : 10.1017/S1464793106007068 (cf. p. 27).

[VM05] Margarita Vigodner et Patricia L Morris. « Testicular expression of small ubiquitin-related modifier-1 (SUMO-1) supports multiple roles in spermatogenesis : Silencing of sex chromosomes in spermatocytes, spermatid microtubule nucleation, and nuclear reshaping ». In : *Developmental Biology* 282.2 (2005), p. 480-492. ISSN : 00121606. DOI : 10.1016/j.ydbio.2005.03.034 (cf. p. 54).

[VDT+18] Trung Nghia Vu, Wenjiang Deng, Quang Thinh Trac et al. « A fast detection of fusion genes from paired-end RNA-seq data ». In : *BMC Genomics* 19.1 (2018), p. 786. ISSN : 1471-2164. DOI : 10.1186/s12864-018-5156-1 (cf. p. 58).

[Wad57] C H Waddington. *The strategy of the genes. A discussion of some aspects of theoretical biology*. English. London : George Allen & Unwin, Ltd., 1957, ix +-262 pp. (Cf. p. 226).

[Wad12] C. H. Waddington. « The epigenotype. 1942. » In : *International journal of epidemiology* 41.1 (2012), p. 10-13. ISSN : 14643685. DOI : 10.1093/ije/dyr184 (cf. p. 27).

[WYH+16] Chun Yan Wang, Pei Fa Yu, Xiao Bing He et al. « $\alpha\beta$ T-cell receptor bias in disease and therapy (Review) ». In : *International Journal of Oncology* 48.6 (2016), p. 2247-2256. ISSN : 17912423. DOI : 10.3892/ijo.2016.3492 (cf. p. 43).

[WDP+07] James Z. Wang, Zhidian Du, Rapeeporn Payattakool et al. « A new method to measure the semantic similarity of GO terms ». In : *Bioinformatics* 23.10 (2007), p. 1274-1281. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btm087 (cf. p. 222).

[WLC+11] Q. Wang, J.-q. Liu, Z. Chen et al. « G-quadruplex formation at the 3' end of telomere DNA inhibits its extension by telomerase, polymerase and unwinding by helicase ». In : *Nucleic Acids Research* 39.14 (2011), p. 6229-6237. ISSN : 0305-1048. DOI : 10.1093/nar/gkr164 (cf. p. 26).

[WHF+15] Ye Wang, Yanpu Han, Enguo Fan et al. *Analytical strategies used to identify the readers of histone modifications : A review*. 2015. DOI : 10.1016/j.aca.2015.06.049 (cf. p. 36).

[WDA+16] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg et al. « The FAIR Guiding Principles for scientific data management and stewardship ». In : *Scientific Data* 3.1 (2016), p. 160018. ISSN : 2052-4463. DOI : 10.1038/sdata.2016.18 (cf. p. 82).

[YMK+16] Ryohei Yatsu, Shinichi Miyagawa, Satomi Kohno et al. « TRPV4 associates environmental temperature and sex determination in the American alligator ». In : *Scientific Reports* 5.1 (2016), p. 18581. ISSN : 2045-2322. DOI : 10.1038/srep18581 (cf. p. 27).

[YVM+18] Ki-Jun Yoon, Caroline Vissers, Guo-li Ming et al. « Epigenetics and epitranscriptomics in temporal patterning of cortical neural progenitor competence ». In : *The Journal of Cell Biology* 217.6 (2018), p. 1901-1914. ISSN : 0021-9525. DOI : 10.1083/jcb.201802117 (cf. p. 37, 38).

[YLQ+10] Guangchuang Yu, Fei Li, Yide Qin et al. « GOSemSim : an R package for measuring semantic similarity among GO terms and gene products ». In : *Bioinformatics* 26.7 (2010), p. 976-978. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btq064 (cf. p. 221).

[YHH+18] Miao Yu, Dali Han, Gary C Hon et al. « Tet-Assisted Bisulfite Sequencing (TAB-seq) ». In : *DNA Methylation Protocols*. Sous la dir. de Jörg Tost. New York, NY : Springer New York, 2018, p. 645-663. ISBN : 978-1-4939-7481-8. DOI : 10.1007/978-1-4939-7481-8_33 (cf. p. 62).

[YR14] Mary A. Yui et Ellen V. Rothenberg. « Developmental gene networks : A triathlon on the course to T cell identity ». In : *Nature Reviews Immunology* 14.8 (2014), p. 529-545. ISSN : 14741741. DOI : 10.1038/nri3702 (cf. p. 45, 46).

[ZMW+12] Jingli A. Zhang, Ali Mortazavi, Brian A. Williams et al. « Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity ». In : *Cell* 149.2 (2012), p. 467-482. ISSN : 00928674. DOI : 10.1016/j.cell.2012.01.056. arXiv : NIHMS150003 (cf. p. 45).

[ZG15] Yingming Zhao et Benjamin A Garcia. « Comprehensive Catalog of Currently Documented Histone Modifications ». In : *Cold Spring Harbor Perspectives in Biology* 7.9 (2015), a025064. ISSN : 1943-0264. DOI : 10.1101/cshperspect.a025064 (cf. p. 34, 35).

[ZT08] Jordanka Zlatanova et Amit Thakar. « H2A.Z : View from the Top ». In : *Structure* 16.2 (2008), p. 166-179. ISSN : 09692126. DOI : 10.1016/j.str.2007.12.008 (cf. p. 30).

ANNEXES

A Annuaire des techniques reposant sur le séquençage à haut débit

Revu, adapté et complété à partir de http://enseqlopedia.com/enseqlopedia/

Détection de faible niveau d'ARN

- CEL-Seq
- CirSeq
- CLaP
- CytoSeq
- Digital RNA Sequencing
- DP-Seq
- Drop-Seq
- Hi-SCL
- InDrop
- MARS-Seq
- Nuc-Seq
- PAIR
- Quartz-Seq
- scM&T-Seq
- SCRB-Seq
- scRNA-Seq
- scTrio-seq
- Smart-Seq
- Smart-Seq2
- snRNA-Seq
- STRT-Seq
- SUPeR-Seq
- TCR-LA-MC PCR
- TIVA
- UMI
- Div-Seq
- FRISCR
- TCR Chain Pairing
- AbPair

Modifications de l'ARN

- ICE
- MeRIP-Seq
- miCLIP-m6A
- Pseudo-Seq
- PSI-Seq

Structure de l'ARN

- CAP-seq
- Cap-Seq
- CIP-TAP
- PARS-Seq
- SPARE
- Structure-Seq / DMS-Seq
- CIRS-Seq
- icSHAPE
- SHAPE-MaP
- SHAPE-Seq

Transcription de l'ARN

- 2P-Seq
- 3'NT Method
- 3P-Seq
- 3Seq
- 3'-Seq
- 5'-GRO-Seq
- BruChase-Seq
- BruDRB-Seq
- Bru-Seq
- CAGE
- CHART
- ChIRP
- ClickSeq
- GRO-seq
- NET-Seq
- PAL-Seq
- PARE-Seq
- PEAT
- PRO-Cap
- PRO-Seq
- RAP
- RARseq
- RASL-Seq
- RNA-Seq
- SMORE-Seq
- TAIL-Seq
- TATL-Seq
- TIF-Seq
- TL-Seq
- 4sUDRB-Seq
- CaptureSeq
- cP-RNA-Seq
- FRT-Seq
- GMUCT
- mNET-Seq

Interactions ARN-Protéine

- AGO-CLIP
- CLASH
- CLIP-Seq / HITS-CLIP
- DLAF
- eCLIP
- hiCLIP
- iCLIP
- miR-CLIP
- miTRAP
- PAR-CLIP
- PIP-Seq
- Pol II CLIP
- RBNS
- Ribo-Seq / ARTSeq
- RIP-Seq
- TRAP-Seq
- TRIBE
- BrdU-CLIP
- HiTS-RAP
- irCLIP

Interactions Protéine-Protéine

- PD-Seq
- ProP-PD / PDZ-Seq

Réarrangements de séquence

- 2b-RAD
- CPT-seq
- ddRADseq
- Digenome-seq
- EC-seq
- hyRAD
- RAD-Seq
- Rapture
- RC-Seq
- Repli-Seq
- SLAF-seq
- TC-Seq
- Tn-Seq / INSeq
- Bubble-Seq
- NSCR
- NS-Seq
- Rep-Seq / Ig-Seq / MAF

Lésions de l'ADN

- BLESS
- DSB-Seq
- GUIDE-seq
- HTGTS
- LAM-HTGTS
- Break-seq
- SSB-Seq

Interactions ADN-protéine

- DNaseI Seq / DNase-Seq
- Pu-seq
- 3-C / Capture-C / Hi-C
- 4C-seq
- 5C
- ATAC-Seq / Fast-ATAC
- CATCH-IT
- Chem-Seq
- ChIA-PET
- ChIPmentation
- ChIP-Seq / HT-ChIP / ChIP-exo / Mint-ChIP
- DamID
- DNase I SIM
- FAIRE-Seq / Sono-Seq
- FiT-Seq
- HiTS-FLIP
- MINCE-Seq
- MNase-Seq / MAINE-Sequcleo-Sequc-Seq
- MPE-Seq
- NG Capture-C
- NOMe-Seq
- ORGANIC
- PAT-ChIP
- PB-Seq
- SELEX / SELEX-Seq / HT-SELEX
- THS-Seq
- UMI-4C
- X-ChIP-Seq

Méthylation de l'ADN

- Aba-Seq
- BisChIP-Seq / ChIP-BS-Seq / ChIP-BMS
- BSAS
- BSPP
- BS-Seq / Bisulfite-Seq / WGBS
- CAB-Seq
- EpiRADseq
- fCAB-Seq
- fC-CET
- fC-Seal
- hMeDIP-Seq
- JBP1-Seq
- MAB-Seq
- MBDCap-seq / MethylCap-Seq / MiGS
- MeDIP-Seq / DIP-Seq
- MIRA
- MRE-Seq / Methyl-Seq
- oxBS-Seq
- PBAT
- redBS-Seq / caMAB-Seq
- RRBS
- RRMAB-Seq
- TAB-Seq
- TAmC-Seq
- T-WGBS

Détection de faible niveau d'ADN

- Safe-SeqS
- scAba-Seq
- scATAC-Seq (Cell index variation)
- scATAC-Seq (Microfluidics variation)
- scBS-Seq
- scM&T-Seq
- scRC-Seq
- SMDB
- smMIP
- G&T-Seq
- DR-Seq
- G&T-Seq
- MALBAC
- MDA
- MIDAS / IMS-MDA / ddMDA
- scMi&T-Seq
- Drop-ChIP / scChIP-Seq
- Duplex-Seq
- MIPSTR
- nuc-Seq / SNES
- OS-Seq

Étude de l'activité des régions régulatrices

- STARR-Seq
- CapStarr-Seq
- MPRA
- CRE-Seq
- MPFD

B Autres analyses et articles

Des analyses bioinformatiques pour d'autres projets ont également été menées pendant ma thèse et ont donné lieu à un article scientifique publié et un autre en révision :

ARTICLE 7 L.T.M. DAO, A.O. GALINDO-ALBARRÁN, J.A. CASTRO-MONDRAGON et al. « Genome-wide characterization of mammalian promoters with distal enhancer functions ». In : *Nature Genetics* 49.7 (2017). ISSN : 15461718. DOI : 10.1038/ng.3884

ARTICLE 8 Aleksandra PEKOWSKA, Guillaume CHARBONNIER, Jean Christophe ANDRAU et al. « H3K4 methylation at enhancers and broad promoters ». In : *Letter to Molecular Cell* ()

Au cours du premier projet, c'est en expérimentant sur la difficulté de collaborer à plusieurs bioinformaticiens (avec Ariel O Galindo-Albarrán, Jaime A Castro-Mondragon et Aurélien Griffon) pour la production d'analyses bioinformatiques reproductibles sans cadre méthodologique bien défini qui m'a incité à réfléchir à la méthodologie présentée en SECTION 3.1.

Le second projet nécessitait de réaliser les analyses d'enrichissement fonctionnel de 252 échantillons de ChIP-Seq ciblant la modification d'histone H3K4me3 dans divers tissus humains sains et cancéreux issus des consortiums Blueprint et Roadmap. Le recours à une approche reproductible et automatisée a permis d'obtenir l'ensemble des résultats plus rigoureusement et moins fastidieusement que par des requêtes manuelles auprès de services en ligne.

nature genetics

Genome-wide characterization of mammalian promoters with distal enhancer functions

Lan T M Dao[1,6,7], Ariel O Galindo-Albarrán[1,7], Jaime A Castro-Mondragon[1], Charlotte Andrieu-Soler[2–4], Alejandra Medina-Rivera[5], Charbel Souaid[1], Guillaume Charbonnier[1], Aurélien Griffon[1], Laurent Vanhille[1], Tharshana Stephen[2,4], Jaafar Alomairi[1], David Martin[4], Magali Torres[1], Nicolas Fernandez[1], Eric Soler[2–4], Jacques van Helden[1], Denis Puthier[1] & Salvatore Spicuglia[1]

Gene expression in mammals is precisely regulated by the combination of promoters and gene-distal regulatory regions, known as enhancers. Several studies have suggested that some promoters might have enhancer functions. However, the extent of this type of promoters and whether they actually function to regulate the expression of distal genes have remained elusive. Here, by exploiting a high-throughput enhancer reporter assay, we unravel a set of mammalian promoters displaying enhancer activity. These promoters have distinct genomic and epigenomic features and frequently interact with other gene promoters. Extensive CRISPR–Cas9 genomic manipulation demonstrated the involvement of these promoters in the *cis* regulation of expression of distal genes in their natural loci. Our results have important implications for the understanding of complex gene regulation in normal development and disease.

Regulation of mammalian gene transcription is accomplished through the involvement of transcription start site (TSS)-proximal (promoter) and TSS-distal (enhancer) regulatory elements[1]. The original definition of a promoter entails the capability to induce local gene expression, whereas the term enhancer implies the property of activating gene expression at a distance. However, this basic dichotomy of *cis*-regulatory elements has been challenged by broad similarities between promoters and enhancers, such as DNA sequence features, chromatin marks, RNA polymerase II (Pol II) recruitment and bidirectional transcription[1–5]. Despite several findings suggesting that promoters might display enhancer activity[6–15], including experimental observations that enhancer elements can work as alternative promoters[16], it is unclear what fraction of promoters is concerned by this property and whether their enhancer activity is involved in distal gene regulation. The advent of high-throughput reporter assays, such as STARR-seq[13], has enabled the identification of enhancer activity solely on the basis of functionality instead of using epigenomics or location criteria[17]. We previously developed CapStarr-seq[18], a strategy coupling capture of a region of interest with STARR-seq, allowing efficient assessment of enhancer activity in mammals. By performing CapStarr-seq in several mammalian cell lines, we found that 2–3% of coding-gene promoters display enhancer activity in a given cell line. In comparison to classical promoters and distal enhancers, these TSS-overlapping enhancers (hereafter referred to as Epromoters) displayed distinct genomic and epigenomic features and were associated with stress response. By using comprehensive CRISPR–Cas9 genomic deletions, we demonstrated that Epromoters are involved in the *cis* regulation of the expression of distal genes in their natural context, therefore functioning as bona fide enhancers. Furthermore, human genetic variation within Epromoters was associated with a strong effect on distal gene expression. We suggest that regulatory elements with dual roles as transcriptional promoters and enhancers might ensure rapid and coordinated regulation of gene expression. These finding will enhance understanding of complex gene regulation in normal development and diseases and of how genetic variation influences the control of gene expression programs.

RESULTS

Mouse TSS-proximal DHSs display enhancer activity

To further decipher the complex relationship between proximal and distal regulatory regions for coding genes, we compared the proportions of enhancer activity for subsets of proximal and distal DNase I–hypersensitive sites (DHSs) in T cell precursors based on our previously published CapStarr-seq experiments performed in the mouse P5424 T cell and NIH-3T3 fibroblast cell lines[13,18] (**Fig. 1a**,**b** and **Supplementary Table 1**). We observed that the proportions of DHSs with enhancer activity were very similar for the proximal (<1 kb from the TSS) and distal subsets in P5424 cells (**Fig. 1c**, left). To avoid artifactual calling of enhancer activity due to sporadic transcription from the vector[19] or initiation from the promoter itself,

[1]Aix-Marseille University, INSERM, TAGC, UMR 1090, Marseille, France. [2]INSERM, UMR 967, CEA/DRF/iRCM, Laboratory of Molecular Hematopoiesis, Université Paris–Diderot, Université Paris–Saclay, Fontenay-aux-Roses, France. [3]Labex GR-Ex, Université Sorbonne Paris Cité, Paris, France. [4]IGMM, CNRS, Université de Montpellier, Montpellier, France. [5]Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Juriquilla, Mexico. [6]Present address: Vinmec Research Institute of Stem Cell and Gene Technology, Hanoi, Vietnam. [7]These authors contributed equally to this work. Correspondence should be addressed to S.S. (salvatore.spicuglia@inserm.fr).

Received 10 January; accepted 1 May; published online 5 June 2017; doi:10.1038/ng.3884

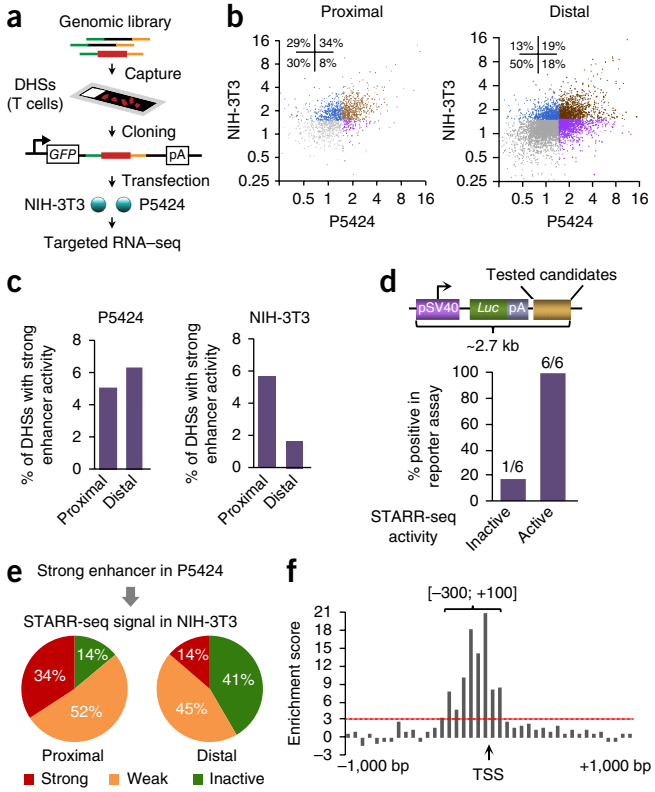


Figure 1 Comparison of proximal and distal DHSs with enhancer activity in two mouse cell lines. (**a**) Schematic of the CapStarr-seq protocol to assess the enhancer activity of promoters in NIH-3T3 and P5424 cells. (**b**) Scatterplots showing the STARR-seq signal ($\log_2$ scale) in P5424 and NIH-3T3 cells for proximal (left; 1,546 regions) and distal (right; 5,605 regions) DHSs. DHSs with enhancer activity in both cell lines (brown) or with activity specific to P5424 (purple) or NIH-3T3 (blue) cells are highlighted. DHSs with no enhancer activity are shown in gray. Quadrant panels show the percentage of regions in each subgroup. (**c**) Percentage of TSS-proximal and TSS-distal DHSs with strong enhancer activity (fold change >3) based on STARR-seq signal in P5424 (left) and NIH-3T3 (right) cells. (**d**) Top, reporter assay constructs. Bottom, summary of luciferase enhancer assays of proximal DHSs defined as active or inactive enhancers by STARR-seq in P5424 cells; detailed results are shown in **Supplementary Figure 1a**. Numbers correspond to the number of positive sites out of those tested. (**e**) Pie charts showing the distribution of enhancer activity in NIH-3T3 cells for the strong enhancers from TSS-proximal and TSS-distal DHSs identified in P5424 cells. (**f**) Distribution of the statistical enrichment of TSS-proximal DHSs for enhancer activity in NIH-3T3 cells. The significantly enriched region around the TSS is highlighted ($P < 0.001$, hypergeometric test).

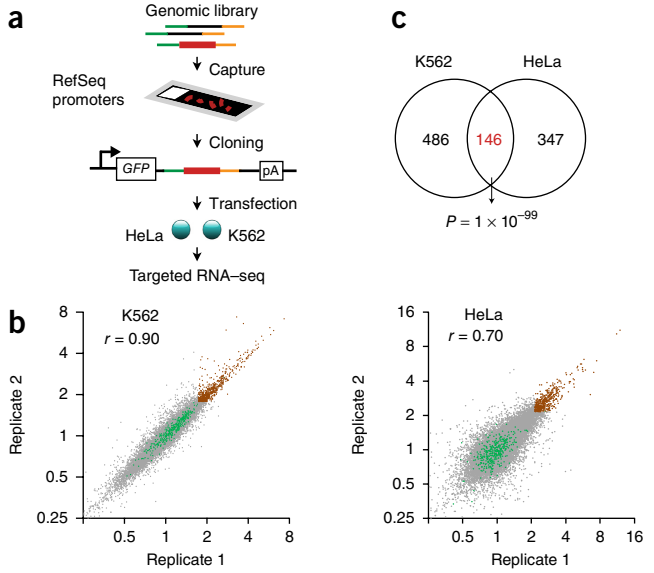


Figure 2 CapStarr-seq with human promoters. (**a**) Schematic of the CapStarr-seq strategy to assess the enhancer activity of human promoters in K562 and HeLa cells. (**b**) Scatterplots showing the correlation of two STARR-seq replicates in K562 (left) and HeLa (right) cells. The data plotted are the fold change in STARR-seq signal over the input signal ($\log_2$ scale). Promoters with enhancer activity in both replicates are shown in brown. Random genomic regions (green) did not display enhancer activity in these assays. (**c**) Venn diagram showing the number of Epromoters found in K562 and HeLa cells. The hypergeometric test *P* value for the overlap between the two sets is shown.

the STARR-seq procedure was implemented to ensure that the transcripts quantified initiated from the synthetic SCP1 promoter and were polyadenylated[9,13,18]. Reporter assays of CapStarr-seq-defined proximal enhancers confirmed their enhancer activity regardless of their orientation (**Fig. 1d** and **Supplementary Fig. 1a**). Distal enhancers identified in the P5424 T cell line were significantly enriched for lymphoid transcription factors, whereas proximal enhancers were generally depleted of these factors (**Supplementary Fig. 1b**), suggesting that the latter differ from classical distal enhancers. Consistently, the percentage of proximal T cell DHSs with enhancer activity in NIH-3T3 cells was higher than that for distal DHSs (**Fig. 1c**, right). Moreover, proximal enhancers in P5424 cells were found to be active more often in NIH-3T3 cells than distal enhancers (**Fig. 1e**) and the proportion of proximal enhancers active in both cell lines was highly significant ($P = 1.8 \times 10^{-106}$, hypergeometric test; **Fig. 1b**), suggesting that proximal enhancers are less specific to tissue type.

Notably, proximal enhancers were over-represented from −300 bp to +100 bp with respect to the TSS (**Fig. 1f**), roughly overlapping the core promoter regions where sense and antisense transcription initiation occurs and transcription factors usually bind[10,20,21]. Collectively, these results suggest that TSS-overlapping regions displaying enhancer activity, here defined as Epromoters, might represent regulatory elements with dual promoter and enhancer functions.

Assessment of the enhancer activity of coding-gene promoters

To characterize Epromoters in an unbiased manner, we performed CapStarr-seq with all promoters of RefSeq-defined human coding genes (−200 to +50 bp with respect to the TSS) in the two ENCODE cell lines K562 and HeLa (**Fig. 2a** and **Supplementary Fig. 2a**,**b**). The enhancer activity of each captured region was calculated as the fold change of the STARR-seq signal over the input signal. We observed high correlation between replicates in both cell lines (**Fig. 2b**). Epromoters were defined as promoters for which the fold change in signal for both replicates was beyond the inflexion point of ranked promoters (Online Methods). Using these stringent criteria, we found 632 (3%) and 493 (2.37%) Epromoters among 20,719 promoters analyzed in K562 and HeLa cells, respectively (**Fig. 2b**,**c** and **Supplementary Table 2**). Remarkably, a highly significant proportion of Epromoters were found in both cell types, suggesting a rather ubiquitous activity. No difference in the percentage of these promoters overlapping CpG islands or in the phylogenetic conservation of these promoters among mammalian species was observed as compated to non-Epromoters (**Supplementary Fig. 2c**).

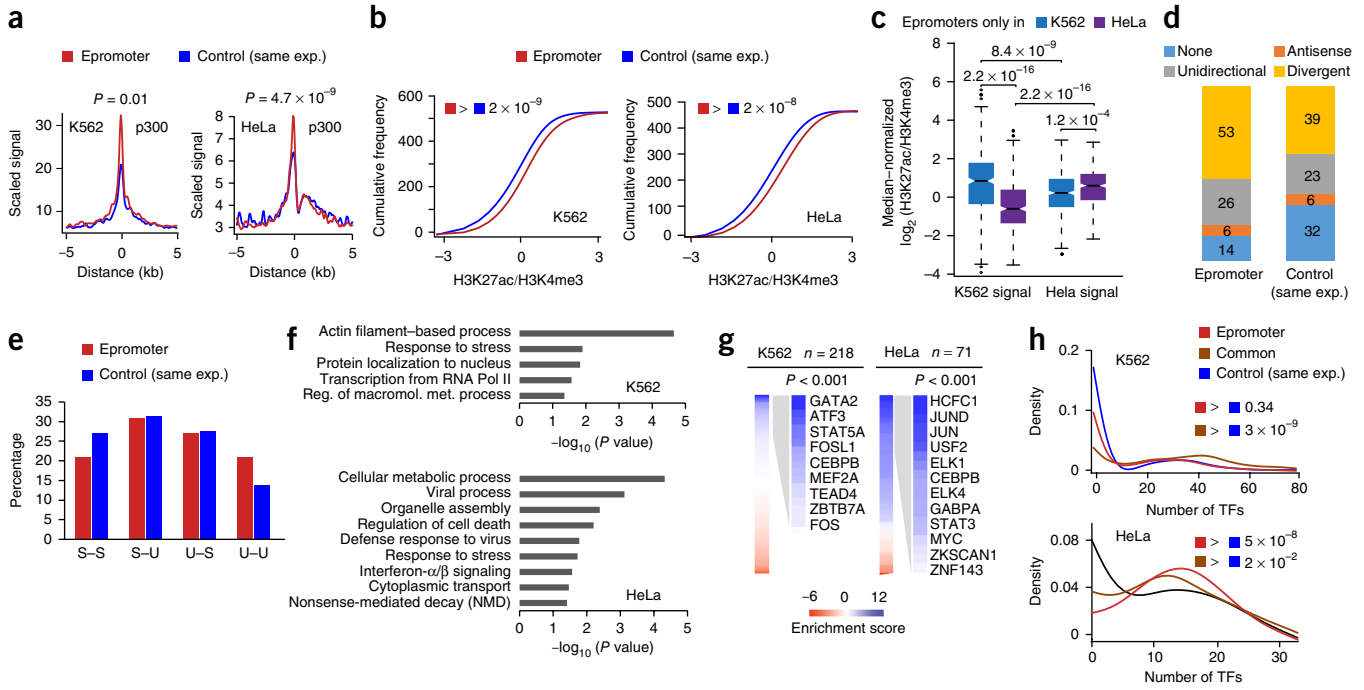


Figure 3 Genomic and epigenomic properties of Epromoters. (**a**) Average profiles of p300 in K562 (left) and HeLa (right) cells centered on the TSSs of Epromoters or control promoters with the same expression pattern for associated genes. Statistically significant differences were calculated for a region centered on the TSS (±250 bp) using two-sided Mann–Whitney *U* tests. (**b**) Cumulative plots showing the H3K27ac/H3K4me3 ratio at Epromoter and control sets in K562 (left) and HeLa (right) cells (Kolmogorov test). (**c**) H3K27ac/H3K4me3 ratios at Epromoters as a function of cell type (Mann–Whitney *U* test). Central values represent the median of the signal, the interquartile range (IQR) corresponds to the 75th to 25th percentile, and whiskers extend to the maximum and minimum values excluding outliers. (**d**) Percentage of the Epromoter and control sets corresponding to the TSS clusters from 5′ GRO–seq data defined by transcript overlap and orientation in HeLa cells. (**e**) Proportion of TSS pairs from each stability class (U, unstable; S, stable) associated with Epromoters and control sequences (with the same expression) in K562 cells. (**f**) Significantly enriched biological processes for Epromoter-associated genes in K562 (top) and HeLa (bottom) cells identified using g:Profiler. (**g**) Transcription factor enrichment (ENCODE data sets) at Epromoters in K562 and HeLa cells (two-sided Mann–Whitney *U* test). (**h**) Density plots showing the number of transcription factors (TFs) per promoter type in K562 (left) and HeLa (right) cells. 'Common' refers to the set of Epromoters active in both cell lines (Kolmogorov test).

Epromoters display specific genomic and epigenomic features

We next compared the epigenomic features of Epromoters with those of a set of matched control promoters chosen from a list of common promoters lacking enhancer activity in all replicates of both cell lines (non-Epromoters) but associated with genes with similar expression levels (**Supplementary Table 2**). Although Epromoters displayed similar levels of DNase I hypersensitivity and histone H3 trimethylation at lysine 4 (H3K4me3) signal as the control promoters, they were generally enriched for the enhancer-associated features monomethylation of histone H3 at lysine 4 (H3K4me1), acetylation of histone H3 at lysine 27 (H3K27ac) and p300 binding (**Fig. 3a** and **Supplementary Fig. 2d**). Consistent with these findings, Epromoters displayed a higher H3K27ac/H3K4me3 ratio (**Fig. 3b**) and were preferentially associated with a strong enhancer state in different ENCODE cell lines (**Supplementary Fig. 2e**). Moreover, Epromoters had a higher H3K27ac/H3K4me3 ratio in the cell type where they were found to be active (**Fig. 3c**). There was no significant bias of RefSeq-defined TSSs at Epromoters, as assessed by cap analysis of gene expression (CAGE) (**Supplementary Fig. 2f,g**), and 94.2% and 95.7% of K562 and HeLa Epromoters, respectively, overlapped with a TSS defined by the FANTOM consortium[22] (**Supplementary Fig. 2h** and **Supplementary Table 2**). However, 42.7% and 18.2% of the Epromoters active in HeLa and K562 cells lacked a TSS in the respective cell line. This might suggest that not all Epromoters are transcriptionally active (see below), although we cannot formally exclude the possibility that some individual cases could actually be promoter-proximal enhancers owing to sites being incorrectly annotated as TSSs. While the majority of Epromoters were found in genes with only one TSS, a substantial proportion were located in genes with two or more TSSs (**Supplementary Fig. 2i**), reminiscent of previous findings suggesting that alternative promoters might work as enhancers[16]. By analyzing 5′ global run-on with sequencing (5′ GRO–seq) data from HeLa cells[20], we found that the proportion of Epromoters with divergent transcripts was higher than that for control promoters ($P = 3.1 \times 10^{-5}$, hypergeometric test; **Fig. 3d**). Moreover, unstable divergent transcripts, which have been shown to be a hallmark of active enhancers[3], were over-represented among K562 Epromoters ($P = 5.8 \times 10^{-8}$, hypergeometric test; **Fig. 3e**). Altogether, the Epromoters defined by STARR-seq activity showed clear chromatin-associated enhancer features.

Gene Ontology (GO) analysis for Epromoter-associated genes primarily showed enrichment for basic processes (**Fig. 3f** and **Supplementary Table 3**), consistent with a previous STARR-seq study in *Drosophila melanogaster* reporting that many promoters of housekeeping genes can function as enhancers[9]. We also observed a significant enrichment ($P < 0.05$) for the cellular stress response in both cell lines. K562 Epromoters were particularly associated with genes encoding actin-binding cytoskeleton proteins, which have been shown to be rapidly and transiently upregulated upon heat shock response[23], whereas HeLa Epromoters were specifically associated with genes involved in type I and II interferon responses. Indeed, the main interferon-related genes were associated with Epromoters in HeLa cells, including *MX1*, *IRF9*, *JUND*, *ISG15*, *OAS* and the IFIT cluster of genes. Epromoter-associated

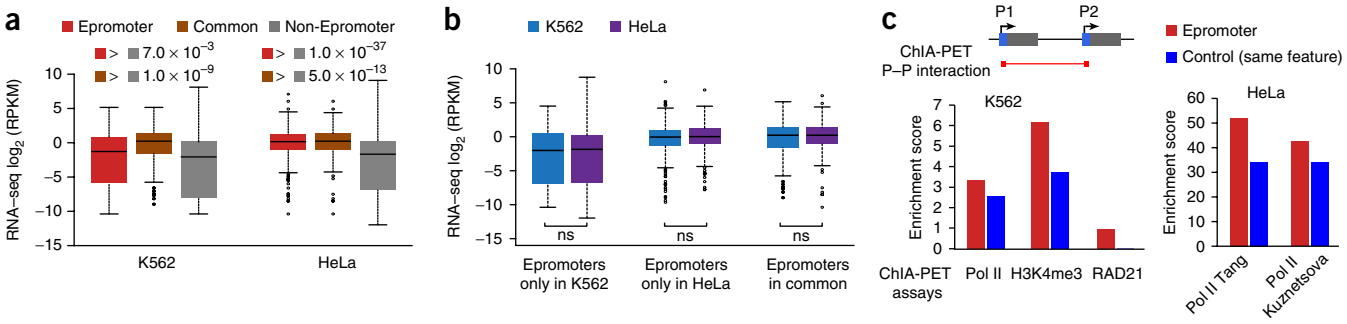


Figure 4 Expression of neighboring genes and promoter–promoter interactions. (**a**,**b**) Box plots comparing the expression levels of Epromoter- and non-Epromoter-associated genes in K562 and HeLa cells (**a**) and the expression of Epromoter-associated genes as a function of cell-line-specific Epromoter activity (**b**). The expression of genes associated with Epromoters active in both cell lines (Common) is also shown. Central values represent the median of the signal, the IQR corresponds to the 75th to 25th percentile, and whiskers extend to the maximum and minimum values excluding outliers (two-sided Mann–Whitney *U* test). ns, not significant. (**c**) Top, schematic of the strategy to identify promoter–promoter (P–P) interactions. Bottom, ChIA-PET enrichment of promoter–promoter interactions for a list of promoters associated with at least one Epromoter or a non-Epromoter from a control set with the same enriched features. The sources[37,38] of the published ChIA-PET data from HeLa cells are indicated by the name of the first author.

genes from HeLa cells were also enriched for transcriptional signatures including interferon- and tumor necrosis factor (TNF)-induced genes (**Supplementary Fig. 3c**). Differences in functional enrichment between K562 and HeLa cells might rely on cell-line-specific contexts. Indeed, interferon response genes are highly expressed in HeLa cells but not in K562 cells (**Supplementary Fig. 3a**,**b**), consistent with the fact that HeLa cells originated from a papillomavirus-infected tumor. We next assessed transcription factor enrichment at Epromoters using ENCODE data (**Fig. 3g**, **Supplementary Fig. 4a**,**b** and **Supplementary Table 4**). Consistent with the GO term enrichments, transcription factors involved in stress/interferon responses such as, JUN, FOS, IRF, ATF/CREB and STAT were enriched at Epromoters. We also found enrichment of specific transcription factor binding sites in general agreement with the transcription factor binding profiles, including strong enrichment for FOS/JUN motifs (**Supplementary Fig. 5a–d**). Moreover, Epromoters harbored a higher density of distinct bound transcription factors (**Fig. 3h**) and motifs (**Supplementary Fig. 5e**), consistent with their enhancer properties[24]. Thus, Epromoters display genomic and epigenomic features associated with enhancer activity. While Epromoters are located close to housekeeping genes, a subset of them might be involved in stress response. In this context, some Epromoters could be required to ensure strong and rapid transcriptional output in response to environmental or intrinsic cellular stimuli.

We next asked whether enhancer and promoter (transcription of the associated gene) activities are correlated for Epromoters. We first observed that Epromoter-associated gene expression was significantly higher than that associated with non-Epromoters (**Fig. 4a**). However, enhancer activity at Epromoters did not strictly correlate with the expression levels of associated genes (**Supplementary Fig. 6a**), and differences in the enhancer activity of Epromoters between the K562 and HeLa cell lines did not correlate with significant differences in gene expression (**Fig. 4b**). This suggests that the promoter and enhancer functions of Epromoters might be partially independent, indicating potential long-range regulation of nearby genes. Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) studies have shown that promoter–promoter interactions are a frequent phenomenon[6]. To test whether Epromoters are preferentially involved in promoter–promoter interactions, we analyzed available ChIA-PET data from the K562 and HeLa cell lines (**Supplementary Table 5**). In both cell lines, promoter–promoter interactions were found more frequently at Epromoters than at control promoters with similar levels of the corresponding histone modifications or transcription factors (**Fig. 4c** and **Supplementary Fig. 6b**). HeLa Epromoters were enriched for HCFC1 and ZNF143 (**Fig. 3g**), two associated factors suggested to be involved in looping[25,26].

Epromoters function as bona fide enhancers

To experimentally address the role of Epromoters in the long-distance regulation of gene expression, we performed CRISPR–Cas9-mediated genomic deletion of the *FAF2* Epromoter, for which clear interactions with the promoters of the *NOP16*, *CLTB* and *RNF44* genes were observed by ChIA-PET in both cell lines (**Fig. 5a** and **Supplementary Fig. 7**). Deletion of the *FAF2* Epromoter (ΔEp.*FAF2*) resulted in significant reduction of *RNF44* expression in both cell lines, while *NOP16* expression was reduced only in HeLa cells (**Fig. 5b**). A decrease in H3K27ac at the *RNF44* promoter in ΔEp.*FAF2* K562 cells was also observed (**Fig. 5c**). We confirmed the interaction between the *FAF2* and *RNF44* promoters by circularized chromosome conformation capture and sequencing (4C–seq) in K562 cells, using either the *FAF2* or *RNF44* promoter region as the viewpoint, and observed almost complete loss of this interaction in the two ΔEp.*FAF2* clones (**Fig. 5d** and **Supplementary Fig. 8a**,**b**). Consistent with these findings, the *FAF2* Epromoter was able to activate the *RNF44* promoter, as demonstrated by luciferase assay (**Fig. 5e**). Note that no luciferase activity was detected for the *RNF44* promoter vector without the *FAF2* Epromoter, ensuring that the observed enhancer activity is not due to spurious transcription[19]. Deletion of the endogenous *RNF44* promoter did not affect *FAF2* expression (**Fig. 5f**), indicating that distal regulation is directional. Moreover, epigenetic marks were correlated between the *FAF2* and *RNF44* loci across different cell lines (**Supplementary Fig. 8c**). To test *in vivo* whether Epromoters might function independently of their orientation, we inverted the *FAF2* Epromoter (including exon 1 of the gene) within its endogenous context in K562 cells (**Supplementary Fig. 7i–k**). Inversion of the *FAF2* Epromoter completely abolished *FAF2* expression and slightly but significantly reduced *RNF44* expression (**Fig. 5g**). However, *FAF2–RNF44* interaction was maintained in the inversion clones (**Supplementary Fig. 8b**) and *RNF44* expression was significantly higher than in the deletion clones (**Fig. 5h**), suggesting that *in vivo* enhancer activity is partially retained with the inverted configuration of the *FAF2* Epromoter. Finally, rescue of *FAF2* expression in either ΔEp.*FAF2* or Inv.Ep.*FAF2* clones did not affect *RNF44* expression levels (**Fig. 5h**), indicating direct regulation of neighboring gene expression by the *FAF2* Epromoter.

To generalize our finding, we targeted three additional Epromoters with promoter–promoter interactions found either in both cell lines

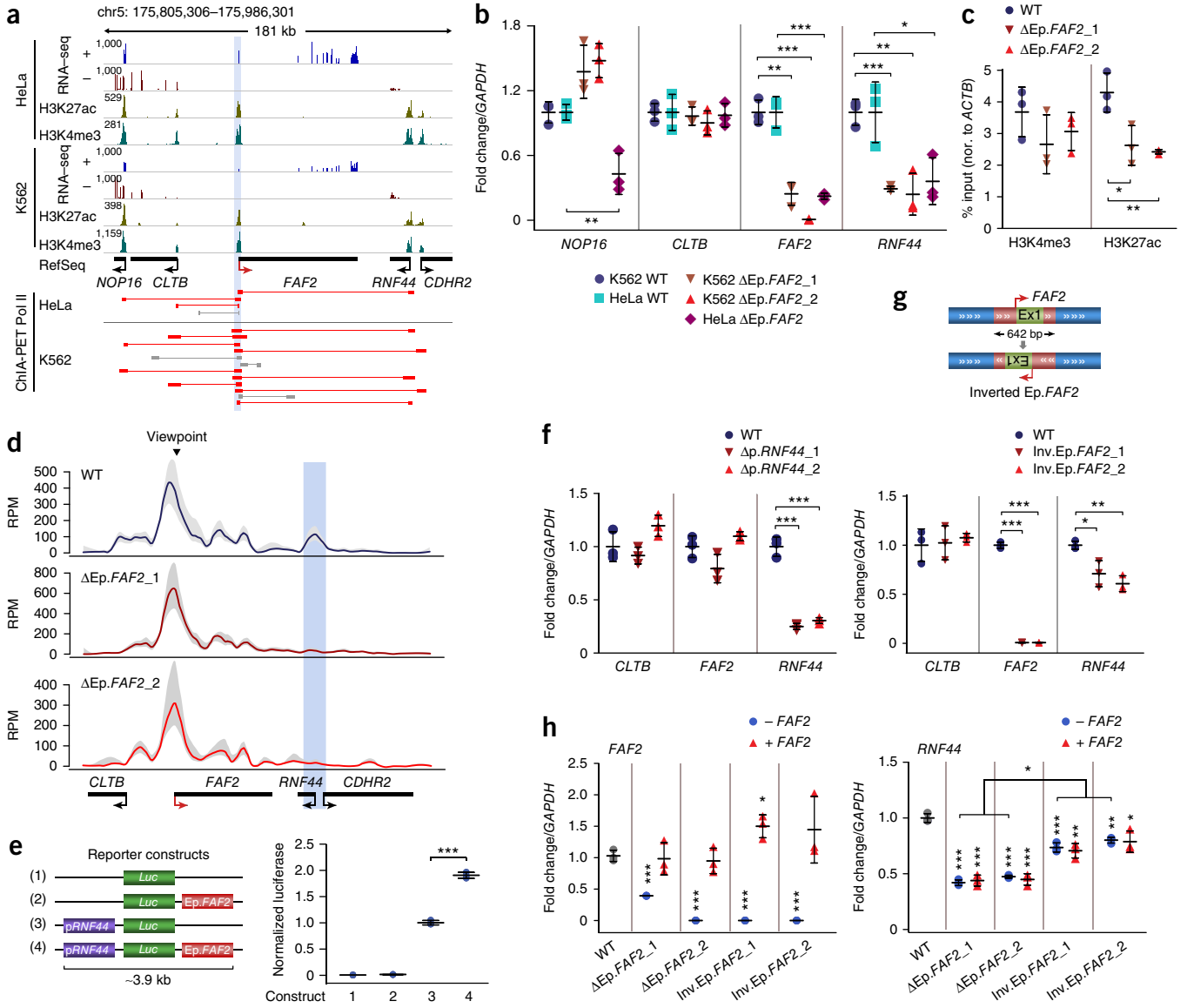


Figure 5 Epromoters function as bona fide enhancers and regulate distal gene expression. (**a**) Genomic tracks for RNA–seq, ChIP–seq and ChIA-PET Pol II around the *FAF2* locus. (**b**) qPCR analysis of gene expression in wild-type (WT) and ΔEp.*FAF2* cell clones (the last numbers indicate the number of the independent clone). (**c**) ChIP–qPCR analysis of H3K4me3 and H3K27ac marks at the *RNF44* promoter in K562 cells. (**d**) 4C–seq analysis of *FAF2* Epromoter interactions in wild-type K562 cells and two ΔEp.*FAF2* clones. The genomic tracks show the LOESS-normalized merge of two technical replicates (see **Supplementary Fig. 8a** for the raw data). RPM, reads per million; gray shading, 40% and 60% quantiles. The *FAF2*–*RNF44* interaction was significant in wild-type cells ($P < 1 \times 10^{-4}$) but not in ΔEp.*FAF2* clones. (**e**) Luciferase reporter assays testing the enhancer activity of the *FAF2* Epromoter coupled with the *RNF44* promoter. (**f**) qPCR analysis of gene expression in wild-type and Δp.*RNF44* K562 clones. (**g**) Top, schematic of knock-in of the inverted *FAF2* Epromoter. Bottom, qPCR analysis of wild-type and Inv.Ep.*FAF2* clones. Note that the intrinsic promoter activity is conserved as increased upstream antisense expression in the Inv.Ep.*FAF2* clones (**Supplementary Fig. 7k**). (**h**) qPCR analysis of the relative gene expression of *FAF2* (left) and *RNF44* (right) in wild-type, ΔEp.*FAF2* and Inv.Ep.*FAF2* clones, in the presence or absence of *FAF2* cDNA. For the graphs in **b**, **c** and **e**–**h**, each point represents one of three independent RNA/cDNA preparations. Error bars, s.d.: $***P < 0.001$, $**P < 0.01$, $*P < 0.1$, two-sided Student's *t* test.

(*CSDE1* and *TAGLN2*) or only in K562 cells (*BAZ2B*). Deletion of the *CSDE1* Epromoter resulted in significant reduction of *BCAS2* and *SIKE1* expression in both cell lines, while *NRAS* expression was reduced only in HeLa cells (**Supplementary Fig. 9a**,**b**). Deletion of the *TAGLN2* Epromoter led to significant reduction of *PIGM* and *PEA15* expression, while *DUSP23* was upregulated (**Supplementary Fig. 9c**,**d**). These results show specific regulation, as no effect was observed on *DCAF8*, another neighboring gene not interacting with the *TAGLN2* Epromoter. Although deletion of the *BAZ2B* Epromoter did not result in loss of *BAZ2B* expression, likely owing to alternative promoter usage, *MARCH7* expression was significantly reduced (**Supplementary Fig. 9f**–**i**). Finally, the presence of CAGE-defined TSSs and spliced transcripts originating from the Epromoter regions (**Supplementary Fig. 9j**) confirmed that these loci are bona fide promoters and not incorrectly annotated distal enhancers.

Epromoters regulate distal interferon response genes

Expression of interacting gene pairs was highly correlated regardless of whether the association involved an Epromoter (**Supplementary Fig. 10a**). We therefore explored the possibility of a coordinated

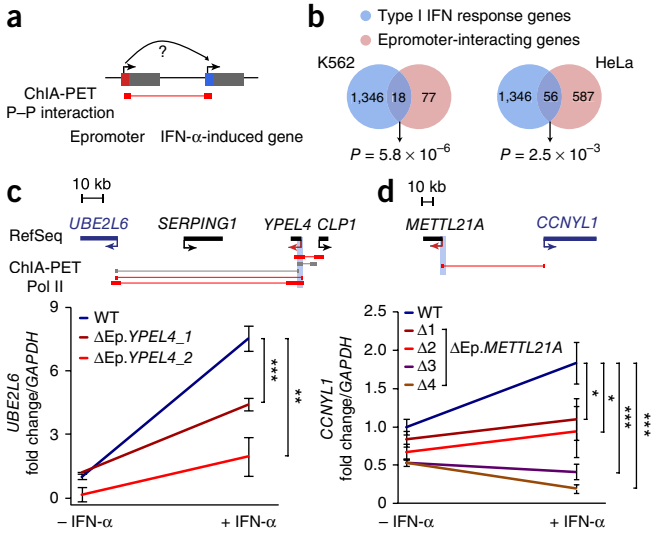


Figure 6 Epromoters are involved in a long-range response to IFN-α signaling. (**a**) Schematic of the strategy to identify IFN-α-induced genes associated with Epromoters combining ChIA-PET and STARR-seq data. (**b**) Venn diagrams showing the overlap between Epromoter-interacting genes and interferon response genes in K562 and HeLa cells (hypergeometric test). (**c**,**d**) qPCR analysis of the expression levels of the Epromoter-interacting genes *UBE2L6* (**c**) and *CCNYL1* (**d**) in wild-type and knockout clones with and without IFN-α stimulation. Error bars, s.d. ($n = 3$ independent RNA/cDNA preparations): $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.1$, two-sided Student's t test. The relative locations of genes and ChIA-PET interactions with Epromoters are shown above; Epromoters are highlighted as red arrows.

response to external stimuli mediated by Epromoters. We initially observed that key interferon response genes were found in interacting clusters associated with HeLa Epromoters (IFIT gene cluster, *ISG15–HES4* and *IRF9–PSME2–RNF31*; **Supplementary Fig. 10b**), suggesting that Epromoters are involved in the coordinated response to interferon signaling and consistent with an active interferon response in these cells (**Supplementary Fig. 3a–c**). To address whether Epromoters are involved in the activation of distal interferon-induced genes, we looked for interferon (IFN)-α-induced genes in promoter–promoter interactions with Epromoters (**Fig. 6a**). We found a significant proportion of Epromoters interacting with interferon response genes in both cell lines (**Fig. 6b** and **Supplementary Table 6**). We reasoned that in K562 cells some Epromoters might be required for proper activation of distally associated interferon response genes. To test this hypothesis, we selected two IFN-α response genes, *UBE2L6* (interacting with the *YPEL4* Epromoter) and *CCNYL1* (interacting with the *METTL21A* Epromoter) that were induced ~7.5- and ~2-fold after IFN-α treatment, respectively (**Fig. 6c**,**d**). Deletion of the interacting Epromoters did not result in consistent changes in *UBE2L6* or *CCNYL1* expression in non-stimulated cells; however, induction of these genes upon IFN-α treatment was severely reduced (**Fig. 6c**,**d** and **Supplementary Fig. 10c–e**). We also noted that *CLP1*, a non-interferon-responsive gene located close to *YPEL4*, displayed significant upregulation in clones in which the *YPEL4* Epromoter was deleted both before and after interferon treatment, suggesting that enhancer–promoter contacts may have been rewired in the Epromoter-knockout clones (**Supplementary Fig. 10d**). Overall, these results show that some Epromoters are involved in the rapid activation of distal genes upon external stress stimuli, supporting a model in which preformed loops between Epromoters and target genes precede gene induction[27].

To further rule out a plausible indirect effect mediated by Epromoter-associated genes, we analyzed allelic expression of wild-type cells and those homozygous and heterozygous for Epromoter deletion for cases where distally regulated genes harbored a SNP within the transcribed region in the K562 cell line. These genes included *PIGM* and *UBE2L6*, which are regulated by the *TAGLN2* and *YPEL4* Epromoters, respectively. In both cases, we found that allelic expression was significantly biased in the heterozygous clones (**Supplementary Figs. 9e** and **10f**), thus suggesting *cis*-specific regulation by the Epromoters.

Genetic variants within Epromoters influence distal genes

Genetic variants lying within Epromoters might influence the expression of distal genes. To address this possibility, we isolated all interacting promoter pairs (using ChIA-PET data) and those that were associated with expression quantitative trait loci (eQTLs) (**Fig. 7a**,**b** and **Supplementary Table 7**). We found that Epromoters more frequently overlapped eQTLs affecting the expression of distal interacting genes and that the eQTLs associated with Epromoters had a significantly stronger effect on distal gene expression than the eQTLs associated with non-Epromoters (**Fig. 7c**). We found eQTLs within three experimentally validated Epromoters (*METTL21A*, *BAZ2B* and *CSDE1*). K562 cells harbor a heterozygous eQTL variant within the *CSDE1* Epromoter (**Fig. 7d–f**) that results in DNase I accessibility and binding of transcription factors with a bias toward the reference allele (**Fig. 7g**). Allelic replacement of the reference allele resulted in decreased expression of *CSDE1* and *SIKE1* (**Fig. 7h**), as predicted by the eQTL association study. Similarly, deletion of the eQTL variant within *BAZ2B* resulted in reduced expression of the distal associated gene *MARCH7* (**Fig. 7i**). To further explore the implications of Epromoter-associated eQTLs, we analyzed *in silico* the probability of affecting transcription factor binding. We observed that SNPs potentially affecting transcription factor binding within Epromoters were biased toward having a positive effect (β) on distal gene expression, whereas this bias was not observed with non-Epromoters (**Fig. 7j**,**k**). Collectively, these results corroborate the functional relevance of eQTL-overlapping Epromoters, raising the intriguing possibility that disease-associated variants lying within Epromoters might directly influence distal gene expression.

DISCUSSION

Here, by implementing a high-throughput reporter assay, we shed light on and characterize a set of mammalian coding-gene promoters carrying both an intrinsic ability to drive local transcription (act as a promoter) and to activate distal gene expression (act as an enhancer). These elements have distinct genomic and epigenomic features, which distinguish them from other promoters and from classical distal enhancers (**Figs. 1**, **3** and **4**). For six of these loci, we demonstrated that they act as bona fide enhancers regulating distal gene expression *in vivo*. Remarkably, some Epromoters were found to regulate the expression of several distal genes (*FAF2*, *CSDE1* and *TAGLN2* Epromoters) over large genomic distances (up to 300 kb in the case of the *TAGLN2* Epromoter), implying that they might function as regulatory hubs. Our results extend and support the increasing amount of evidence pointing to a unified model of transcriptional regulation, highlighting broad similarities between enhancers and promoters[1–5]. Furthermore, previous studies based on the frequency of promoter–promoter interactions[6,12,14] or epigenetic features[7,10] suggested that some promoters might display enhancer function. Consistent with our findings, previous reporter assays also showed enhancer activity from TSS-proximal regions[9,11,13]. It is also worth noting that several

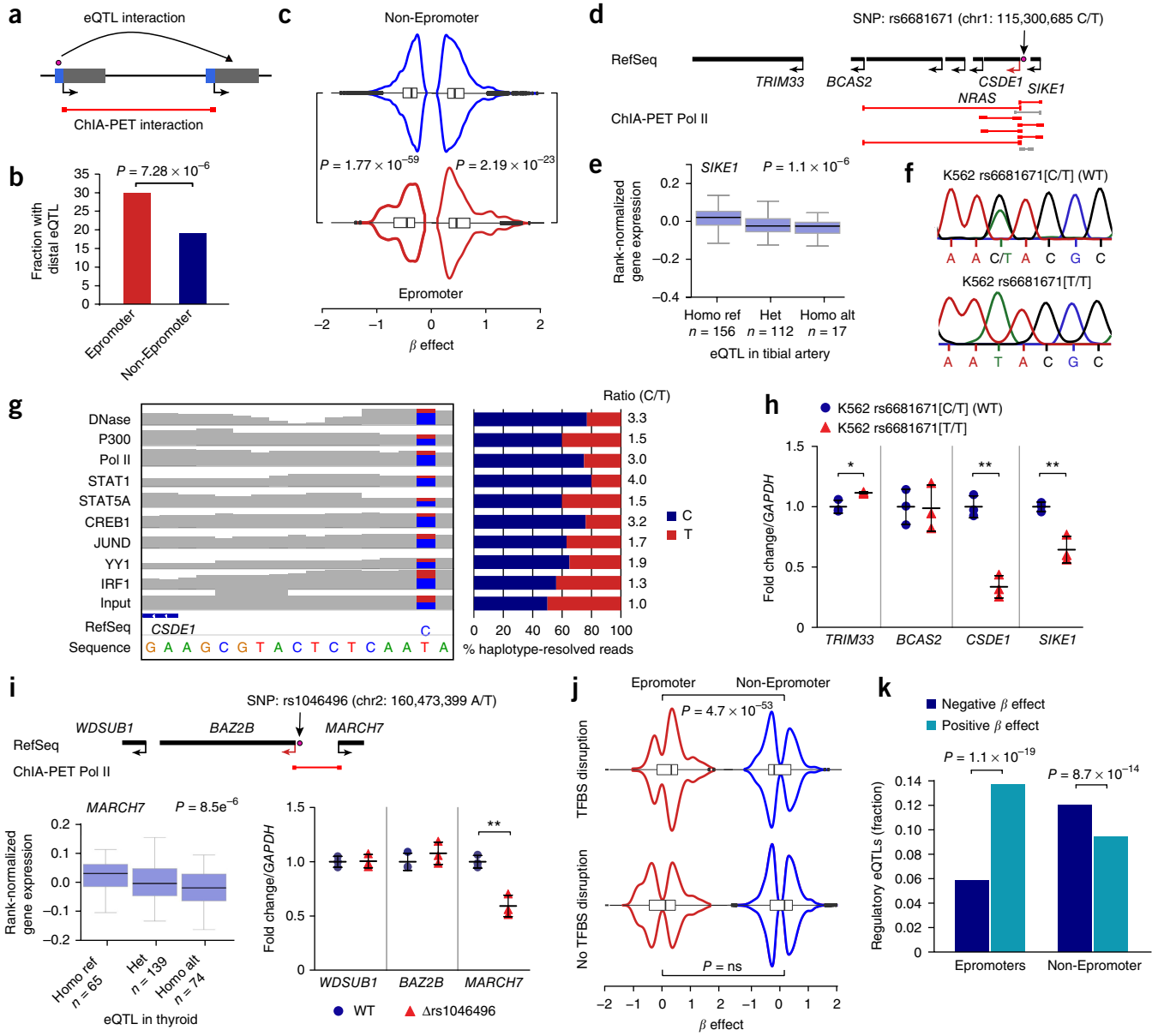


Figure 7 eQTL association within Epromoters. (**a**) Schematic of the eQTLs assessed. (**b**) Frequency of promoters having eQTLs associated with distal gene expression. Statistical significance was assessed by testing for equality of proportions. (**c**) Effects associated with eQTLs lying within promoter pairs with ChIA-PET interactions. Statistical significance was assessed independently for negative and positive scores using two-sided Mann–Whitney *U* tests. (**d**) Schematic of the eQTL SNP (rs6681671) within the *CSDE1* Epromoter associated with *SIKE1* expression. (**e**) eQTL data retrieved from the GTEx Portal. Ref, reference; Alt, alternate. Central values represent the median of the signal, the IQR corresponds to the 75th to 25th percentile, and whiskers extend to the maximum and minimum values excluding outliers. (**f**) Sequence chromatograms of wild-type and mutant K562 clones. (**g**) Coverage tracks from IGV (left) and histograms (right) showing the frequency of haplotype-resolved reads at SNP rs6681671 from the indicated ENCODE data in K562 cells. (**h**) qPCR analyses of gene expression in wild-type cells and eQTL mutants. (**i**) The eQTL SNP within the *BAZ2B* Epromoter associated with *MARCH7* expression is shown as in **d**, **e** and **h**. Δrs1046496 denotes deletion of SNP rs1046496 in K562 cells. For **h** and **i**, error bars show s.d. ($n = 3$ independent RNA/cDNA preparations): $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.1$, two-sided Student's *t* test. (**j**) Effects depending on whether the eQTL disrupts a transcription factor binding site (TFBS). Statistical significance was assessed by a one-sided Mann–Whitney *U* test and corrected for multiple testing using the Benjamini–Hochberg method. (**k**) Fraction of regulatory eQTLs (affecting transcription factor binding) with positive and negative β values. Statistical significance was assessed by Fisher's exact test.

well-characterized enhancers of rapidly induced genes, including metallothioneins, histones of early cleavage stages, viral immediate-early genes (from SV40 and some cytomegaloviruses and retroviruses), heat-shock genes and the antiviral interferon genes, are located very close to the TSS[8]. Our study clearly shows that reporter-assay-based approaches can lead to the identification of TSS-overlapping promoters with bona fide enhancer activity *in vivo*.

It is possible that previous studies deleting large genomic regions overlapping promoters have underestimated the potential enhancer function of these regulatory elements (for example, see ref. 28). To our knowledge, only two studies have reported dual promoter and enhancer functions for the same regulatory elements in their endogenous context in mammals. Kowalczyk *et al.* showed that intragenic enhancers frequently act as alternative, tissue-specific promoters,

although these promoters produced a class of noncoding transcript[16]. Another study, published while this manuscript was under review, reported frequent distal *cis* regulation by loci associated with long noncoding RNAs (lncRNAs) and, to a lesser extent, coding genes[15]. Interestingly, using genetic manipulations in mouse embryonic stem cells, the authors demonstrated that these effects did not require the specific transcripts themselves, but instead involved general processes associated with their production, including enhancer-like activity of the gene promoters, the process of transcription and splicing of the transcript. On the basis of these findings, it is plausible that some of the experimentally validated Epromoters might function by other processes than enhancer-like activity. Further studies based on our catalog of Epromoters will be needed to precisely characterize the mechanisms by which these elements regulate distal gene expression.

Could it be that some of the Epromoters identified in this study are actually incorrectly annotated as promoter-proximal enhancers? The selection of captured TSS-encompassing regions was based on the annotation of coding-gene transcripts by RefSeq. Despite this conservative approach, we cannot completely rule out the possibility of erroneous TSS calls, leading to incorrectly annotated promoter-proximal enhancers. The vast majority of the tested regions overlapped with a CAGE-defined TSS. Moreover, the experimentally validated Epromoters (with the exception of *YPEL4*) did overlap with CAGE TSSs identified in the corresponding cell lines and were associated with spliced and polyadenylated transcripts of the nearest gene, confirming that these particular loci are bona fide promoters (**Supplementary Fig. 9j**). The analyses of CAGE-based TSSs also found that a substantial number of Epromoters did not display CAGE signal in the cell line where they were active (**Supplementary Fig. 2h**), in line with the poor correlation between Epromoter activity and expression of the closest gene (**Supplementary Fig. 6a**). However, we also found good correlation between gene pairs of interacting promoters involving at least one Epromoter (**Supplementary Fig. 10a**). This apparent contradiction might be explained by the existence of two types of Epromoters. One type might coordinately regulate the expression of several genes, including the closest one, therefore displaying simultaneous promoter and enhancer activities. For example, in the case of the *FAF2* Epromoter, expression of the *FAF2* and *RNF44* genes is positively correlated across different cell types (**Supplementary Fig. 8c**). Another type might display independent promoter and enhancer activities; in these cases, an active Epromoter could be associated with a silent or weakly expressed gene. For example, in the case of the *YPEL4* Epromoter, the *YPEL4* gene is not expressed in K562 cells, but the Epromoter regulates the expression of *UBE2L6* (**Fig. 6c**,**d**). This is reminiscent of a previous work showing that the same genomic region can have the epigenetic features of an enhancer or a promoter in different tissues[7].

In the current model of transcription factories, the regulatory regions of neighboring genes are clustered together and each contributes to the expression of multiple genes by increasing the local concentration of regulatory factors and RNA polymerases[29]. In this context, multigene interaction complexes have provided a structural framework for the postulated transcription factories[6]. Our results showing that Epromoters interact more frequently with other distal promoters (**Fig. 4**) and that eQTLs associated with Epromoters have a significantly stronger effect on distal gene expression (**Fig. 7**) support a model in which Epromoters have a key role within transcription factories. Whether Epromoter–promoter interactions rely on mechanisms similar to those previously shown for enhancer–promoter interactions[30] and what the specific contribution of Epromoters to the functioning of transcription factories is will need to be investigated in the future.

We found that a significant proportion of Epromoters interacted with interferon response genes in both cell lines analyzed (**Fig. 6**). Interferon response genes are not induced at baseline in K562 cells, suggesting the existence of preformed chromatin loops preceding gene induction of interferon response genes, in line with previous findings showing that TNF-α-responsive enhancers are already in contact with their target promoters before signaling[27]. This is illustrated by the examples of the *YPEL4* and *METTL21A* Epromoters, which were found to interact with the promoters of distal IFN-α-responsive genes in unstimulated K562 cells, thus preceding gene activation. Further studies will be required to identify the transcription factors and (epigenetic) mechanisms involved in these interactions.

URLs. ENCODE, https://www.encodeproject.org/; R Core Team, https://www.R-project.org/; Reactome: interferon αβ signaling, http://www.broadinstitute.org/gsea/msigdb/cards/REACTOME_INTERFERON_ALPHA_BETA_SIGNALING.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank J.-C. Andrau and J. Imbert for critical reading of the manuscript. We thank the IBiSA 'Transcriptomics and Genomics Marseille-Luminy' (TGML) platform for sequencing of CapStarr-seq samples and the cell biology platform for management of cell culture. Work in the laboratory of S.S. was supported by recurrent funding from INSERM and Aix-Marseille University and by specific grants from the European Union's FP7 Programme (282510-BLUEPRINT), ARC (PJA 20151203149) and A*MIDEX (ANR-11-IDEX-0001-02). L.T.M.D., A.G. and G.C. were supported, respectively, by Vietnam International Education Development (911), CONACYT and FRM.

AUTHOR CONTRIBUTIONS

L.T.M.D. and S.S. conceptualized and designed the experiments. L.T.M.D. performed most experimental work. A.O.G.A. performed most bioinformatics analyses. J.A.C.-M. and J.v.H. performed motif analysis. C.A.-S., T.S., D.M. and E.S. performed 4C–seq experiments and analyses. C.S., A.G. and L.V. performed and analyzed data from mouse CapStarr-seq. J.A., M.T. and N.F. contributed to CRISPR screening and analyses of allelic expression. G.C. and D.P. performed ChIA-PET analyses. A.M.R. performed eQTL analysis. All authors contributed to reading, discussion and commenting on the manuscript. L.T.M.D. and S.S. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Kim, T.K. & Shiekhattar, R. Architectural and Functional Commonalities between Enhancers and Promoters. *Cell* **162**, 948–959 (2015).
2. Andersson, R. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *BioEssays* **37**, 314–323 (2015).
3. 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).
4. Pekowska, A. *et al*. H3K4 tri-methylation provides an epigenetic signature of active enhancers. *EMBO J.* **30**, 4198–4210 (2011).
5. Koch, F. *et al*. Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat. Struct. Mol. Biol.* **18**, 956–963 (2011).
6. Li, G. *et al*. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).
7. Leung, D. *et al*. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**, 350–354 (2015).
8. Schaffner, W. Enhancers, enhancers - from their discovery to today's universe of transcription enhancers. *Biol. Chem.* **396**, 311–327 (2015).

9. Zabidi, M.A. *et al.* Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2015).
10. 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).
11. Nguyen, T.A. *et al.* High-throughput functional comparison of promoter and enhancer activities. *Genome Res.* **26**, 1023–1033 (2016).
12. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
13. Arnold, C.D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
14. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582–597 (2015).
15. Engreitz, J.M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452–455 (2016).
16. Kowalczyk, M.S. *et al.* Intragenic enhancers act as alternative promoters. *Mol. Cell* **45**, 447–458 (2012).
17. Dailey, L. High throughput technologies for the functional discovery of mammalian enhancers: new approaches for understanding transcriptional regulatory network dynamics. *Genomics* **106**, 151–158 (2015).
18. Vanhille, L. *et al.* High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nat. Commun.* **6**, 6905 (2015).
19. Nejepinska, J., Malik, R., Moravec, M. & Svoboda, P. Deep sequencing reveals complex spurious transcription from transiently transfected plasmids. *PLoS One* **7**, e43283 (2012).
20. Duttke, S.H. *et al.* Human promoters are intrinsically directional. *Mol. Cell* **57**, 674–684 (2015).
21. Roy, A.L. & Singer, D.S. Core promoters in transcription: old problem, new insights. *Trends Biochem. Sci.* **40**, 165–171 (2015).
22. Forrest, A.R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
23. Mahat, D.B., Salamanca, H.H., Duarte, F.M., Danko, C.G. & Lis, J.T. Mammalian Heat Shock Response and Mechanisms Underlying Its Genome-wide Transcriptional Regulation. *Mol. Cell* **62**, 63–78 (2016).
24. Hardison, R.C. & Taylor, J. Genomic approaches towards finding cis-regulatory modules in animals. *Nat. Rev. Genet.* **13**, 469–483 (2012).
25. Michaud, J. *et al.* HCFC1 is a common component of active human CpG-island promoters and coincides with ZNF143, THAP11, YY1, and GABP transcription factor occupancy. *Genome Res.* **23**, 907–916 (2013).
26. Whalen, S., Truty, R.M. & Pollard, K.S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* **48**, 488–496 (2016).
27. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).
28. Li, Y. *et al.* CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS One* **9**, e114485 (2014).
29. Feuerborn, A. & Cook, P.R. Why the activity of a gene depends on its neighbors. *TIG* **31**, 483–490 (2015).
30. Kagey, M.H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430–435 (2010).

ONLINE METHODS

Cell culture. K562 (CCL-243), a chronic myelogenous leukemia cell line, and HeLa-S3 (CCL-2.2), a cervical carcinoma cell line, were obtained from the American Type Culture Collection (ATCC) and maintained in RPMI (Gibco) supplemented with 10% FBS (Gold, PAA) at 37 °C, 5% CO_2. The P5424 T cell line[31] was cultured as described previously[18]. Cells were passaged every 2–3 d and routinely tested for mycoplasma contamination. For cell stimulation, 1×10^6 K562 cells were incubated with IFN-α (Sigma, SRP4594) at 50 ng/ml for 6 h.

Mouse CapStarr-seq. Enhancer activity in the mouse P5424 and NIH-3T3 cell lines was retrieved from our previously published CapStarr-seq data[18]. DHS genomic regions were separated into TSS distal (>1 kb) and proximal (<1 kb) while keeping the previous definition of enhancer activity (**Supplementary Table 1**).

Luciferase reporter assays. For the reporter assays related to **Figure 1c** and **Supplementary Figure 1a**, proximal-defined DHS regions overlapping TSSs were selected on the basis of CapStarr-seq activity in the P5424 cell line. The tested candidates were amplified from mouse genomic DNA and cloned downstream of the luciferase gene in the pGL3-Promoter vector (Promega) at the BamHI site. For the reporter assays related to **Figure 5e**, the human *RNF44* promoter (1,294 bp, chr5:176,537,245–176,538,538) and/or *FAF2* Epromoter (661 bp, chr5:176,447,822–176,448,482) was amplified from K562 genomic DNA and cloned into the pGL3-Basic vector (Promega). The *RNF44* promoter was cloned upstream of the luciferase gene at the MluI–BglII sites, and the *FAF2* Epromoter was cloned downstream of the luciferase gene at the SalI site. For cell transfection, a total of 1×10^6 P5424 or K562 cells were cotransfected 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 described previously[18], and conditions for K562 cells are described below (human CapStarr-seq). Twenty-four hours after 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 increase in relative luciferase signal over the pGL3-Promoter vector (**Supplementary Fig. 1a**) or *RNF44*-pGL3-Basic vector (**Fig. 5e**) with s.d. Student's *t* tests (one-sided, unpaired) from three independent transfections were used to calculate significance.

Human CapStarr-seq. Construction of the human promoter library is detailed in the **Supplementary Note**. The principle of CapStarr-seq was described previously[18]. The detailed step-by-step protocol is accessible on Protocol Exchange[32]. The human promoter library was transfected into K562 and HeLa cells using the Neon Transfection System (Thermo Fisher Scientific; pulse voltage 1,450 V and 1,005 V, pulse width 10 and 35 ms, pulse number 3 and 2 for K562 and HeLa cells, respectively). For each replicate, 30×10^6 cells were transfected with 150 μg of library; two independent transfection replicates were performed for each cell line. The transfected and non-transfected (plasmid input) libraries were single-end sequenced on the Illumina NextSeq 500 platform, and reads were mapped to the hg19 reference genome using standard procedures. **Supplementary Table 8** summarizes the number of sequenced and mapped reads for each sample. The coverage of each genomic region was calculated using BEDTools (v2.17.0), and the ratio of the CapStarr-seq coverage over the input (fold change) was computed for each sample. Promoter regions with enhancer activity were defined by determining the inflexion point of the ranked fold change (**Supplementary Table 2a**). Epromoters were defined as promoters displaying enhancer activity in both replicates. A common set of non-Epromoters was also defined as promoters lacking enhancer activity in all replicates of both cell lines. STARR-seq-positive controls displayed enhancer activity in our assays (**Supplementary Fig. 2a**).

Flow cytometry. We primarily observed enhanced GFP expression from the pooled promoter library as compared to the empty vector by FACS analysis (**Supplementary Fig. 2b**). A total of 5×10^6 K562 or HeLa cells were transfected with 25 μg of the empty STARR-seq screening vector[13] or the promoter library using the Neon Transfection System (Thermo Fisher Scientific) with the conditions described above. Twenty-four hours after electroporation, GFP expression was assessed on a FACSCalibur (BD Biosciences). Data were analyzed and visualized with FlowJo software.

RNA transcription and selection of the control set. Transcript quantification by RPKM (K562 and HeLa cell lines, four samples each) was obtained from the ENCODE Consortium (**Supplementary Table 9**). The data were normalized using the Normalizer package[33] with the quartiles $-\log_2$ option, and the mean of the four samples was obtained. A control (with the same expression) for each cell line was obtained by comparing Epromoters to promoters without enhancer activity (using transcription values for the nearest gene), and a list was generated of the same number of observations using a tool developed in house. The expression levels of genes associated with Epromoters and control sets in each cell line were compared to each other or to CapStarr-seq fold changes in signal and graphed using R software (R Core Team).

Epigenomic analysis. ChIP–seq data for the H3K4me3, H3K4me1 and H3K27ac histone marks, as well as DNase–seq data, were obtained from the ENCODE Consortium (**Supplementary Table 9**). Median average profiles were generated by extracting ChIP–seq signal from wiggle files for the 5-kb regions centered on TSSs. To test whether the differences between different classes of promoters were statistically significant, we first extracted the average signal for the top 25% of the signal in 2-kb regions centered on TSSs. A two-sided Mann–Whitney *U* test was then performed for each pair of promoter sets.

TSS analyses. To define promoter classes, clusters of 5′ GRO–seq transcripts from HeLa cells were obtained from Duttle *et al.*[20]. The clusters overlapping a 500-bp region extended from the promoter coordinates were retrieved. Bidirectional coding genes (TSS closer than 1.5 kb and in the opposite direction) were omitted. Each promoter was defined as a function of the orientation of the overlapping clusters of 5′ GRO-seq transcripts: unidirectional, only one transcript in the same direction as the gene; divergent, two RNA fragments in opposite directions; antisense, only one transcript in the opposite direction as the gene. Definition of TSS pairs as a function of RNA stability (UU, unstable–unstable; US, unstable–stable; SS, stable–stable) in K562 cells was obtained from Core *et al.*[3]. The TSS pairs overlapping a 500-bp region extended from the promoter coordinates were retrieved. Further analyses of TSSs and comparison with CAGE data are provided in the **Supplementary Note**.

Functional enrichment. GO enrichment in biological processes and pathways was assessed using g:Profiler[34] and default options (**Supplementary Table 3**). For the statistical background, we used the list of all genes associated with the capture promoters. Enrichment scores were calculated using the g:GOSt native method. Enrichment analysis for transcriptomic signatures was performed using GREAT[35] with all capture promoters as the background. Only gene signatures involved in TNF and interferon responses are shown in **Supplementary Figure 3c**.

To analyze the expression of type I interferon response genes, transcript quantification data (FPKM) for 23 cell lines (including HeLa and K562 cells) were obtained from the ENCODE Consortium (**Supplementary Table 9**) and normalized as described above. The FPKM values of genes involved in the 'Reactome: interferon αβ signaling' pathway were graphed using R software in a cumulative plot (**Supplementary Fig. 3a**). A Kolgomorov test was then performed to compare the HeLa and K562 cell lines. Genes in the 'Reactome: interferon αβ signaling' pathway that were differentially expressed in HeLa cells relative to the remaining 22 cell lines were identified by performing Significance Analysis of Microarrays (SAM) with TMEV (4.9)[36] software using a delta value of 0.5.

Transcription factor enrichment and density. ChIP–seq data (wiggle and peak files) from 71 (56 unique) and 218 (116 unique) transcription factors for the HeLa and K562 cell lines, respectively, were obtained from the ENCODE Consortium (**Supplementary Table 9**). To test whether the differences between Epromoters and control promoters (with the same expression) were statistically significant, we quantified the ChIP–seq signal from −200 to +50 bp with respect to the TSS. A Mann–Whitney *U* test was then performed for each pair of promoter sets. An enrichment score was calculated using the following

formula: $-\log_{10}$ (*P* value) if fold change >1 or $\log_{10}$ (*P* value) if fold change <1. A heat map of the scores was generated using Multiple Experiment Viewer[36]. We considered transcription factors to be enriched if they had a fold change >1.2 and $P < 0.001$. The average profiles for significantly enriched transcription factors were generated by extracting ChIP–seq signal from wiggle files for the 5-kb regions centered on TSSs. To assess the number of transcription factors bound per promoter (transcription factor density), the overlap of transcription factor peaks with Epromoters and control promoters (same expression) was assessed using BEDTools. The presence (1) or absence (0) of overlapping transcription factors for each promoter was summed and the density of transcription factors for each promoter was graphed using R software. A Kolgomorov test was then performed for each pair of promoter sets.

Motif analysis in Epromoters. Epromoter sequences from K562 and HeLa cells were scanned with a non-redundant collection of TFBMs (**Supplementary Note**) to detect over-represented and positionally biased motifs relative to control sequences (non-Epromoters). We detected motifs over-represented in Epromoters relative to non-Epromoters with the program matrix-enrichment (default parameters), which computes the cumulative distributions of scores for a given motif and computes the significance of over-representation at each possible score threshold with the binomial law. In addition to assessing global over-representation, we ran position-scan, which runs a chi-squared homogeneity test to detect motifs whose positional distribution differs between two sequence sets. We tuned the position-scan parameters to detect motifs showing a specific peak of enrichment near the core promoter (from −250 to +50 with respect to the TSS) of Epromoters relative to non-Epromoters. For graphical representation, the positional distributions of predicted sites were drawn on an extended region (±1 kb relative to the TSS), whereas the chi-squared test was restricted to the core promoter using a bin width of 50 bp and scanning with a threshold of $P \leq 1 \times 10^{-3}$. The background model was a first-order Markov chain trained with dinucleotide frequencies from all human core promoters.

Computations of ChIA-PET enrichment scores for promoter–promoter interactions. Pol II ChIA-PET interactions from HeLa and K562 cells were obtained from published data[37,38] and ENCODE Consortium data (**Supplementary Table 9**), respectively. ChIA-PET fragments for which the two mates intersected a 1-kb region encompassing two distinct TSSs were selected to define promoter–promoter interactions (**Supplementary Table 5**). Control sets were subsets of promoters without enhancer activity in both cell lines, as defined above. For each mark, each Epromoter was associated with a control promoter with the closest ChIP–seq signal computed from ENCODE Consortium data (**Supplementary Table 9**) to create a control list matched to the Epromoter list for signal distribution. To obtain enrichment scores, the fraction of promoters with promoter–promoter interactions was computed. Next, the number of interacting promoters labeled as Epromoter or control promoter was retrieved. ChIA-PET interactions mediated by H3K27ac, H3K4me2 and H3K4me1 were not significant for any set and are not displayed in **Figure 4c**. The corresponding enrichment scores were computed from hypergeometric tests using the following formula: $-\log_{10}$ (*P* value).

Gene expression correlation for interacting gene pairs. RNA–seq quantification data (FPKM) for 23 cell lines were retrieved from the ENCODE Consortium (**Supplementary Table 9**) and normalized as described above. Pearson's correlation between coding-gene pairs on the same chromosome and having a ChIA-PET interaction in K562 or HeLa cells (**Supplementary Table 5**) was assessed using R software (R Core Team). Correlation scores for gene pairs involving at least one Epromoter or only non-Epromoters were graphed using R software. A control set containing shuffled gene pairs from the ChIA-PET interacting pairs was also plotted.

CRISPR–Cas9 genome editing. Targeted Epromoter and promoter regions were defined by CapStarr-seq and DNase–seq peaks ranging from 410 bp to 1,255 bp in length (**Supplementary Fig. 7b–h**, left). For the knockout experiments, the general strategy is shown in **Supplementary Figure 7a**. Two gRNAs were designed for each end of the targeted region using the CRISPRdirect tool[39]. The gRNAs were cloned into a gRNA cloning vector (Addgene, 41824) as previously described[40]. Two million cells were transfected with 15 µg of the hCas9 vector (Addgene, 41815) and 7 µg of each gRNA using the Neon Transfection System (Thermo Fisher Scientific). Three days after transfection, the bulk of transfected cells were plated in 96-well plates at limiting dilution (0.5 cells per 100 µl per well) for clonal expansion. After 10–14 d, individual cell clones were screened for homologous allele deletion by direct PCR using Phire Tissue Direct PCR Master Mix (Thermo Fisher Scientific) according to the manufacture's protocol. Forward and reverse primers were designed bracketing the targeted regions, allowing for the detection of knockout or wild-type alleles. Clones were considered to have undergone homologous allele deletion if they had at least one deletion band of the expected size and no wild-type band (**Supplementary Fig. 7b–h**, right). If more than two cell clones were obtained for a given locus, the most precise deletion was chosen. All gRNAs and primers are listed in **Supplementary Table 10**. The generation of clones in which the *FAF2* Epormoter was inverted and eQTL SNPs were mutated is described in the **Supplementary Note**.

Gene expression. Total RNA was extracted using TRIzol reagent (Thermo Fisher Scientific). 3 µg of RNA was then treated with DNase I (Ambion) and reverse transcribed into cDNA using Superscript VILO Master Mix (Thermo Fisher Scientific). Real-time PCR was performed using Power SYBR Master Mix (Thermo Fisher Scientific) on a Stratagene Mx3000P instrument. Primer sequences are listed in **Supplementary Table 10**. Gene expression was normalized to that of *GAPDH*. Relative expression was calculated by the ΔC_T method, and all data shown are reported as the fold change over the control. For each cell clone, the Student's *t* test was performed (unpaired, two-tailed, 95% confidence interval) from three independent RNA/cDNA preparations. Data are represented with s.d. For conventional RT–PCR, one-twentieth of the synthesized cDNA was used as the template for one reaction; PCRs were performed with Phusion polymerase (Thermo Fisher Scientific), $T_m = 60$ °C, 30 cycles.

***FAF2* rescue experiments.** Human *FAF2* cDNA was purchased from Origene (SC100662). K562 cell clones in which the *FAF2* Epromoter was knocked out or inverted were transfected with 2 µg of *FAF2* cDNA plasmid, and samples were collected 24 h after transfection for gene expression analysis as described above.

Allelic expression. Genetic variants within the transcribed regions of the *PIGM* (chr1:160,000,435) and *UBE2L6* (chr11:57,319,339) genes were identified by visual assessment of RNA–seq data from the K562 cell line using the IGV tool (version 2.3.67)[41]. PCR primers containing Illumina adaptors were designed flanking each variant (**Supplementary Table 10**). cDNAs from wild-type K562 clones and clones with homozygous and heterozygous deletion of the *TAGLN2* and *YPEL4* Epromoters were amplified using *PIGM*- and *UBE2L6*-specific primers, respectively. In the case of *UBE2L6*, the cDNA was generated from IFN-α-treated cells. A second PCR was performed using NEBNext Multiplex Oligos for Illumina (New England BioLabs), the product was subjected to single-end sequencing on the Illumina NextSeq 500 platform and reads were mapped to the hg19 reference genome using standard procedures. Allelic frequency was computed using the IGV tool.

Haplotype-resolved analysis of DNase–seq and ChIP–seq data. Transcription factors for which a ChIP–seq peak in K562 cells (ENCODE Consortium) overlapped the eQTL SNP rs6681671 in the *CSDE1* Epromoter were selected. BAM files from corresponding ChIP–seq data, along with DNase–seq data and input, were directly retrieved with the IGV tool, and the frequency of the haplotype-resolved reads was manually computed. Only samples with at least ten reads were selected.

Chromatin immunoprecipitation and qPCR. Generation of ChIP samples is described in the **Supplementary Note**. ChIP eluates and input were assayed by real-time PCR (Stratagene Mx3000P instrument) in a 20-µl reaction with one-thirtieth of the elution material using Power SYBR Master Mix (Thermo Fisher Scientific). The primers used in the real-time PCR assays are listed in **Supplementary Table 10**. Data represent the percentage of input normalized to *ACTB* with s.d. Student's *t* test (two-tailed, unpaired) was used to test for significance from three independent chromatin preparations.

4C analysis. 4C–seq experiments were carried out as described[42–44]. 4C libraries were prepared using NlaIII–DpnII enzyme combinations for the *FAF2* and *RNF44* promoters. Primer sequences are listed in **Supplementary Table 10**. For the *FAF2* viewpoint, two technical replicates each of one wild-type K562 clone and two ΔEp.*FAF2* clones were analyzed. For the *RNF44* viewpoint, one wild-type K562 clone, two ΔEp.*FAF2* clones and one Inv.Ep.*FAF2* clone were analyzed. Samples were sequenced and used for downstream analysis as independent replicates and as a merged data set. 4C–seq data processing was performed as described[45] using the NCBI human assembly GRCh37 (hg19), and detailed analysis and visualization were carried out using r3Cseq and FourCseq software[46,47]. For a visible data profile, normalized RPM data were smoothed via a running-mean approach and quantiles (40%, 50% and 60%) were further smoothed and interpolated with the R loess function using Basic4Cseq[48].

Distal association with interferon response. Human type I interferon response genes were retrieved from Interferome database v2.01 (ref. 49). We then selected the interferon response genes distally interacting with an Epromoter on the basis of ChIA-PET data (**Supplementary Table 5**). The list of Epromoters distally interacting with interferon response genes is provided in **Supplementary Table 6**.

eQTL analysis. eQTL data were obtained from GTEx project portal version 6 and lifted over to hg19 coordinates to match capture promoter data. Using GenomicRanges[50], capture promoter coordinates were extended 1.5 kb to each side to capture overlapping eQTLs that could be mechanistically related to these promoters. ChIA-PET promoter–promoter pairs were obtained as described below. Promoter–promoter pairs were annotated using capture promoters and eQTL overlaps to determine long- and close-range interaction effects between pairs. We were able to annotate 4,310 of 7,825 pairs (**Supplementary Table 7**). Customized R scripts were used to analyze the relationship between eQTL β value (effect size) and long- and close-range gene promoter interactions in the annotated promoter–promoter pairs and to determine whether eQTLs were located within the extended region of an Epromoter or a non-Epromoter. Taking only eQTLs affecting the distal gene in the pair, the β-value bimodal distributions of these eQTLs were split into negative and positive values by fitting a two-component mixture model (R mixtool package[51]) and looking for the cutoff where the probability of a negative value being generated by the left distribution was ≥0.5. To test whether Epromoter-associated β values were stronger than the ones associated with non-Epromoters, we independently compared negative and positive β-value sets using a one-tailed non-parametric Wilcoxon rank-sum test (wilcox.test R function) and corrected for multiple testing using the Benjamini–Hochberg method (p.adjust R function). The statistical analyses to predict the impact of eQTL SNPs on transcription factor binding sites is detailed in the **Supplementary Note.**

Statistics. All experiments were performed using at least three independent samples or transfections. R/Bioconductor or GraphPad Prism 6.0 was used for statistical analysis. For comparisons in Venn diagram representations, a hypergeometric test was performed. Unless otherwise indicated in the figure legends, for comparisons between two groups of equal sample size and small n (like in qPCR dot plots), an unpaired two-tailed Student's t test was performed; for comparisons between two groups of equal sample size and large n (as in box-plot representations), a two-tailed Mann–Whitney U test was performed. For comparisons of two distributions, a Kolmogorov–Smirnov test was performed. $P < 0.05$ was considered to be statistically significant, and error bars represent s.d. Investigators were not blinded to sample identity.

Data availability. All custom scripts have been made available at https://github.com/arielgalindoalbarran/Epromoters. Human CapStarr-seq and 4C data generated during the current study are available in the Gene Expression Omnibus (GEO) under accessions GSE83296 (**Supplementary Table 8**) and GSE98194, respectively. Mouse CapStarr-seq data analyzed during the current study were published previously[18] and are available in GEO under accession GSE60029. All public data sets and primers used are described in **Supplementary Tables 9** and **10**, respectively.

31. 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. USA* **92**, 7420–7424 (1995).
32. Dao, L.T.M., Vanhille, L., Griffon, A., Fernandez, N. & Spicuglia, S. CapStarr-seq protocol. *Protocol Exchange* http://dx.doi.org/10.1038/protex.2015.096 (2015).
33. Glusman, G., Caballero, J., Robinson, M., Kutlu, B. & Hood, L. Optimal scaling of digital transcriptomes. *PLoS One* **8**, e77885 (2013).
34. Reimand, J. *et al.* g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44**, W1 W83-9 (2016).
35. McLean, C.Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
36. Saeed, A.I. *et al.* TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374–378 (2003).
37. Kuznetsova, T. *et al.* Glucocorticoid receptor and nuclear factor kappa-b affect three-dimensional chromatin organization. *Genome Biol.* **16**, 264 (2015).
38. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163**, 1611–1627 (2015).
39. 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).
40. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
41. Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
42. Stadhouders, R. *et al.* Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nat. Protoc.* **8**, 509–524 (2013).
43. Stadhouders, R. *et al.* HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. *J. Clin. Invest.* **124**, 1699–1710 (2014).
44. Vieux-Rochas, M., Fabre, P.J., Leleu, M., Duboule, D. & Noordermeer, D. Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain. *Proc. Natl. Acad. Sci. USA* **112**, 4672–4677 (2015).
45. Stadhouders, R. *et al.* Control of developmentally primed erythroid genes by combinatorial co-repressor actions. *Nat. Commun.* **6**, 8893 (2015).
46. Thongjuea, S., Stadhouders, R., Grosveld, F.G., Soler, E. & Lenhard, B. r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. *Nucleic Acids Res.* **41**, e132 (2013).
47. Klein, F.A. *et al.* FourCSeq: analysis of 4C sequencing data. *Bioinformatics* **31**, 3085–3091 (2015).
48. Walter, C., Schuetzmann, D., Rosenbauer, F. & Dugas, M. Basic4Cseq: an R/Bioconductor package for analyzing 4C-seq data. *Bioinformatics* **30**, 3268–3269 (2014).
49. Rusinova, I. *et al.* Interferome v2.0: an updated database of annotated interferon-regulated genes. *Nucleic Acids Res.* **41**, D1040–D1046 (2013).
50. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).
51. Benaglia, T., Chauveau, D., Hunter, D.R. & Young, D.S. mixtools: An R Package for Analyzing Finite Mixture Models. *J. Stat. Softw.* **32**, 1–29 (2009).

Title: H3K4 methylation at enhancers and broad promoters

Pękowska Aleksandra[1], Charbonnier Guillaume[2-3], Andrau Jean-Christophe[4], Spicuglia Salvatore[2-3,*]

[1]Lymphocyte Nuclear Biology, NIAMS, NIH, Bethesda, MD 20892, USA
[2]Aix-Marseille University, Inserm, TAGC, UMR1090, Marseille, France.
[3]Equipe Labélisée Ligue Contre le Cancer, Marseille, France.
[4] Institut de Génétique Moléculaire de Montpellier, University of Montpellier, CNRS, Montpellier, France
* Lead Contact

1513 words

In issue 72 of Molecular Cell, Shah *et al*. determined the specificity of widely used antibodies recognizing different levels of methylation of histone 3 at lysine 4 (H3K4me) (Shah *et al*., 2018). The authors presented a series of calibrated ChIP-seq experiments and firmly concluded that a significant part of the published studies that report on H3K4me has been “flawed” due to cross-reactivity of antibodies. Considering the potential weight of these conclusions, we investigated the source of the discrepancy between the previous papers and that of Shah *et al*.

Using ChIP-seq data from primary tissues, we and others uncovered that tissue-specific and cell-identity genes display a unique pattern of histone modifications featuring broad enrichment of H3K4me2/me3 that extends from the transcription start site (TSS) into the gene-body (Benayoun *et al*., 2014; Chen *et al*., 2015; Pekowska *et al*., 2010; Suzuki *et al*., 2017; Zacarias-Cabeza *et al*., 2015). Contrasting with these results, Shah *et al*. observed that, in K562, such loci were related to metabolic processes. The authors concluded that cross-reactivity between antibodies used in the past had led to the erroneous association of tissue-specific genes with gene-body enrichment of H3K4 methylation.

To address this issue, we reanalyzed published H3K4me2 data to identify genes with high level of H3K4me2 in the gene-body, as previously described (Pekowska *et al*., 2010) (cluster 1; **Figure S1A**). Independently of the antibody used to detect H3K4me2 (Abcam or Thermofisher) or the normalization approach (uncalibrated *versus* calibrated ChIP-seq), the loci marked with gene-body enrichment of H3K4me2 in K562 cells were primarily associated with functional terms related to metabolic processes (**Figure S1B**), consistent with Shah *et al*.. Remarkably, the same analysis in primary B cells (i.e. Abcam; ENCODE) revealed a specific enrichment for genes involved in lymphoid and hematopoietic functions (**Figure S1B**). Next, we identified genes associated with H3K4me3 broad domains in a subset of primary tissues and cancer cell lines. In agreement with the published findings, genes displaying broad H3K4me3 signal around the TSS region were primarily implicated in cell-type specific functions reflecting the tissue-of-origin of the samples (**Figure S1C**). For instance, loci associated with broad H3K4me3 domains in CD3+ T cells and T-acute lymphoblastic leukemia cell lines (DND41 and Jurkat) were primarily enriched in gene ontologies related to T cell functions, while those identified in the brain and neuroblastoma cell lines (Be2C and Sknmc) were involved in neuronal functions. This analysis also revealed that a small subset of cancer cell lines, including K562, did not follow this trend (**Figure S1C, highlighted**). In these outlier samples, genes with a broad H3K4me3 pattern were primarily related to metabolic processes. Thus, neither the antibody nor the ChIP calibration procedure can comprehensibly account for the differences between the published results and the analyses presented by Shah *et al*. Rather, the discrepancy can be attributed to a yet undetermined aspect of chromatin landscape of few cancer cell lines, including K562.

A high ratio of H3K4me1/H3K4me3 accompanied by H3K27ac have been used as a proxy for active enhancers (Creyghton *et al*., 2010; Heintzman *et al*., 2009; Rada-Iglesias *et al*., 2011). However, the concept of a strict dichotomy between H3K4me1 and H3K4me3 enrichments and enhancer function was challenged by the following observations. First, the presence of H3K4me3 is fully compatible with enhancer activity (Core *et al*., 2014; Ernst *et al*., 2011; Henriques *et al*., 2018; Koch *et al*., 2011; Mikhaylichenko *et al*., 2018; Outchkourov *et al*., 2013; Pekowska *et al*., 2011; Rennie *et al*., 2018; Wang *et al*., 2008). Second, the level of H3K4me3 is positively correlated with enhancer strength and RNA level (Andersson *et al*., 2014; Core *et al*., 2014;

Henriques *et al*., 2018; Ibrahim *et al*., 2018; Vanhille *et al*., 2015). Moreover, the gain of H3K4me3 at enhancers is linked to their oncogenic activation in cancer (Li et al., 2019; Shen et al., 2016). It is now widely recognized that, akin promoters, active enhancers bind RNA-Polymerase II (Pol2) and produce enhancer RNAs (eRNAs) (Andersson *et al*., 2014; Core *et al*., 2014; De Santa *et al*., 2010; Henriques *et al*., 2018; Kim *et al*., 2010; Koch *et al*., 2011). Thus, it has been postulated that the differences in the combinatorial pattern of H3K4me simply reflect different transcription levels between the two types of elements, leading to an unified model of distal and proximal regulatory elements (Andersson et al., 2015; Kim and Shiekhattar, 2015; Medina-Rivera et al., 2018).

Shah *et al*., stated that there is little evidence for H3K4me3 enrichment at putative enhancers. They observed that, when averaged across the entire set of the putative regulatory elements, the level of H3K4me3 is substantially lower at enhancers than at promoters (Figure 4C and 4H of Shah *et al*.). This is actually in line with published reports in that only a fraction of enhancers is associated with H3K4me3. However, supported by no direct evidence, the authors concluded that the apparent enrichment of H3K4me3 at enhancers arises as a consequence of “signal leakage” from lower methyl forms due to poor quality of the antibodies. Yet, the antibody that displayed the highest specificity towards H3K4me3 also yielded the most pronounced signal at distal regulatory elements (Figure 4H of Shah *et al*.), suggesting a genuine enrichment of this modification at enhancers. It is also important to stress that, unlike declared in Shah *et al.*, we did not state that, in primary T cells, H3K4me1 and H3K4me3 are similarly enriched at promoters and intergenic TBP sites. Rather, we concluded that both marks were present at these elements without putting any emphasis on the quantitative comparison of the levels of H3K4me3 at promoters and enhancers (Koch et al, 2011). We also note that a rigorous and thus meaningful comparison between Koch *et al.*, in primary T cells and Figure 7A in K562 from Shah et al. is virtually impossible. Their analysis is not based on the same selection criteria for TBP sites and did not consider Pol2-Ser5P or histone modification enrichments. It is in fact critical to take these variables into account, as TBP is known to bind at many Pol2-independent regions.

To determine the extent of the overlap between distal DNaseI hypersensitive sites (DHS) and H3K4me3, we considered the list of DHS in K562 (ENCODE) as well as the high confidence H3K4me3 tracks provided by Shah *et al.* using their most specific antibody. We found that 537 of 9,155 distal DHS overlapped high confidence H3K4me3 peaks. This included elements experimentally validated as active enhancers in K562 (Gasperini et al., 2019) (**Figure S1D**). Consistent with this result, a recent publication has independently validated the presence of H3K4me3 at enhancers, also using Shah *et al*. data (Li et al., 2019).

Previous work from us and others have shown that active enhancers feature relatively high levels of H3K4me3. Thus, we next assessed whether the H3K4me3-enriched distal DHS represented *bona fide* active enhancers in K562. Given the established association between H3K4me3, Pol2 binding and eRNA expression at active enhancers, we stratified distal DHS by the level of eRNA, as determined by GRO-seq, or Pol2 binding. We found that enhancers producing eRNAs or bound by Pol2 were 8 and 10 times more likely to overlap H3K4me3 peaks than elements without detectable transcription or Pol2 peaks, respectively (**Figures S1E and G;** P-val.<$2.2x10^{-16}$, Fisher's exact test). Moreover, increased eRNA levels (**Figure S1F**) or Pol2 binding (**Figure S1H**) were significantly associated with H3K4me3 enrichment. Thus, although H3K4me3 was found in a relatively small subset of distal elements, its presence constitutes a genuine marker of their *cis*-regulatory activity.

Shah *et al*. also failed to detect a significant link between the H3K4me3/H3K4me1 ratio at enhancers and chromatin contacts between promoters and enhancers and hence concluded that the level of H3K4me3 does not report on enhancer activity. Such approach requires two primary assumptions to hold. First, that the H3K4me3/H3K4me1 ratio is related to enhancer activity. Second, that enhancer activity can be extrapolated by the number of ChIA-PET contacts it establishes with promoters. None of these hypotheses have been explicitly tested in the literature. Thus, the lack of a correlation between these two features does not appear sufficient to reject the potential association between the presence of H3K4me3 and enhancer activity. Finally, to explain why H3K4me3 was observed by others at enhancer elements, Shah *et al.*, argued that the use of crosslinking procedure might have resulted in a reduced specificity of antibodies. Yet, this explanation seems unlikely

since in Pekowska *et al*. , the bulk of the experiments was performed under native conditions (Pekowska *et al*., 2011). It is worth highlighting here that in the previous studies (e.g. (Pekowska et al., 2011) the association of H3K4me3 with active enhancers was assessed using primary models of cell differentiation and not cancer cell lines, which do not constitute optimal models to define the nature of epigenetic dynamics during development.

In sum, we show that the failure to reach conclusions put forward by other groups is not, as the authors claimed, caused by a poor performance of the antibodies used but rather a consequence of the experimental design in Shah *et al*., embodied by the use of a single cancer cell line (K562) as well as unappropriated analyses.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and one figure

AUTHOR CONTRIBUTIONS

P.A., C.G. and S.S. analyzed data. P.A., A.J.C. and S.S. interpreted results and wrote the letter.

References

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T.*, et al.* (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455-461.

Andersson, R., Sandelin, A., and Danko, C.G. (2015). A unified architecture of transcriptional regulatory elements. Trends in genetics : TIG *31*, 426-433.

Benayoun, B.A., Pollina, E.A., Ucar, D., Mahmoudi, S., Karra, K., Wong, E.D., Devarajan, K., Daugherty, A.C., Kundaje, A.B., Mancini, E., Hitz, B.C., Gupta, R., Rando, T.A., Baker, J.C., Snyder, M.P., Cherry, J.M., and Brunet, A. (2014). H3K4me3 breadth is linked to cell identity and transcriptional consistency. Cell *158*, 673-688.

Chen, K., Chen, Z., Wu, D., Zhang, L., Lin, X., Su, J., Rodriguez, B., Xi, Y., Xia, Z., Chen, X., Shi, X., Wang, Q., and Li, W. (2015). Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. Nature genetics *47*, 1149-1157.

Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature genetics *46*, 1311-1320.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., Boyer, L.A., Young, R.A., and Jaenisch, R. (2010).

Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci USA, DOI: 10.1073.
De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.L., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS biology *8*, e1000384.
Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B.E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43-49.
Gasperini, M., Hill, A.J., McFaline-Figueroa, J.L., Martin, B., Kim, S., Zhang, M.D., Jackson, D., Leith, A., Schreiber, J., Noble, W.S., Trapnell, C., Ahituv, N., and Shendure, J. (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell *176*, 377-390 e319.
Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., Ching, K.A., ntosiewicz-Bourget, J.E., Liu, H., Zhang, X., Green, R.D., Lobanenkov, V.V., Stewart, R., Thomson, J.A., Crawford, G.E., Kellis, M., and Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108-112.
Henriques, T., Scruggs, B.S., Inouye, M.O., Muse, G.W., Williams, L.H., Burkholder, A.B., Lavender, C.A., Fargo, D.C., and Adelman, K. (2018). Widespread transcriptional pausing and elongation control at enhancers. Genes Dev *32*, 26-41.
Ibrahim, M.M., Karabacak, A., Glahs, A., Kolundzic, E., Hirsekorn, A., Carda, A., Tursun, B., Zinzen, R.P., Lacadie, S.A., and Ohler, U. (2018). Determinants of promoter and enhancer transcription directionality in metazoans. Nat Commun *9*, 4472.
Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P.F., Kreiman, G., and Greenberg, M.E. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182-187.
Kim, T.K., and Shiekhattar, R. (2015). Architectural and Functional Commonalities between Enhancers and Promoters. Cell *162*, 948-959.
Koch, F., Fenouil, R., Gut, M., Cauchy, P., Albert, T.K., Zacarias-Cabeza, J., Spicuglia, S., de la Chapelle, A.L., Heidemann, M., Hintermair, C., Eick, D., Gut, I., Ferrier, P., and Andrau, J.C. (2011). Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. Nature structural & molecular biology *18*, 956-963.
Li, Q.L., Wang, D.Y., Ju, L.G., Yao, J., Gao, C., Lei, P.J., Li, L.Y., Zhao, X.L., and Wu, M. (2019). The hyper-activation of transcriptional enhancers in breast cancer. Clin Epigenetics *11*, 48.
Medina-Rivera, A., Santiago-Algarra, D., Puthier, D., and Spicuglia, S. (2018). Widespread Enhancer Activity from Core Promoters. Trends Biochem Sci.
Mikhaylichenko, O., Bondarenko, V., Harnett, D., Schor, I.E., Males, M., Viales, R.R., and Furlong, E.E.M. (2018). The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. Genes Dev *32*, 42-57.
Outchkourov, N.S., Muino, J.M., Kaufmann, K., van Ijcken, W.F., Groot Koerkamp, M.J., van Leenen, D., de Graaf, P., Holstege, F.C., Grosveld, F.G., and Timmers, H.T. (2013). Balancing of histone H3K4 methylation states by the Kdm5c/SMCX histone demethylase modulates promoter and enhancer function. Cell reports *3*, 1071-1079.
Pekowska, A., Benoukraf, T., Ferrier, P., and Spicuglia, S. (2010). A unique H3K4me2 profile marks tissue-specific gene regulation. Genome research *20*, 1493-1502.

Pekowska, A., Benoukraf, T., Zacarias-Cabeza, J., Belhocine, M., Koch, F., Holota, H., Imbert, J., Andrau, J.C., Ferrier, P., and Spicuglia, S. (2011). H3K4 tri-methylation provides an epigenetic signature of active enhancers. The EMBO journal *30*, 4198–4210.
Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279-283.
Rennie, S., Dalby, M., Lloret-Llinares, M., Bakoulis, S., Dalager Vaagenso, C., Heick Jensen, T., and Andersson, R. (2018). Transcription start site analysis reveals widespread divergent transcription in D. melanogaster and core promoter-encoded enhancer activities. Nucleic acids research.
Shah, R.N., Grzybowski, A.T., Cornett, E.M., Johnstone, A.L., Dickson, B.M., Boone, B.A., Cheek, M.A., Cowles, M.W., Maryanski, D., Meiners, M.J., Tiedemann, R.L., Vaughan, R.M., Arora, N., Sun, Z.W., Rothbart, S.B., Keogh, M.C., and Ruthenburg, A.J. (2018). Examining the Roles of H3K4 Methylation States with Systematically Characterized Antibodies. Molecular cell *72*, 162-177 e167.
Shen, H., Xu, W., Guo, R., Rong, B., Gu, L., Wang, Z., He, C., Zheng, L., Hu, X., Hu, Z., Shao, Z.M., Yang, P., Wu, F., Shi, Y.G., Shi, Y., and Lan, F. (2016). Suppression of Enhancer Overactivation by a RACK7-Histone Demethylase Complex. Cell *165*, 331-342.
Suzuki, H.I., Young, R.A., and Sharp, P.A. (2017). Super-Enhancer-Mediated RNA Processing Revealed by Integrative MicroRNA Network Analysis. Cell *168*, 1000-1014 e1015.
Vanhille, L., Griffon, A., Maqbool, M.A., Zacarias-Cabeza, J., Dao, L.T.M., 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.
Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q., and Zhao, K. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. Nature genetics *40*, 897-903.
Zacarias-Cabeza, J., Belhocine, M., Vanhille, L., Cauchy, P., Koch, F., Pekowska, A., Fenouil, R., Bergon, A., Gut, M., Gut, I., Eick, D., Imbert, J., Ferrier, P., Andrau, J.C., and Spicuglia, S. (2015). Transcription-Dependent Generation of a Specialized Chromatin Structure at the TCRbeta Locus. J Immunol *194*, 3432-3443.

Supplemental Experimental Procedures

Functional enrichment of genes associated with H3K4me3 broad domains

H3K4me3 peaks associated with gene promoters in a subset of samples from ENCODE and Roadmap Epigenomics were retrieved from *Chen et al.* (Chen et al., 2015). Genes associated with H3K4me3 peaks broader than 4 kb were extracted for each sample and used for GO Biological Process enrichment analysis using clusterProfiler (Yu et al., 2012). For each sample we retrieved the most significantly enriched term related to the biological processes. Source code for the analysis is available here (https://github.com/guillaumecharbonnier/mw-broad).

Gene annotation

RefSeq annotation of promoters from the GRCh38 assembly of the human genome was retrieved from ensemble using *getBM* function from the *biomaRt* package.

Identification of enhancers

Genomic coordinates of DNase I hypersensitive sites identified in two independent samples of K562 cells were obtained from ENCODE. The list of DHS sites was further limited to instances detected in both replicates, the discrimination was made based on the genomic overlap between the two lists of DHS peaks. DHS peaks that overlapped H3K4me1 peaks (Shah et al., 2018) and that did not overlap promoters (+/- 2kb around the annotated transcription starts) or genes (for this analysis gene coordinates were extended by 2 kb in both directions) were considered as enhancers throughout the study.

Estimation of the transcriptional activity of enhancers

Global run on coupled with high throughput sequencing (GRO-seq) data was used to estimate transcriptional activity at enhancer elements. Process GRO-seq tracks were downloaded from GEO and the signal at enhancers integrated (GRO-seq score). Classification of enhancers into expression classes was based on the distribution of the GRO-seq scores and we identified elements with no GRO-seq reads, elements with low expression (GROs-eq score <2 but > than 0) and elements with high level of GRO-seq signal (>2).

Analysis of H3K4me3 signal at different classes of enhancers

We considered the HMD track from Shah et al., obtained using the most specific H3K4me3 antibody. We integrated the signal at enhancers that overlap Pol2 peaks or enhancers classified based on the GRO-seq signal.

Clustering of H3K4me2 profiles

We considered the protein coding genes. Extended promoter regions were defined as regions -2+8kb around the annotated transcription start sites. We collapsed loci based on the gene symbol annotation. In cases whereby multiple TSS were annotated and the breadth of the TSS region expanded 2kb., we considered the TSS that was located at the shortest distance with respect to the transcription termination site. We kept only instances longer then 8kb that did not overlap any other loci identified at this point. The 10kb regions were next divided into 101 bins. We considered the processed published "bigwig tracks", for each of the 101 bins defined above, we retrieved all the bins from the ChIP-seq tracks that intersected the bin and integrated the ChIP-seq signal.

We noticed that in the K562 data, irrespectively of the antibody used in the ChIP procedure, a handful of genes displayed unusually high level of H3K4me2 modification. To account for that we assessed the distribution of the summed H3H4me2 signal in all the considered data and removed genes found within the top 1 percentile for any experiment.

We clustered the H3K4me2 patterns using R function *kmeans,* the expected number of clusters was set to five (Pekowska et al., 2010). The clusters were manually named according to their gene body enrichment on H3K4me3 (gene body, TSS or Negative). We performed the GO term analysis on the gene body enriched cluster, using the *runTest* function from the R/Bionconducotr *topGO* package, the nodeSize was set to 10.

Summary of the data considered in our study

Data	Dataset	File name
H3K4me1 peaks	GSM2773406	GSM2773406
H3K4me3 peaks	GSM2773406	GSM2773406_AR16-4_TF-PA5-40086_High_Confidence_Peaks.bed
H3K4me3 track	GSM2773406	GSM2773406_AR16-4_TF-PA5-40086_mQ20_L200_genome_coverage.bedGraph
PolII peaks	ENCODE	ENCFF516BJP_optimal_idr_thresholded_peaks_reps_1_and_2.bed
DHS replicate 1	ENCODE	ENCFF100IJK_DHS_Stam_2018_rep1.bed
DHS replicate 2	ENCODE	ENCFF339PKG_DHS_Stam_2018_rep2.bed
GRO-seq	ENCODE	GSM1480325_K562_GROseq_minus.bigWig

		GSM1480325_K562_GROseq_plus.bigWig
H3K4me2 track	GSM2773406	GSM2773400_AR16-5_TF-710796_mQ20_L200_genome_coverage.bigwig
H3K4me2 HMD	GSM2773406	GSE103543_AR16_H3K4me2-Corrected_HMD.bedgraph
H3K4me2 B cells	ENCODE	ENCFF945VIU_B_cell_rep1_rep2.bigWig
Broad H3K4me3 peaks	(Chen et al., 2015)	(https://media.nature.com/original/nature-assets/ng/journal/v47/n10/extref/ng.3385-S4.xls)

References

Chen, K., Chen, Z., Wu, D., Zhang, L., Lin, X., Su, J., Rodriguez, B., Xi, Y., Xia, Z., Chen, X., Shi, X., Wang, Q., and Li, W. (2015). Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. Nature genetics *47*, 1149-1157.

Pekowska, A., Benoukraf, T., Ferrier, P., and Spicuglia, S. (2010). A unique H3K4me2 profile marks tissue-specific gene regulation. Genome research *20*, 1493-1502.

Shah, R.N., Grzybowski, A.T., Cornett, E.M., Johnstone, A.L., Dickson, B.M., Boone, B.A., Cheek, M.A., Cowles, M.W., Maryanski, D., Meiners, M.J., Tiedemann, R.L., Vaughan, R.M., Arora, N., Sun, Z.W., Rothbart, S.B., Keogh, M.C., and Ruthenburg, A.J. (2018). Examining the Roles of H3K4 Methylation States with Systematically Characterized Antibodies. Molecular cell *72*, 162-177 e167.

Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284-287.

Figure S1. H3K4 methylation at broad promoters and enhancers

(A) Average profiles for H3K4me2 around gene TSSs using ENCODE data from K562 cell line and human B-cells. Five clusters were generated using k-means clustering of H3K4me2 signal distribution as previously described (Pekowska et al., 2010). The cluster with gene body enrichment of H3K4me2, centered on the TSS, or with low signal were denoted “Gene body”, “TSS” and “Negative”, respectively.

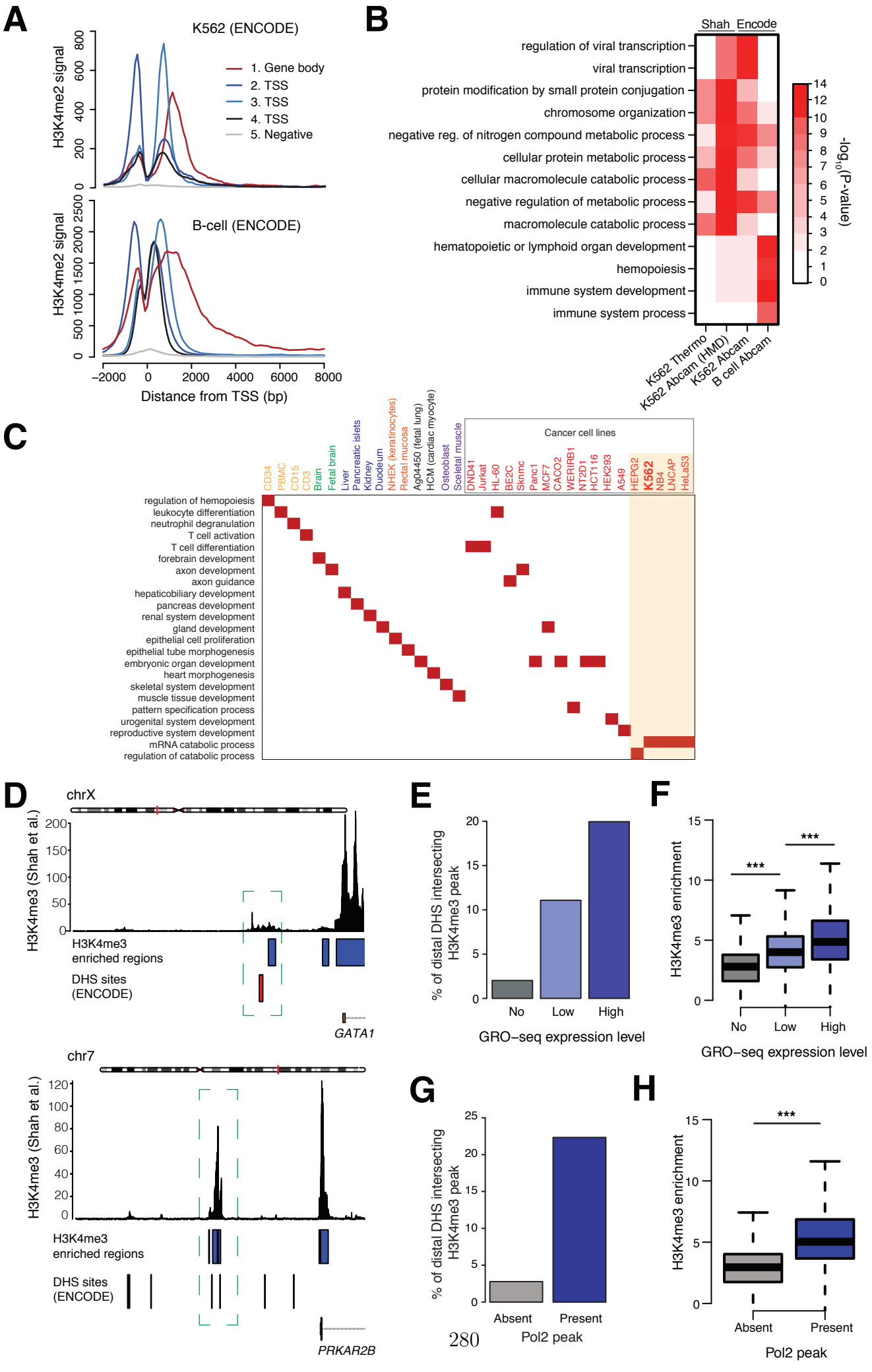
(B) Heatmap showing the GO term enrichments for Biological Processes found in the “Gene body” H3K4me2 clusters identified in the datasets from Shah et al. or ENCODE and using the indicated antibodies. The top 3 most significant terms for each condition was selected.

(C) Top GO term enrichments for Biological Processes identified from the list of H3K4me3 broad associated genes from a subset of ENCODE and ROADMAP using pre-processed data from (Chen et al. 2018). The most significant term for each sample is indicated. The cancer cell lines enriched in metabolic terms is highlighted in yellow.

(D) Screenshots of *GATA1* and *PRKAR2B* loci along with their cognate enhancers, recently validated through CRISPRi screens in K562 cells (Gasperini et al. 2019). At both loci, enhancers are marked by a significant enrichment of H3K4me3 (PA5-40086, Shah et al. 2018).

(E-G) Percentage of distal DHS overlapping H3K4me3 peaks (HMD data from Shah et al. 2018) with respect to different levels of GRO-seq signal (E) or the presence of Pol2 (G).

(F-H) Boxplots showing the H3K4me3 enrichment (HMD data from Shah et al. 2018) with respect to different levels of GRO-seq signal (F) or the presence of Pol2 (H).


A
K562 (ENCODE)
H3K4me2 signal
1. Gene body
2. TSS
3. TSS
4. TSS
5. Negative
B-cell (ENCODE)
Distance from TSS (bp)
B
Shah
Encode
regulation of viral transcription
viral transcription
protein modification by small protein conjugation
chromosome organization
negative reg. of nitrogen compound metabolic process
cellular protein metabolic process
cellular macromolecule catabolic process
negative regulation of metabolic process
macromolecule catabolic process
hematopoietic or lymphoid organ development
hemopoiesis
immune system development
immune system process
-log10(P-value)
K562 Thermo
K562 Abcam (HMD)
K562 Abcam
B cell Abcam
C
Cancer cell lines
regulation of hemopoiesis
leukocyte differentiation
neutrophil degranulation
T cell activation
T cell differentiation
forebrain development
axon development
axon guidance
hepaticobiliary development
pancreas development
renal system development
gland development
epithelial cell proliferation
epithelial tube morphogenesis
embryonic organ development
heart morphogenesis
skeletal system development
muscle tissue development
pattern specification process
urogenital system development
reproductive system development
mRNA catabolic process
regulation of catabolic process
D
chrX
H3K4me3 (Shah et al.)
H3K4me3 enriched regions
DHS sites (ENCODE)
GATA1
chr7
PRKAR2B
E
% of distal DHS intersecting H3K4me3 peak
No
Low
High
GRO−seq expression level
F
H3K4me3 enrichment
***
GRO−seq expression level
G
Absent
Present
Pol2 peak
H
H3K4me3 enrichment
***
Pol2 peak