# Etude bioinformatique de l'épigénome au cours de la différenciation des lymphocytes T et des leucémies

Mohamed Belhocine

# UNIVERSITE D'AIX-MARSEILLE

ECOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE
Technologies Avancées pour le Génome et la Clinique
INSERM U1090

## THESE DE DOCTORAT

MENTION GENOMIQUE ET BIOINFORMATIQUE
Présentée par

**Mohamed BELHOCINE**

Pour obtenir le grade de docteur de l'Université Aix-Marseille

# Etude bioinformatique de l'épigénome au cours de la différenciation des lymphocytes T et des leucémies

Soutenue publiquement le 13 Décembre 2016 devant le jury composé de:

| | | |
|---|---|---|
| Pr. Jacques VAN HELDEN | AMU | Président du Jury |
| Pr. Daniel GAUTHERET | Université Paris-Sud | Rapporteur |
| Dr. Bernard MARI | IPMC | Rapporteur |
| Dr. Denis PUTHIER | TAGC | Examinateur |
| Pr. Vahid ASNAFI | Hôpital Necker | Co-directeur de thèse |
| Dr. Salvatore SPICUGLIA | TAGC | Directeur de thèse |

# UNIVERSITE D'AIX-MARSEILLE

ECOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE
Technologies Avancées pour le Génome et la Clinique
INSERM U1090

## THESE DE DOCTORAT

MENTION GENOMIQUE ET BIOINFORMATIQUE
Présentée par

**Mohamed BELHOCINE**

Pour obtenir le grade de docteur de l'Université Aix-Marseille

# Etude bioinformatique de l'épigénome au cours de la différenciation des lymphocytes T et des leucémies

Soutenue publiquement le 13 Décembre 2016 devant le jury composé de:

| | | |
|---|---|---|
| Pr. Jacques VAN HELDEN | AMU | Président du Jury |
| Pr. Daniel GAUTHERET | Université Paris-Sud | Rapporteur |
| Dr. Bernard MARI | IPMC | Rapporteur |
| Dr. Denis PUTHIER | TAGC | Examinateur |
| Pr. Vahid ASNAFI | Hôpital Necker | Co-directeur de thèse |
| Dr. Salvatore SPICUGLIA | TAGC | Directeur de thèse |

A quoi bon soulever des montagnes
quand il est si simple de passer par-dessus ?
Boris Vian

# Résumé

Les analyses génomiques à grande échelle ont modifié notre compréhension de l'organisation du génome. En particulier, des études récentes ont mis en évidence qu'au moins 70% du génome humain est transcrit et produit une myriade d'ARN non codants. Parmi eux des classes d'ARN moins bien caractérisées, les longs ARN non codants (ou lncRNA). Bien que la fonction de la grande majorité des lncRNA reste largement inexplorée, il semble que ces transcrits jouent un rôle important dans plusieurs processus physiologiques, telles que la régulation épigénétique de l'expression génique ou la maturation des ARN messagers. De plus, plusieurs travaux récents indiquent qu'ils seraient impliqués dans le développement de pathologies humaines.

Cependant, il est difficile de distinguer entre un rôle fonctionnel ou le fait qu'ils soient le reflet d'une séquence régulatrice. En effet, très souvent ils sont produits à partir des séquences régulatrices des gènes codants. Néanmoins, au début de ma thèse j'ai utilisé des données de RNA-Seq sens-spécifique pour identifier les transcrits divergents dans les tissus primaires de souris. J'ai utilisé aussi des données ChIP-Seq afin d'analyser leurs caractéristiques épigénétiques. Nous avons trouvé que la transcription divergente est associée de manière significative à des gènes liés à la régulation de la transcription et le développement. Ces promoteurs présentent des caractéristiques génétiques et épigénétiques uniques. Ces résultats suggèrent fortement un lien fonctionnel entre la transcription divergente et la régulation des gènes codant pour des facteurs de transcription et de développement et ainsi pourraient illustrer un mécanisme supplémentaire pour affiner l'expression des gènes. Cette étude a donné lieu à une publication dont je suis cosignataire en premier auteur. (Lepoivre, Belhocine et al. BMC Genomics 2013)

Dans un deuxième temps, je me suis intéressé à l'identification et la caractérisation des lncRNA au cours du développement des lymphocytes T humains et dans les leucémies aigues lymphoblastiques T (LAL-T). Tout d'abord, j'ai identifié des nouveaux lncRNA non annotés dans les bases de données disponibles et défini leurs structures génomiques. Ensuite, et après avoir complété les bases d'annotation avec ses nouveaux lncRNA, j'ai appliqué des approches statistiques pour quantifier leur expression et identifier ceux qui sont (dé)régulés dans un contexte normal ou leucémique en utilisant des données de séquençage à haut débit directionnelles publiques ou produites dans le laboratoire.

Dans un troisième temps, je me suis interrogé sur les approches les mieux adaptées permettant de prioritiser les lncRNA ayant un rôle fonctionnelle. En effet, les approches actuelles qui consistent à attribuer un rôle aux lncRNA en fonction des gènes voisins ou co-régulés (culpabilité par association) ne sont pas satisfaisantes. Ainsi j'ai émis l'hypothèse que les données épigénomiques, en particulier les informations qualitatives, pourraient permettre d'identifier les lncRNA les plus relevant dans un contexte biologique normal ou pathologique. Cette hypothèse est basée sur des observations réalisées par le laboratoire et développées au cours de ma thèse. En effet, nous avons relevés un lien direct entre un profil épigénomique particulier (profils H3K4me2/3 élargis) et la fonction des gènes. Ce concept a été récemment repris et validé expérimentalement par d'autres laboratoires. Au cours de ma thèse, je me suis attaché à étudier le mécanisme moléculaire à l'origine de ce marquage épigénomique ainsi qu'à développer un pipeline bioinformatique permettant d'identifier les gènes (codant ou non codant) associés à des profils H3K4me2/3 étendus. Ainsi, j'ai mis en évidences que ces profils étendus étaient directement dépendants d'un processus transcriptionnel impliquant des nouveaux mécanismes de régulation. Cette étude a donné aussi lieu à une publication dont je suis cosignataire en premier auteur. (Zacarias, Belhocine et al. Journal of Immunology 2015).

En conséquence, durant la dernière partie de ma thèse, j'ai combiné les données d'expression (RNA-seq) et épigénomique (ChIP-seq) au cours du développement des lymphocytes T humains et dans les LAL-T afin d'identifier les lncRNA dérégulés et associés à un profil H3K4me2/3 étendu et ainsi prioritiser les lncRNA qui auraient la plus forte probabilité de jouer un rôle dans le développement T et/ou dans la leucémogenèse. Ces analyses bioinformatiques nous ont permis de sélectionner certains candidats très prometteurs qui font actuellement l'objet d'une validation expérimentale. Ces résultats font l'objet d'un manuscrit en préparation.
Cette nouvelle approche devrait s'avérer très utile dans d'autres modèles développementaux et/ou pathologiques et pourrait être utilisée comme outil de prioritisation des candidats les plus relevant dans des approches plus globales.

Mots clés : LncRNAs, hématopoïèse, Leucémies et H3K4me3

# Abstract

High throughput genomic analyses have changed our understanding of genome organization. In particular, recent studies indicate that at least 70% of the human genome is transcribed into a myriad of both coding and non-coding RNAs. Among the non-coding RNA classes are the less well characterized long non-coding RNAs (or LncRNAs). Although the function of the vast majority of this LncRNAs remains largely unexplored, it appears these transcripts play an important role in many physiological processes, such as epigenetic regulation of gene expression or the processing of messenger RNAs. In addition, several recent studies indicate their involvement in numerous human diseases.

However, it is difficult to distinguish between a functional role and the reflection of a regulatory sequence. Indeed, very often LncRNAs are produced from the regulatory sequences of coding genes. Nevertheless, at the beginning of my thesis I used RNA-Seq data to identify divergent transcripts in mouse primary tissues. I also used the ChIP-Seq data to analyze their epigenetic characteristics. The results demonstrated that divergent transcription was significantly associated with genes related to transcription regulation and development. These promoters have unique genetic and epigenetic characteristics. The results strongly suggest a functional link between the divergent transcription and regulation of genes coding for transcription factors and development, thus illustrating an additional mechanism to refine gene expression. This study resulted in a publication of which I am joint-first author. (Lepoivre, Belhocine et al. BMC Genomics 2013)

In a second phase, I was interested in the LncRNAs identification and characterization during the development of human T lymphocytes and in T acute lymphoblastic leukemia (T-ALL). First, I have identified new non-annotated LncRNAs and defined their genomic structures using the public databases. After completing the LncRNAs catalogue with this LncRNAs newly identified, I applied statistical approaches to quantify their expression and identify those that are regulated in a normal or leukemic context using high throughput sequencing data public or produced in the laboratory.

Subsequently, I determined the most appropriate approach to prioritize the functional role of LncRNAs. Indeed, current approaches which assign functional roles to LncRNAs based on the role of their neighbors or co-regulated genes (guilt by association) are not satisfactory. So I suggested that the epigenomic data, in particular, qualitative information, could identify the key LncRNAs involved in a normal biological versus a pathological context. This hypothesis is based on observations made by the laboratory and developed during my thesis. Indeed, we identified a direct link between a particular epigenomic profile (extended profiles of H3K4me2/3) and gene function. This concept was experimentally validated by other laboratories. During my thesis, I focused on studying the molecular mechanism behind this epigenomic marking and developed a bioinformatics pipeline to identify genes (coding or non-coding) associated with the extended profiles of H3K4me2/3. Evidence generated through the pipeline demonstrated that these extended profiles were directly dependent on specific transcriptional process involving new regulatory mechanisms. This study also gave rise to a publication of which I am joint-first author. (Zacarias, Belhocine et al. Journal of Immunology 2015).

Consequently, during the last part of my thesis, I compared the expression (RNA-seq) and epigenomic (ChIP-seq) data of developing human T cells to that of T acute lymphoblastic leukemia (LAL-T) to identify LncRNAs associated with deregulation and with an extended H3K4me2/3 profile. This approach prioritized LncRNAs that would have the highest probability of playing a role in T cell development and / or in leukemogenesis. These bioinformatics analyses allowed us to select promising candidates that are currently undergoing experimental validation (manuscript in preparation).

In conclusion, this body of work has resulted in a more comprehensive approach to determining the true functional role of LncRNAs in both normal biological context and in human disease.

Keywords: LncRNAs, hematopoiesis, Leukemia and H3K4me3.

# Remerciements

Tout d'abord je tiens à remercier chaleureusement les membres de mon jury de thèse : Pr. Jacques VAN HELDEN, Pr. Daniel GAUTHERET et Dr. Bernard MARI qui ont accepté sans hésitation d'évaluer mon travail.

Je tiens à remercier particulièrement le Dr. Spicuglia Salvatore, mon directeur de thèse qui m'a accompagné de près pendant ces huit longues années et de qui j'ai beaucoup appris. Sur le plan scientifique tout d'abord car tu as su m'enseigner les bases du domaine de recherche et tu m'as permis de me former progressivement et de développer mes compétences bioinformatiques et biologiques. Ton implication à la fois scientifique et humaine au cours de ces années a créé une ambiance motivante et des discussions scientifiques et personnelles toujours très intéressantes. Tu as aussi su me pardonner mes approximations et mon manque d'organisation. Grâce à ta patience, et à ton insistance pour me convaincre de lire davantage, j'ai appris plus que je n'aurais imaginé. Je te souhaite le meilleur et je ne doute pas que l'on se recroisera dans le futur.

Je tiens à remercier aussi le Pr. Asnafi Vahid mon co-directeur de thèse, pour ta très grande disponibilité, ton écoute et ton suivi exemplaire tout au long de ces quatre années de thèse. J'ai pu apprendre beaucoup auprès de toi, tant sur le plan scientifique et professionnel que sur le plan personnel à travers tes conseils et tes encouragements. Je souhaite à tous les doctorants de pouvoir être encadrés par une personne comme toi.

Je remercie également Pr. Macintyre Elizabeth, directrice du laboratoire d'oncohématologie à l'hôpital Necker des enfants malades à Paris et Pr. Catherine NGUYEN, directrice du TAGC à Marseille, pour m'avoir accueilli au sein de leurs laboratoires respectifs et pour m'avoir accordées leurs confiances pendant ces années.

Toute ma gratitude va à mes chers professeurs durant mon cursus universitaire, mais une mention très spéciale, du côté bioinformatique, va au Dr Puthier Denis, tu as été pour moi durant ces années mon enseignant, ensuite, mon encadreur puis mon collègue pour enfin devenir un cher ami. Je n'oublierai jamais tous ces moments où tu m'as sauvé des situations délicates, bien sûr je parle de programmation.

Coté paillasse, je remercie tout d'abord Laurent Vanhille et Eve-Lyne Mathieu des collègues estimés avec qui j'ai partagé des beaux moments au laboratoire. Que ce soit pendant les pauses café ou les longues journées de travail au laboratoire, votre sérieux et votre humour inimitable ont beaucoup apporté à l'équipe.

Je remercie également tous les membres de l'équipe Salva : Lan, Marco, Guillaume, Ariel, Aurélien, Charbel, Jaafar, Saadat, Idir, Yasmine et Wiam qui ont créé une ambiance vraiment appréciable au sien de l'équipe. Et également, l'équipe Vahid à Paris : Amélie, Aurore, Agata, Patrick, Jonathan et tous les autres avec qui j'ai passé des moments formidables qui m'ont fait oublier le froid de Paris.

Je remercie AMidex pour les trois ans de financement de thèse et également la Fondation ARC pour la recherche sur le cancer d'avoir financé ma 4ème année de thèse.

Je tiens aussi à remercier mes amis qui sont toujours là pour moi, Amine, Abdou, Billel, Redoune, Hsan, Firas et Mohammed, désolé si j'oublie quelqu'un mais vous êtes tous dans mon cœur.

Je remercie, particulièrement, mon cher ami Idir Kacel qui m'accompagne depuis de nombreuses années où j'ai apprécié tous les moments de travail et de détente avec toi. Tu m'as soutenu pendant toute ces années, et tu m'a toujours fais confiance. Sans ton aide je n'aurais certainement pas pu poursuivre les études qui m'intéressaient et prendre autant de plaisir dans un métier aussi passionnant. Si la distance nous sépare maintenant, sache que tu me manque beaucoup et que tu seras toujours accueilli avec grand plaisir si l'envie te prend de me rejoindre à Dubaï.

Je tiens à remercier tous les autres membres des laboratoires TAGC et Necker. Merci de votre soutien et de toutes les soirées qui m'ont fait oublier le quotidien

Et pour finir, je remercie chaleureusement mes parents, Boudjema et Fatiha, pour m'avoir si bien réussie déjà, mais également pour leur soutien immuable même après 27 ans d'étude… Promis je m'arrête là ! Mes frères : Abd-El-Kader, Sif El-Dine, Nabil et Abd-El-Raouf. Mon unique sœur Amina et mes belles sœurs Kamilla et Soumiya, sans oublier : Hiba, Loudjayn et Omar Rassim. Je remercier aussi ma belle-famille : Mama Sahar, Amou Akram, Ali, Abd Alah, Ghadir, Abir et Lama.

Et non je ne vous ai pas oublié, de façon très particulière je tiens à remercier ma femme Nour et ma fille Joury qui m'ont soutenu et avec qui j'ai vécu au plus près pendant cette période. Leur amour a été ma source d'énergie que je puise tous les jours ; leur patience à mes longues journées sur l'ordinateur et surtout pour leur courage de m'avoir suivi partout où je déménage Paris, Marseille et maintenant Dubaï notre nouveau départ …… Inch'Allah.

Enfin je tiens à dédicacer ce manuscrit à mes grand parents Ben Ali, Sayeh, Massouda et spécialement ma grand-mère Nedjma, parti trop tôt, mais dont je sais qu'elle aurait été fière de moi.

# Table des matières

# Avant-propos

L'étude des organismes vivants ne cesse de dévoiler la diversité et la complexité des différentes formes de vie. Les premiers biologistes ont décrit les caractéristiques morphologiques des espèces, puis les ont recensées et classées par groupes en fonction de leurs similarités. En parallèle, l'étude de leurs caractéristiques progressait et les biologistes s'intéressaient aux mécanismes de plus en plus fondamentaux, révélant petit à petit la complexité des êtres multicellulaires. La découverte de la cellule est à la base des connaissances accumulées sur le vivant. En effet, depuis l'axiome principal de la 'théorie cellulaire' (Schleiden et Schwann, en 1838), la communauté scientifique s'efforce de comprendre les mécanismes qui régissent le comportement cellulaire.

La découverte de l'ADN, de sa structure, puis du code génétique qui permet de traduire les codons (triplets de nucléotides) en acides aminées, représentent une révolution pour la compréhension du vivant (Avery et al., 1944, Nirenberg and Leder, 1964). Les chercheurs ont désormais un moyen simple et universel pour interpréter le code génétique. Plus tard, le séquençage des génomes permet d'établir la cartographie des gènes (lesquels codent l'information pour la synthèse de protéines (International Human Genome Sequencing, 2004), et d'autres unités de séquence (ou d'information) qui seront détaillées dans l'introduction biologique de ce manuscrit.

Diverses fonctions sont associées aux protéines et aux gènes correspondants. Un contrôle fin de leur expression est nécessaire pour maintenir un état cellulaire stable et permettre aux cellules d'assurer leurs fonctions. Ainsi, chaque cellule peut exhiber un profil d'expression génétique différent, et peut l'adapter en fonction de signaux particuliers, ou de phénomènes extérieurs (changement de milieu, stress physique, chimique…). Nous verrons, que la régulation de l'expression des gènes est complexe et peut faire intervenir des mécanismes à divers niveaux. Du plus tardif comme la régulation de la population d'ARN ou de protéines (exosome, protéasome), au plus précoce par la régulation de la transcription des gènes, qui nous intéresse dans ce manuscrit.

Les dernières avancées des technologies de séquençage et d'analyse ont révélé une complexité insoupçonnée du génome humain. En effet, des analyses génomiques à grande échelle ont mis en évidence l'existence d'une multitude de longs ARN non codants (LncRNAs). Bien que la fonction de la majorité de ces LncRNAs reste à l'heure actuelle inexplorée, il semble clair que plusieurs de ces transcrits jouent un rôle important dans la régulation de l'expression génique et sont impliqués dans diverses pathologies, notamment le cancer.

Au cours de ma thèse, j'ai été amené à développer des pipelines d'analyse bioinformatiques afin d'intégrer de nombreuses données hétérogènes issues de techniques à haut débit, principalement du séquençage. Cette approche intégrative m'a permis d'étudier les mécanismes qui régissent la régulation des gènes dans un contexte normal et pathologique. Plus précisément, je me suis intéressé à la régulation de la transcription chez l'homme et la souris en étudiant les caractéristiques et l'activité des promoteurs et des enhancers.

# CHAPITRE I : Généralités

## 1. Le génome des eucaryotes :

### 1.1. La structure du génome :

Le génome est l'ensemble du matériel génétique d'une espèce codée sous forme d'acide désoxyribonucléique (ADN). On distingue le génome nucléaire contenu dans le noyau et le génome non-nucléaire : mitochondrial ou chloroplastique. Chez les eucaryotes, le noyau est organisé en de multiples territoires tel que le nucléole, la lamina, les Paraspeckles, les corps PML (*Promyelotic Leukemia Protein*) et les corps Cajal (Lanctot et al., 2007).

Le matériel génétique est replié et compacté sous forme de structures non aléatoires. Ces structures sont très importantes dans le contrôle des processus transcriptionnels. On peut distinguer plusieurs échelles de compaction qui s'appuient sur des mécanismes distincts, d'abord à l'échelle chromatinienne par le biais des histones, puis à l'échelle plus grande grâce à des structures d'ordre supérieur.

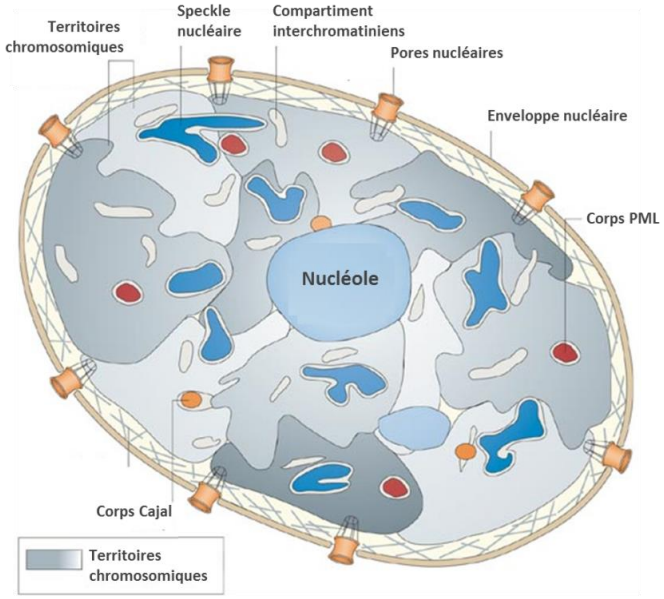



**Fig. 1: Organisation fonctionnelle du noyau : les différents compartiments nucléaires (Lanctot et al., 2007)**

Chez les eucaryotes, en général, les fonctions fondamentales telles que le développement, le métabolisme ou la différenciation cellulaire nécessitent l'intervention des complexes ARN/Protéines qui contrôlent l'expression génique afin d'assurer son bon déroulement. Ces complexes contrôlent l'organisation de l'ADN au sein de la chromatine et qui permettrait de réguler la transcription. Ainsi, l'ADN structuré en double hélice de 2nm de diamètre subit une première étape de compaction afin de former l'unité structurale de la chromatine : Le nucléosome (Luger et al., 1997) (Fig2).

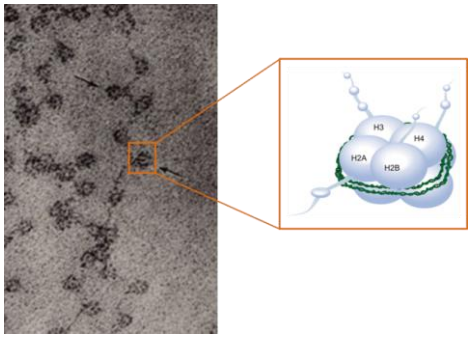



**Fig. 2 : Structure chromatinienne : Une image de microscopie électronique illustre la structure en 'collier de perles' de la chromatine. Les flèches pointent sur des nucléosomes autour desquels la molécule d'ADN est enroulée. Les portions d'ADN entre chaque nucléosome sont appelés 'linker DNA'. Marque d'échelle : 10nm. D'après (Olins and Olins, 1974). Un zoome sur la structure canonique des nucléosomes : elle est composée de deux tétramères formé par l'assemblage des histones H2A H2B H3 et H4. Les nucléosomes interagissent avec l'ADN et les queues des histones sont accessibles à l'environnement nucléaire. D'après (Marks et al., 2001).**

## 1.2. La structure de la chromatine :

Le nucléosome est composé d'un octamère d'histone (H2A, H2B, H3 et H4) entouré d'un fragment d'ADN de 146 paires de bases (pb) (Luger et al., 1997). Des observations réalisées par microscopie électronique indiquent que les nucléosomes s'organisent en structure de type « collier de perle » (Fig. 2). Le niveau de la condensation de la chromatine dépend de la concentration de l'histone Linker H1, ainsi, on parle soit d'euchromatine ou soit d'hétérochromatine. Ces deux niveaux correspondent, respectivement, à une structure décondensée transcriptionnellement active ou à une structure dense transcriptionnellement inactive (Meshorer and Misteli, 2006) (Fig. 3).

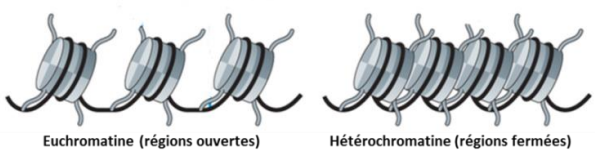



**Fig. 3 : Etats de la chromatine en fonction de la densité en nucléosomes : l'euchromatine (à gauche), état relâché de la chromatine et l'hétérochromatine (à droite), état compacté de la chromatine. Ainsi, l'ADN est plus accessible lorsque la chromatine est à l'état d'euchromatine.**

Par contre, l'empaquetage de l'ADN en nucléosomes raccourcit la longueur de la fibre de l'ADN d'environ sept fois. En d'autres termes, un morceau d'ADN qui fait 1 mètre de

long va devenir une fibre de chromatine «collier de perles » de seulement 14 centimètres de long. Malgré ce raccourcissement, la chromatine est encore beaucoup trop longue pour tenir dans le noyau, qui est typiquement seulement 10 à 20 microns de diamètre. Par conséquent, la chromatine est en outre enroulée en une fibre plus épaisse encore plus courte, d'environ 30 nanomètres de diamètre (Robinson and Rhodes, 2006) (Fig. 4).

Au fil des ans, il y a eu beaucoup de spéculations concernant la manière dont les nucléosomes sont pliées en fibres de 30 nm (Woodcock, 2005). Une partie du problème réside dans le fait que la microscopie électronique a du mal à discerner les nucléosomes après que la fibre est formée. Ainsi, la fibre 30 nm peut être très irrégulière et pas tout à fait uniforme (Bednar et al., 1998). Fait intéressant, l'histone H1, très importante dans la stabilisation de structures du nucléosome, est aussi très importante pour la structure d'ordre supérieur des fibres 30 nm. Ainsi, la formation des fibres de 30 nm est forment favorisée quand H1 est présent en abondance.

Lorsque les cellules eucaryotes se divisent, l'ADN génomique doit être également divisé en deux cellules filles. Pour ce faire, l'ADN devient très compacté dans les chromosomes métaphasiques classiques qui peuvent être vus avec un microscope optique. Une fois qu'une cellule est divisée, ses chromosomes se déroulent à nouveau.

Si on compare la longueur des chromosomes en métaphase à celle de l'ADN nu, le taux de compaction de l'ADN dans les chromosomes métaphasiques est d'environ 10000 fois (en fonction du chromosome). Cela peut être considéré comme semblable à prendre une corde aussi longue qu'un terrain de football et la compacter à moins de deux centimètres et demi. Ce niveau de compaction est réalisé par pliage de façon répétée des fibres de chromatine dans une hiérarchie de boucles multiples et des bobines (Robinson and Rhodes, 2006, Bednar et al., 1998) (Fig. 4). Exactement comment cela est accompli difficile de le dire, mais nous savons que la phosphorylation de l'histone H1 joue un rôle très important. En effet, ceci reste à explorer dans les années à venir.

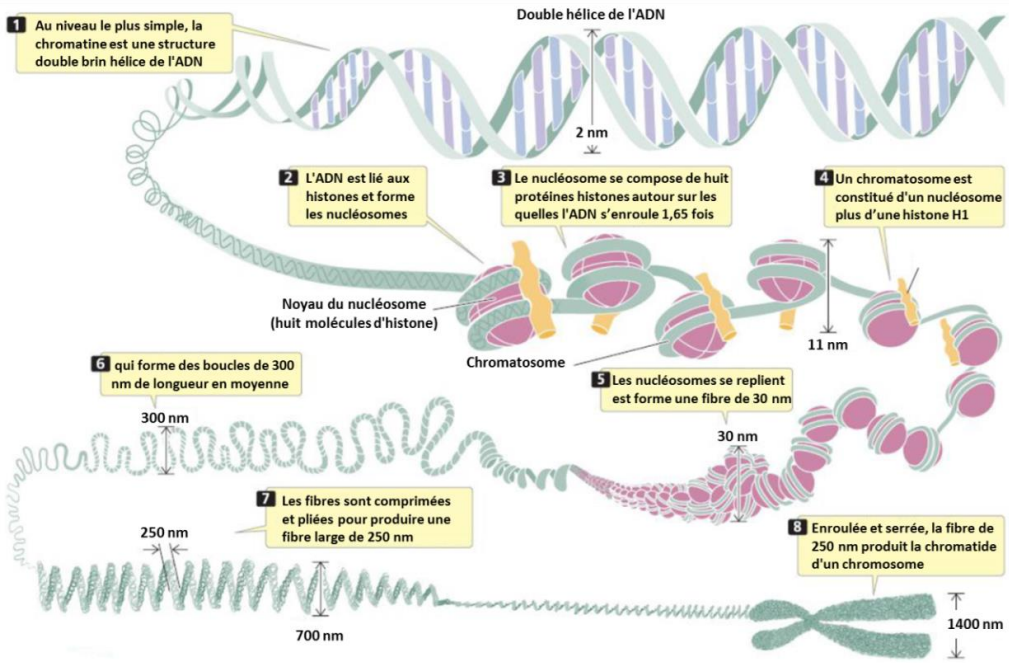



**Fig. 4 : Vue de la chromatine à différentes échelles.**
**Cette image schématise la conformation de l'ADN sous différents états de compaction. D'abord libre (1), puis apparié aux nucléosomes 'en collier de perles' pour former la fibre de 11nm (2-4), l'agencement en fibre de 30nm (5), et finalement les structures d'ordre supérieur qui permettent une compaction extrême de la chromatine et l'agrégation en chromosomes (6-8).** 

# I.2 - La régulation de l'expression des gènes:

## 2.1 L'importance de la régulation des gènes:

Un organisme possède des milliards de cellules avec un génome identique, alors que la diversité des types cellulaires est considérable. Avec entre 210 et 411 types cellulaires différents, la régulation de l'expression des gènes doit être finement régulée de façon spécifique afin d'assurer cette grande diversité (Vickaryous and Hall, 2006). Ainsi, deux niveaux de régulation s'imposent, une régulation qualitative pour déterminer le gène à exprimer, et une régulation quantitative pour définir exactement le niveau d'expression de chaque gène. En conséquence, on peut classer les gènes en trois catégories : (i) les gènes tissu-spécifique exprimés dans un nombre très restreint de type cellulaire ; (ii) les gènes de ménage qui s'expriment en général dans plusieurs types cellulaires à des niveaux constant et (iii) les gènes ubiquitaires qui s'expriment aussi dans différents types cellulaires mais à des niveaux variables (Kouadjo et al., 2007, Velculescu et al., 1999).

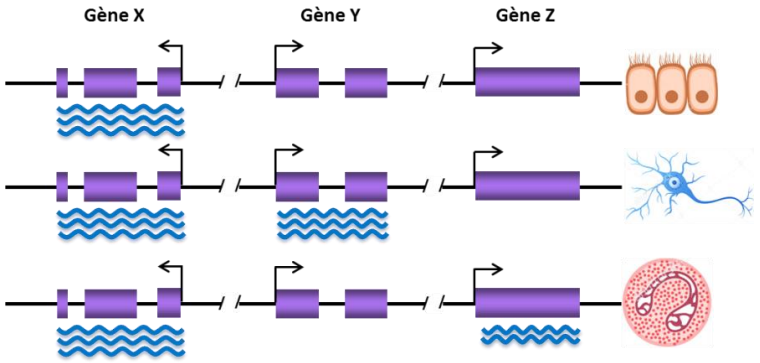



**Fig. 5 : Le modèle de la régulation des gènes : La notion des gènes de ménage et les gènes tissu-spécifique. Le gène X: gène de ménage, les gène Y, Z : gènes tissu-spécifique.**

De ce fait, chaque cellule possède un profil d'expression unique en fonction de la combinaison des gènes exprimés et leurs taux d'expression. Un exemple de cette régulation fine représenté par les premières étapes de la différenciation cellulaire lors de l'embryogenèse qui va permettre les trois principaux feuillets embryonnaires. Ces feuillets vont à leurs tours se différencier et se spécialiser progressivement afin de produire les centaines de types cellulaires qui constituent un organisme.

Enfin, cette régulation de l'expression dépend d'un grand nombre de processus et mécanismes différents qui peuvent intervenir lors de chaque étape des processus de la transcription, la post-transcription, la traduction et la post-traduction.

## 2.2 Les éléments régulateurs de la transcription :

En plus de la régulation de la structure chromatinienne en eu/hétérochromatine, la transcription génique est contrôlée par des séquences d'ADN régulatrices et des facteurs protéiques clés (Les Facteurs de Transcription – FT). Ces régulateurs peuvent se localiser plus ou moins proches aux sites d'initiation de la transcription des gènes (TSS). En effet, nous pouvons définir des régions régulatrices en *Cis* ou en *Trans*. Ainsi, les régions en *Cis* situées au niveau du site TSS sont appelées les promoteurs, alors que les régions distales sont définis en fonction de leur action en enhancers, insulateurs ou silenceurs (Fig. 6).

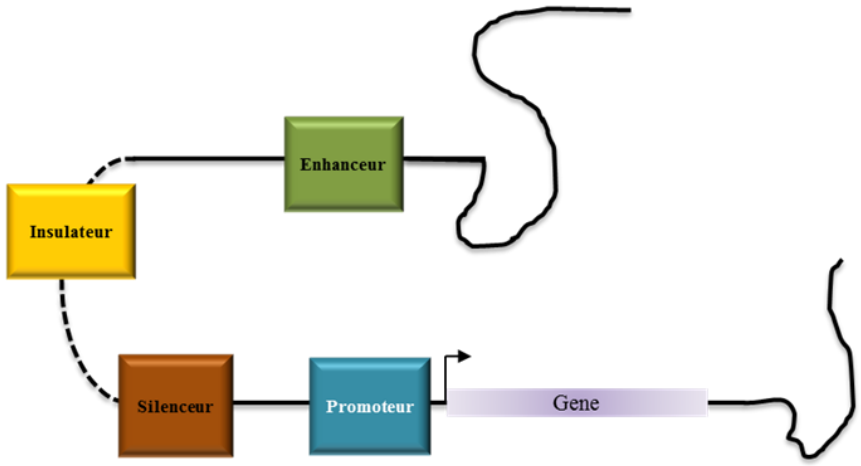



**Fig. 6 : Les différentes classes d'éléments régulateurs : promoteur, *enhancer*, *silencer* et *insulator*.**

### a) Les Promoteurs :

Les promoteurs des gènes sont des régions très bien caractérisées, car elles sont à la base de la régulation de l'expression. On distingue deux grandes catégories de promoteurs : (i) les promoteurs ubiquitaires qui promeuvent l'expression des gènes de manière systématique comme par exemple les gènes de ménage. (ii) Les promoteurs inductibles qui provoquent la transcription du gène en réponse à un stimulus dans un tissu ou un stade développement particulier. (Fig. 5)

Par définition, les promoteurs correspondent aux séquences minimales nécessaires à l'initiation du processus de la transcription (Heintzman et al., 2007). Ils sont constitués de plusieurs motifs de reconnaissance pour les éléments du complexe basal de transcription telles que la boîte TATA, le motif *Initiator* (Inr), les motifs spécifiques aux facteurs TFIIB, le « *Downstram* Promotor *Elements* » (DPE), le « *Downstram Core Element* » (DCE) et sans oublier des régions denses en di-nucléotides CG (Les ilôts CpG) (Roy and Singer, 2015). (Fig. 7)

En plus de cette région basale (autour de +/-35 nucléotides), les promoteurs sont constitués d'une région relativement proximale située à environ 100 ~ 200 pb en amont du TSS. Cette région englobe des sites de reconnaissance très spécifiques de recrutement des facteurs de stimulation ou d'inhibition de la transcription (Ong et al. 2011).

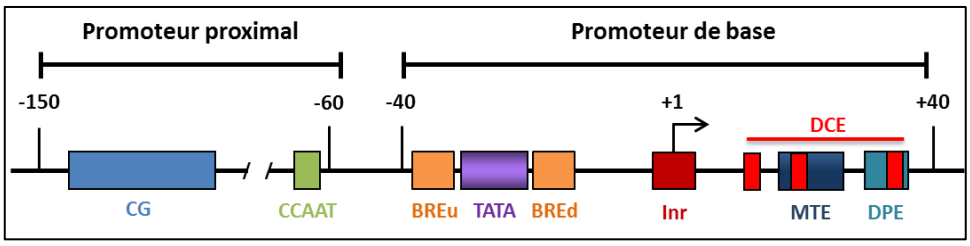



**Fig. 7 : Liste non exhaustive d'éléments régulateurs identifiés au niveau du promoteur proximal et du promoteur de base des gènes : BREu et BREd (*TFIIB recognition Elements upstream* et *downstream*), boîte TATA, Inr (*Initator*), MTE (*Motif Ten Element*), DPE (*Downstream Promoter Element*).**

## b) Les enhancers :

Les enhancers sont définis comme des éléments de séquence capable de provoquer l'activation d'un gène distant. Ces séquences assurent leur fonction indépendamment de leur orientation et de leur localisation sur le génome (Banerji et al., 1983, Gillies et al., 1983). Ils sont caractérisés par une importante densité de motifs de liaison aux FT et sont relativement bien conservés au cours de l'évolution (Rada-Iglesias et al., 2011, Ovcharenko et al., 2005). Le rôle des enhancers est crucial lors du développement embryonnaire et la différenciation cellulaire (Visel et al., 2009). Les altérations génétiques des enhancers expliquent de nombreuses maladies génétiques dues à leur rôle très important dans les contrôles spatiaux/temporaux de l'expression des gènes.

Des études récentes utilisant des approches à grande échelle ont permis d'observer la présence de l'ARN polymérase II active sur ces régions enhancers. Ces observations confirmées par l'identification des transcrit naissant sur ces enhancers ainsi dite d'ARN enhancer (ou eRNA)(Wang et al., 2001, Kim et al., 2010).

Les modèles proposés suggèrent que la plupart des enhancers fassent intervenir des boucles sur l'ADN pour rapprocher les facteurs de transcription du promoteur. Ces boucles ont été mises en évidence grâce à la technique de Capture de Conformation Chromatinienne (3C) (Drissen, Palstra et al. 2004).

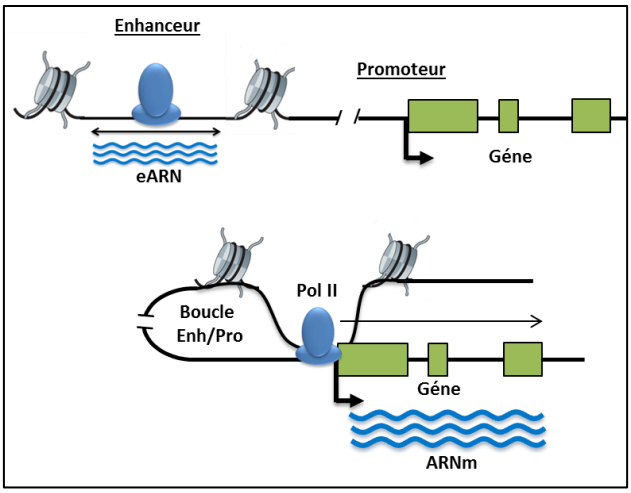



**Fig. 8 : Modèle d'activation de la transcription au niveau du promoteur par la création d'une boucle avec l'enhanceosome.**

### c) Les *silencers* :

De la même manière que les enhancers, les silencers sont des éléments de séquences impliqués dans la répression de l'expression génique. Ils sont capables d'agir indépendamment de leur position. Les silencers interagissent avec des facteurs de répression de la transcription (Dellino et al., 2004).

### d) Les *insulators*:

Les *insulators* définissent les frontières des domaines topologiques tels que les promoteurs et les enhancers/silencers. Ils sont très riches en protéine CTCF dont son rôle sera dans la formation des boucles d'interaction à longue distance entre les insulateurs. En conséquence, ces boucles forment des partitionnements dans la chromatine de sorte que les enhancers ou silencers d'une boucle ne puissent pas agir sur les promoteurs d'une autre boucle (Geyer, 1997, Gerasimova et al., 2000).

Les *insulators*, comme les autres éléments régulateurs, interagissent avec des FT, dont certains sont communs avec les promoteurs. Enfin, les altérations génétiques des *insulators* peuvent induire des perturbations dans l'expression génique provoquant des maladies graves.

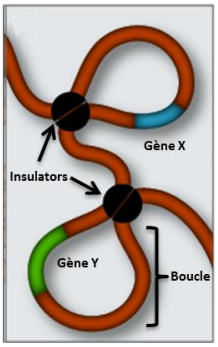



**Fig.9 : Représentation schématique d'un *insulator* et les barrières génétiques : Les gènes X et Y sont normalement confinés spatialement dans des boucles séparées. Ses boucles sont formées par des interactions à longue distance entre les *insulators*.**

### e) Les TIP, les broad promoters et les différentes régions classées selon la méthylation de l'ADN :

Divers autres types de régions, moins fréquemment citées dans la littérature, ont également été décrits ces dernières années. On retrouve par exemple les TIP (Transcription Initiation Platform) qui sont définis comme des régions de plusieurs centaines de paires de bases généralement situées en amont de gènes tissu-spécifiques et qui possèdent un niveau élevé d'ARN Polymérase II (Koch et al., 2011).

Les promoteurs étendus (Broad) correspondent quant à eux à des promoteurs qui possèdent de larges domaines pour les marques épigénétiques H3K4me2 et H3K4me3 (Di et tri-méthylation de la lysine 4 de l'histone 3) (Pekowska et al., 2010, Benayoun et al., 2014, Chen et al., 2015a). Je vais décrire ces domaines avec plus de détails dans le chapitre II (Les domaines étendus de la méthylation Histone 3 lysine 4).

Finalement, de nombreuses régions génomiques sont définies en fonction de leur niveau de méthylation (Xie et al., 2013). On retrouve ainsi les UMR (*UnMethylated Region*), les LMR (*Low Methylated Regions*), les PMD (*Partially Methylated Regions*) et les FMR (*Fully Methylated Regions*). Enfin, les DMV (*DNA Methylation Valley*) correspondent à de grandes régions génomiques (de 5kb à 68kb) non méthylées pouvant contenir un ou plusieurs gènes.

En général, ces régions manquent de caractérisation fonctionnelle. Leur identification en majeur partie est basée uniquement sur des analyses bioinformatiques, leur fonction d'activateurs transcriptionnels reste à confirmer expérimentalement.

### f) Les LCR :

Les LCR (*Locus Control Regions*) sont des régions génomiques regroupant les éléments régulateurs (enhancers, silencers, insulators) nécessaires à la régulation des gènes d'un locus particulier (Li et al., 2002). Ces régions peuvent intervenir à longue distance sur le locus et indépendamment de leur position dans le génome. L'ensemble des éléments régulateurs qui composent ces régions est essentiel pour définir le profil d'expression du locus cible. Ainsi, leur fonction est fortement liée à leur composition particulière. Plusieurs modèles tentent d'expliquer le mécanisme d'action des LCR sur les locus cibles, comme par exemple : le modèle de *looping*, du *tracking* ou du *linking* (Bulger and Groudine, 2010). Un point commun à ces différents modèles est la capacité du LCR à altérer la conformation de la chromatine et ainsi, ouvrir le locus cible et permettre la transcription des gènes cibles.

Les LCR sont généralement associés à des loci et gènes tissu-spécifiques. Comme est le cas avec le premier LCR à avoir été identifié, et l'un des mieux caractérisé, le LCR du locus de la β-*globin* qui se trouve entre -25kb à -6kb en amont du gène (Grosveld et al., 1987). Les modifications ou la suppression de cette région régulatrice provoquent un dysfonctionnement dans l'expression des gènes du locus et l'apparition chez les patients d'une hémoglobinopathie nommée « thalassémie ».

### g) Les super-enhancers :

Les super-enhancers sont de larges régions génomiques (jusqu'à 50kb), décrites pour la première fois en 2013 (Whyte et al., 2013), qui sont composées d'un ou plusieurs enhancers et qui sont fortement enrichies en facteurs de transcription, cofacteurs et/ou des modifications épigénétiques en particulier la modification H3K27ac. Cet enrichissement est d'ailleurs utilisé comme critère de distinction entre les super-enhancers et les enhancers classiques lors de leur identification.

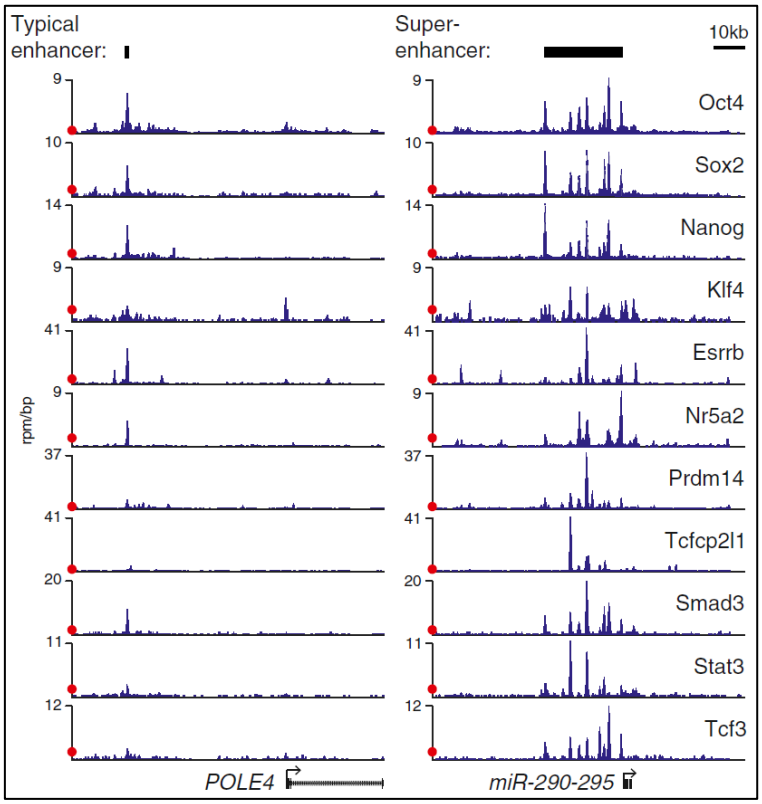



**Fig. 10: Les profils de liaison de ChIP-seq pour les facteurs de transcription indiqués aux locus POLE4 (*enhancer* canonique) et miR-290-295 (*super enhancer*) dans les cellules ESC. (Hnisz et al., 2013).**

Les gènes associés aux super-enhancers sont généralement plus tissu-spécifiques et présente une plus forte expression que ceux associés aux enhancers

classiques. Il semble aussi qu'ils soient également plus sensibles à la perte d'un facteur de transcription que les enhancers classiques. Cependant, l'existence des super-enhancers comme nouveau type de régions régulatrices reste soumise à débat dans la communauté scientifique (Pott and Lieb, 2015). Ainsi, leur définition, leur identification et leur distinction par rapport aux enhancers classiques sont basées sur des méthodes d'enrichissement en marques épigénétiques et non sur des bases fonctionnelles. En effet, leur définition est plus ou moins similaire aux types de régions déjà décrites auparavant comme les LCR.

Enfin, Une base de données (dbSUPER) a récemment été développée afin de répertorier l'ensemble des super-enhancers identifiés chez l'Homme (96 types cellulaires actuellement) et la souris (5 types cellulaires)(Khan and Zhang, 2016).

### h) Les facteurs de transcription :

Les FT sont essentiels à la régulation de l'expression génique et jouent un rôle important dans diverses fonctions cellulaires en particulier lors du développement et la différenciation cellulaire. Ce sont des protéines avec, en général, deux domaines essentiels, le premier permet la fixation sur l'ADN et le deuxième intervient dans l'interaction protéine/protéine. Grâce à ces domaines, les FT ont été classés en sous familles telles que : le *zinc Finger*, le *Forkhead* et le *Helix_loop-helix* …etc.(Wingender, 2013).

La localisation des sites de fixation des FT par rapport au gène cible a permis l'établissement de deux groupes (Roeder, 1996) : (i) les Facteurs Généraux de la Transcription (FGT ou GTF en anglais) qui se fixent sur les régions promotrices et constituent la machinerie de l'initiation de la transcription ; (ii) les autres facteurs qui se fixent aux autres éléments régulateurs cités en haut, participent à la stimulation ou la répression de la transcription des gènes cibles.

## 2.3 Les étapes de la transcription (par la RNA-Polymerase II) :

La transcription des gènes par l'ARN polymérase II se résume en trois principales étapes : l'initiation, l'élongation et la terminaison. Ces différentes étapes sont régulées par l'intervention des différents éléments topologiques décrits en haut à savoir les promoteurs, les enhancers, les silencers et les insulateurs et aussi par les facteurs transcriptionnels assurant le bon déroulement et une coordination très fine de la transcription.

### 2.3.1 L'initiation de la transcription:

La première étape de la transcription est le recrutement et l'assemblage séquentiel du complexe de pré-initiation de la transcription (PIC) qui contient l'ARN Pol II et tous les facteurs généraux (excepté TFIID). Ainsi, TFIIH permet l'ouverture de la structure de la structure double brins de l'ADN (Holstege et al., 1998, Yan et al., 1997). L'ARN Pol II réalise alors la synthèse des premiers dNTPs et entre dans une étape d'initiation abortive où elle effectue des allers-retours (mécanisme connu sous le nom *backtracking*). Lors de cette étape l'initiation peut encore être avortée (Fig. 11A-C) (Holstege et al., 1997).

### 2.3.2 L'élongation :

#### a) L'élongation précoce :

L'engagement de la transcription se produit après la formation d'ARN naissant d'une dizaine de nucléotides. L'ARN pol II s'échappe du promoteur en se dissociant du PIC. En effet, une bulle de transcription se forme où l'ARN pol II a verrouillé l'ARN naissant et le brin matrice de l'ADN (Zawel and Reinberg, 1995). En même temps, la TFIIH grâce à son domaine kinase CDK7 phosphoryle la sérine 5 du domaine carboxyl terminal (CTD) de l'ARN Pol II. Cette phosphorylation permet le recrutement des protéines impliquées dans l'élongation (Gerber and Shilatifard, 2003, Hampsey and Reinberg, 2003).

Des études ont démontré l'existence d'une phase de pause à environ 20 à 50 pb en aval du TSS chez les gènes liés au développement ou répondant à des stimuli. Cette phase de pause est due au recrutement du complexe DSIF-NELF qui se lie de façon stable à l'ARN Pol II. Les auteurs suggèrent que cette phase permet une réponse rapide notamment pour les gènes de choc thermique ou du développement.

A l'issue de cette pause deux suites sont possible : soit le relargage et l'annulation de l'initiation ; ou soit l'engagement dans l'élongation productive (Fig. 11D) (Zeitlinger et al., 2007).

#### b) L'élongation productive :

Le relâchement de la pause et la transition vers la transcription productive sont assurés par le recrutement du complexe P-TEFb (*Positive Transcription Elongation Factor b*) qui par la phosphorylation du complexe DSIF-NELF produit le relargage de la protéine NELF et la levée du blocage (Fujinaga et al., 2004, Romanov et al., 2008).

En parallèle, le complexe P-TEFb assure aussi par son activité kinase la phosphorylation de la sérine 2 du domaine CTD de l'ARN Pol II, et ainsi, permet le recrutement d'autre complexe nécessaire à la stabilisation de l'ARN Pol II. En conséquence, L'ARN pol II va pouvoir polymériser la chaîne d'ARN pour des dizaines voire même des centaines de kilo-bases avant de se dissocier de l'ADN (la phase de la terminaison) (Fig. 11E-F)(Marshall and Price, 1995, Peterlin and Price, 2006).

### 2.3.3 La terminaison de la transcription :

Le signal de la terminaison de la transcription est amené par la séquence terminatrice de l'ADN qui apporte aussi la matrice de la polyadénylation de la molécule d'ARN (Proudfoot, 2000, Proudfoot et al., 2002, Proudfoot, 2004). Deux modèles sont proposés afin de décrire le mécanisme de la terminaison (Buratowski, 2005) : (i) le modèle allostérique, qui explique la terminaison par le changement des facteurs liés à l'ARN pol II et détachement de l'ARN Pol II de l'ADN, ce changement est du à l'apparition de la séquence de la polyadénylation (Logan et al., 1987); (ii) l'autre modèle dit de « torpille » qui suggère que l'ARN est libéré par clivage et cela lorsque le complexe de la transcription rencontre le site de la polyadénylation (Connelly and Manley, 1988). Cependant, des études pan-génomiques, basées sur le séquençage à haut débit, ont observées que la machinerie de la transcription subsiste souvent plusieurs kilo-bases après la séquence terminatrice. Ce qui suggère de préférence le modèle « torpille » (Fig. 11G) (Buratowski, 2005).

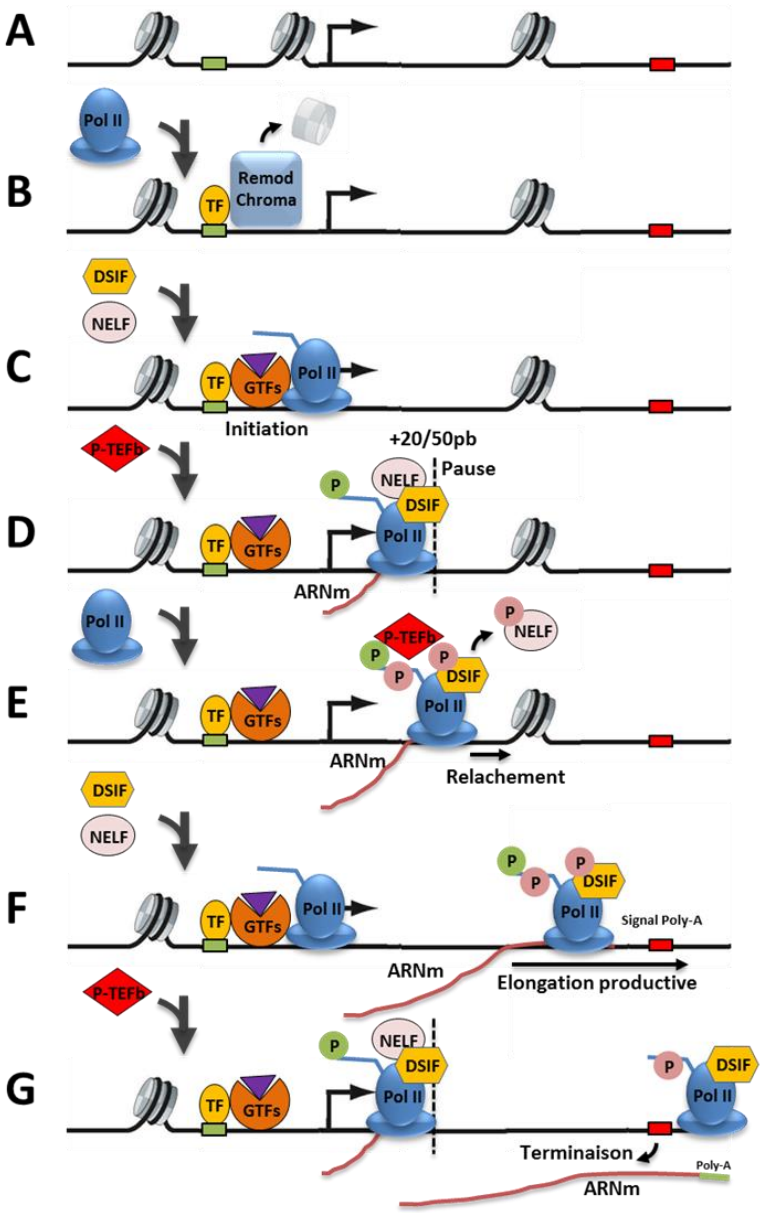



**Fig. 11 : Les étapes d'initiation de la transcription.**
**Cette figure illustre les principales étapes de l'initiation de la transcription. Apres la formation du PIC, l'initiation de la transcription et l'ouverture des brins d'ADN provoque la transcription des deux première paires de bases à partir du brin matrice. L'éventuelle transcription abortive sur le promoteur est à l'origine de petits transcrits avant l'engagement de la transcription et la transition vers la phase d'élongation. Une pause est encore décrite en aval du promoteur sur certains gènes après la transcription des premières 20 à 50 paires de bases. Le relâchement de la pause permet alors la transition définitive vers l'élongation et enfin la terminaison est marquée par le signal de la poly-Adénilation et le relâchement de l'ARNm. D'après (Margaritis and Holstege, 2008). La phosphorylation de la sérine 5 et la sérine 2 sont représentées respectivement par des ronds de couleur vert et rose.**

# I.3 - La régulation épigénétique:

Le terme épigénétique a été introduit en 1957 par Conrad Hal Waddington qui définit les modifications affectant les propriétés de l'ADN ou des protéines y associées (en particulier les histones) et transmissibles au cours des divisions cellulaires (Jaenisch and Bird, 2003).

La première modification épigénétique à avoir été décrite est la méthylation de l'ADN sur les résidus de cytosine qui sont suivis par un résidu guanine (di-nucléotide CpG). La méthylation de l'ADN est présente dans 60 à 80% des CpG du génome humain et impacte très significativement la stabilité du génome, l'expression des gènes (Smith and Meissner, 2013).

La deuxième marque épigénétique la plus étudiée est la méthylation et l'acétylation des protéines d'histone. En 2008, plus de 60 modifications post-transcriptionnelles d'histone ont été identifiées (Kouzarides, 2007). Ces modification sont étroitement liées aux mécanismes de la transcription et en régulant les éléments régulateurs, décrit dans les sections précédentes, et ainsi apportent une couche supérieure supplémentaire à la régulation transcriptionnelle.

## 3.1. La méthylation de l'ADN :

La méthylation de l'ADN est modulée par des enzymes nommées les ADN Méthyl-Transférases (DNMTs) (Deaton et al., 2011). Ces méthylations sont distribuées entre les régions riches et pauvres en CpG. Chez l'homme, on estime que près de 70% des promoteurs des gènes sont associés à ces régions riches en CpG, appelées aussi les ilots CpG (Bird, 1986). La méthylation de ces ilots répond à des schémas spécifiques pour chaque tissu ou type cellulaire (Ehrlich et al., 1982). En général, la méthylation des ilots CpG est considérée comme une activité répressive durable. On peut citer par exemple le mécanisme mis en jeu lors de l'inactivation des chromosomes X ou encore la régulation de l'expression allélique des gènes (Edwards and Ferguson-Smith, 2007). Hors les ilots CpG, la méthylation de l'ADN participe à la stabilisation de la structure de l'ADN et aussi à la régulation répressive de la transcription (Bird, 2002).

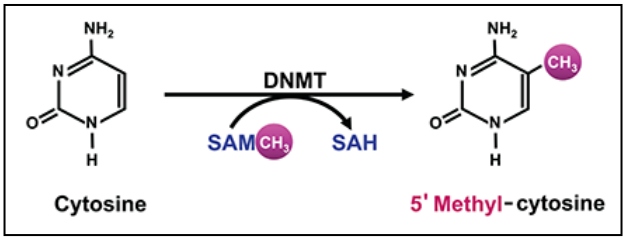



**Fig. 12 : Méthylation du carbone C5 du résidu cytosine par une DNMT (*DNA methyltransferase*).**

Plusieurs modèles tentent d'expliquer le mécanisme de la répression par la méthylation des promoteurs. Il est admis que certains facteurs deviennent incapables de se lier à leurs sites de fixation sur l'ADN après la méthylation de ces derniers, en particulier les ilots CpG. Inversement, elle peut créer une forte compétition pour ces sites de fixation avec des nouveaux facteurs inhibiteurs, et ainsi, empêcher l'activation de la machinerie transcriptionnelle par privation de certains éléments clés du complexe d'initiation de la transcription. Il est suggéré aussi que ces facteurs recrutent des enzymes responsables du remaniement local de la chromatine et donc provoquent un environnement défavorable à l'enclenchement de la transcription (Klose and Bird, 2006).

## 3.2. Les modifications post-transcriptionnelles des histones :

Les mécanismes de compaction de l'ADN et la dualité hétérochromatine/euchromatine sont des modèles élégants qui, en se focalisant sur la structure, permettent d'expliquer l'accessibilité de l'information au sein du noyau. Toutefois, ce modèle structurel souffre d'un manque d'arguments au niveau moléculaire et en particulier au niveau la régulation transcriptionnelle.

Comme vu dans les sections précédentes, les histones ont des structures globalement conservées mais exhibent certaines différences, notamment au niveau des queues N- et C- terminales qui dépassent de leurs structures. Les résidus composant ces queues sont sujets à de nombreuses modifications post-traductionnelles. Ces nombreuses modifications changent les propriétés physico-chimiques des histones (Allis et al., 2007). Il s'agit de méthylations, acétylations, sumoylations, ubiquitinations, ADP-ribosylations, et phosphorylations des sérines, lysines et arginines. Les différentes combinaisons de ces modifications peuvent être reconnues par des protéines spécifiques et accompagnent la réponse transcriptionnelle qui varie selon la position sur le génome.

Parmi les différentes modifications possibles sur les résidus d'histones, la méthylation et l'acétylation des résidus des queues N-terminales d'histones ont été les premières à être caractérisées et leurs rôles dans la régulation sont donc mieux connus que les autres modifications plus récemment mises en évidence (Wolffe, 1998, Spencer and Davie, 1999).

### a) La méthylation :

Dans les queues N-terminales des histones, deux résidus peuvent être méthylés, l'arginine et la lysine. La méthylation de la lysine peut s'observer sous trois degrés (mono- di- ou tri-méthylés) et cela grâce à l'action des enzymes « Histones lysine méthyl-transférases » (HKMT)(Bannister and Kouzarides, 2011). D'abord considérées comme permanentes, ces modifications sont en réalité réversibles et de nombreuses histones lysine déméthylases' (HKDM) ont peu à peu été identifiées, et se sont révélées

(tout comme les HKMT) très spécifiques quant aux lysines ou au type (degré) de méthylation qu'elles affectent (Tsukada et al., 2006, Whetstine et al., 2006). Bien sûr, Les arginines ont aussi des protéines méthyl-transférases propres : les PRMTs.

**b) <u>L'acétylation :</u>**

L'acétylation des résidus d'histones fait intervenir des groupes de protéines particuliers et spécifiques de cette marque : les histones acétyl-transférases (HATs) et les histones déacétylases (HDACs), respectivement, responsable de l'acétylation ou déacétylation. Touchant également les lysines, cette modification est très dynamique (demi-vie de l'ordre de quelques minutes) (Brown et al., 2000, Allis et al., 2007).

Ces méthylation/acétylation sont associées à une activation de la transcription comme c'est le cas pour l'acétylation des histones H3/H4 et les méthylations des lysines K4 et K36 de l'histone H3 (Pokholok et al., 2005). Ainsi, on retrouve au niveau des régions promotrices les marques H3K4me3, H3K9ac et H3K14ac, tandis que la marque H3K36me3 est présente sur l'ORF des gènes activement transcrits. A l'inverse, la déacétylation des histones et la méthylation des lysines K9 et K27 de l'histone H3 sont associés à une répression de la transcription (Li et al., 2007). Aussi, les marques H3K27me3 et H3K9me3 sont retrouvées exclusivement sur les gènes réprimés.

La variété des modifications, ainsi que leur divers processus fonctionnels, ont conduit à proposer l'hypothèse du « code histone », qui délimiterait des modules chromatiniens sur le génome. Ainsi, des combinatoires particulières ont été attribuées à ces « modules d'information » précédemment décrits sur l'ADN (promoteurs, enhancers…) et à leur état (réprimé/activé, compacté/relâché…).

Une étude devenue référence, (Barski et al., 2007), où les auteurs ont tenté de caractériser ce code d'histone. En effet, ils ont généré des cartes à haute résolution pour la distribution de 20 histones modifications (méthylations de la lysine et de l'arginine), ainsi que les différentes variantes d'histone H2A.Z, l'ARN polymérase II, et la protéine de liaison CTCF à travers le génome humain en utilisant la technologie de séquençage à haut débit. Les analyses ont mis en évidence des modèles typiques de méthylations d'histones spécifique aux promoteurs, insulateurs, silenceurs et aux régions transcrites. Par exemple, les mono-méthylations de H3K27, H3K9, H4K20, H3K79 et H2BK5 sont toutes liées à l'activation des gènes, alors que les tri-méthylations de H3K27, H3K9 et H3K79 sont liées à la répression.

### 3.2.1 Les modèles d'action des modifications d'histone sur la régulation:

D'une part, les modifications des histones peuvent avoir différentes conséquences. Il est probable qu'un changement structurel des histones survienne suite à la modification post-traductionnelle d'un de ses résidus. La charge positive des histones responsable de l'affinité avec l'ADN (chargé négativement) peut être altérée, notamment par l'acétylation des lysines. Ainsi, une étude a démontré un effet cumulatif global de l'acétylation des lysines en conséquence aux mutations indépendantes des lysines de la queue d'histone H4 (Dion et al., 2005). Les auteurs suggèrent une déstabilisation des nucléosomes sur l'ADN et par conséquent une accessibilité accrue à l'ADN, ce modèle permet d'expliquer l'apparente activation transcriptionnelle globalement corrélée avec l'acétylation des histones.

D'une autre part, un grand nombre de protéines est capable de reconnaitre et d'interagir avec les résidus méthylés ou acétylés et exercer divers rôles. La reconnaissance se fait grâce à des modules qui peuvent être communs ou spécifiques à chaque modification (Li et al., 2007, Liu et al., 2012). Ce modèle réduit les modifications d'histones à un système de balisage de l'ADN, responsable de la reconnaissance et du recrutement de facteurs spécifiques sur la structure locale de la chromatine. Cette représentation n'est cependant pas réductrice car elle permet une grande variété de réponses associées à chaque marque, dépendant du complexe protéique qui s'y associe. Par ailleurs, certaines modifications peuvent au contraire défavoriser la liaison des facteurs. Comme l'illustre l'exemple de la protéine HP1 qui n'est plus capable de reconnaitre la tri-méthylation de la lysine 9 de l'histone H3 suite à la phosphorylation de la sérine 10, ce qui perturbe, d'une façon indirecte, la structure de la chromatine (Fischle et al., 2005).

### 3.2.2 Les signatures épigénétiques des modules régulateurs:

De même, une localisation préférentielle se verra associée à chaque marque épigénétique, bien que celle-ci ne soit qu'indicative. En effet, plusieurs marques sont à la fois présentes au niveau des enhancers et des promoteurs, mais dans une moindre mesure (ou inversement). Notamment, la marque H3K4me1 est majoritairement présente au niveau des enhancers mais on la retrouve aussi largement au niveau des promoteurs (Fig. 13). De même, la marque H3K4me3 qui est essentiellement associée aux promoteurs est en fait également présente au niveau des enhancers actifs (Pekowska et al., 2011).

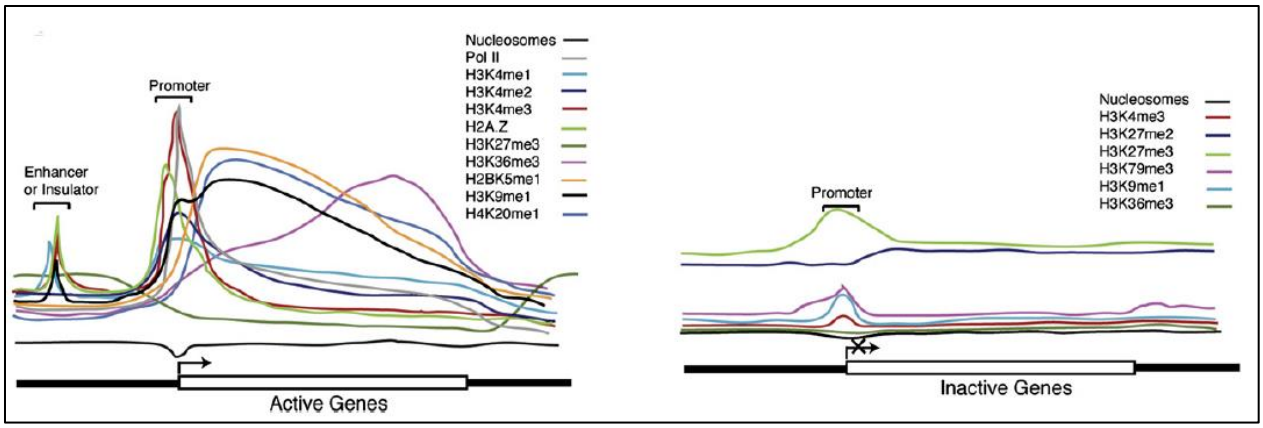



**Fig. 13 : Les signatures caractéristiques des modifications épigénétiques associées aux insulateurs, promoteurs et enhancers dans les gènes inactifs (droite) et actifs (gauche) (Barski et al., 2007).**

Une classification des modifications post-traductionnelles des histones peut ainsi être établie en se basant sur leur effet et leur localisation préférentielle dans le génome ; les plus courantes sont présentées dans le tableau 1.

**Tableau 1. Modifications post-traductionnelles des histones.**

| Modifications | Histones | Sites | Enzymes | Fonctions proposées |
|---|---|---|---|---|
| **Acétylation** | **H3** | Lys9 | Inconnu | dépôt des histones |
| | | | Gcn5, SRC-1 | Activation de la transcription |
| | | Lys27 | Gcn5 | Activation de la transcription |
| **Méthylation** | **H3** | Lys4 (mono,di et tri) | Set1 *(S. cerevisiae)* | euchromatine permissive (di-Me) |
| | | | Set 7/9 (vertebrates) | Activation de la transcription (tri-Me) |
| | | | MLL, ALL-1 | Activation de la transcription |
| | | | Ash1 *(D. melanogaster)* | Activation de la transcription |
| | | Lys9 (Tri) | Suv39h, Clr4 | silence transcriptionnelle (tri-Me) |
| | | | G9a | répression de la transcription empreinte génomique |
| | | | SETDB1 | répression de la transcription (tri-Me) |
| | | | Dim-5 *(N. crassa)*, Kryptonite *(A. thaliana)* | méthylation del'ADN (tri-Me) |
| | | | Ash1 *(D. melanogaster)* | Activation de la transcription |
| | | Lys27 (Tri) | Ezh2 | silence transcriptionnelle |
| | | | | inactivation X (tri-Me) |
| | | | G9a | silence transcriptionnelle |
| | | Lys36 (Tri) | Set2 | Activation de la transcription (élongation) |
| | | Lys79 (Tri) | Dot1 | Euchromatine |
| | | | | Activation de la transcription (élongation) |
| | | | | réponse *checkpoint* |

### 3.2.3 La dynamique des marques épigénétiques dans les enhancers:

Ainsi, les modifications épigénétiques sont des éléments incontournables de la régulation transcriptionnelle en agissant directement sur la structure de la chromatine. Leur combinaison permet en fait de définir de façon très dynamique des régions génomiques actives ou inactives qui vont influer sur l'expression des gènes, comme les promoteurs et les enhancers (Fig. 14). La question se pose alors de savoir s'il existe un lien quantitatif entre le nombre de marques déposées au niveau des éléments régulateurs et leur activité, ou bien si cette activité repose uniquement sur un lien qualitatif basé sur les marques épigénétiques ?

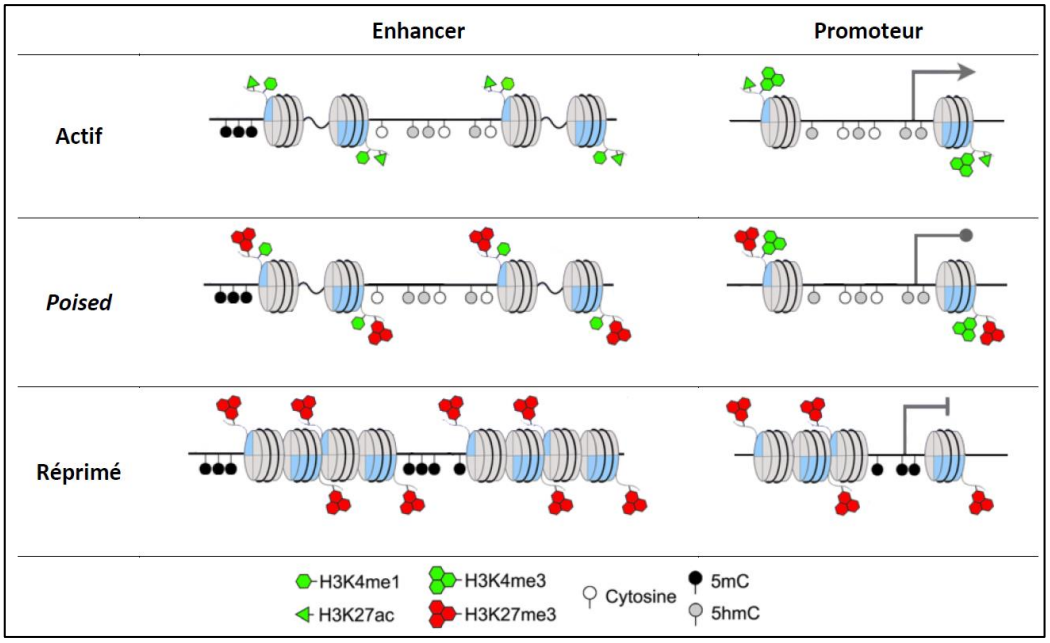



**Fig. 14 : Marques épigénétiques généralement associées aux enhancers (gauche) et promoteurs (droite) en fonction de leur état d'activité (Boland et al., 2014).**

# I.4 - La génération des données à haut débit et les bases de données :

## 4.1. Les techniques et les outils d'analyse:

Le séquençage à haut débit est aujourd'hui devenu incontournable pour l'identification et la caractérisation des transcrits et des éléments régulateurs à l'échelle du génome entier. De nombreuses techniques très diverses ont été développées pour analyser par exemple le transcriptome ou la composition des régions génomiques en modifications épigénétiques, sites de fixation ou encore pour étudier la conformation de la chromatine et ainsi identifier les interactions entre les différents éléments régulateurs.

### 4.1.1 Le ChIP-seq :

#### a) Définition :

Le ChIP-seq (*Chromatin ImmunoPrecipitation followed by sequencing*) est une technique d'analyse des interactions entre les protéines et l'ADN. Elle permet de détecter et d'identifier les séquences d'ADN liées *in vivo*, de façon directe ou indirecte, par une protéine donnée. Cette technique permet donc l'étude des sites de fixation des facteurs de transcription, des cofacteurs, de l'ARN Polymérase II ou bien des modifications d'histone. Le ChIP-seq est en fait une amélioration technique de la technique précédente ChIP-on-chip (*Chromatin ImmunoPrecipitation on chip*). Cette amélioration apporte une plus haute résolution dans la position des sites de fixation identifiés (Rhee and Pugh, 2011).

La ChIP-seq permet d'étudier en théorie tous les sites de fixation d'une protéine à l'échelle du génome entier, et ce de façon non biaisé. La technique de ChIP-seq a l'avantage d'avoir un faible taux de faux négatifs par contre, elle souffre d'un problème de spécificité (Shlyueva et al., 2014). En effet, l'analyse d'une expérience de ChIP-seq identifie généralement plusieurs dizaines de milliers de sites de fixation dans le génome pour une protéine donnée, dont beaucoup d'entre eux semblent être des faux positifs. La détection de ces faux sites pourrait s'expliquer par une trop faible spécificité des anticorps utilisés lors de l'immuno-précipitation ou par l'affinité importante des protéines, notamment des facteurs de transcription, pour l'ADN. Ceux-ci se fixeraient alors plus ou moins aléatoirement et transitoirement sur toute région d'ADN accessible. L'utilisation d'un échantillon contrôle ou de réplicats biologiques permet alors de limiter ce nombre de faux positifs.

Une des utilisations de cette technique est notamment d'identifier les promoteurs et les enhancers dans le génome pour un type cellulaire donné. En pratique, il est possible d'identifier des régions génomiques riches en sites de fixation ou modifications épigénétiques susceptibles d'être des éléments régulateurs. Une autre analyse possible est l'identification des sites de fixation de l'ARN Polymérase II qui permet quant à elle de caractériser les régions transcrites de la chromatine comme les promoteurs, les gènes, ou également les enhancers actifs (les eRNA) (Andersson et al., 2014, Lai and Shiekhattar, 2014).

### b) Le principe expérimental :

Le principe du ChIP-seq repose sur la récupération par immuno-précipitation des fragments d'ADN sur lesquels une protéine d'intérêt est fixée. Ainsi, Le séquençage et l'alignement de ces fragments sur le génome permettent ensuite de localiser les sites de fixation de cette protéine (Fig. 15).

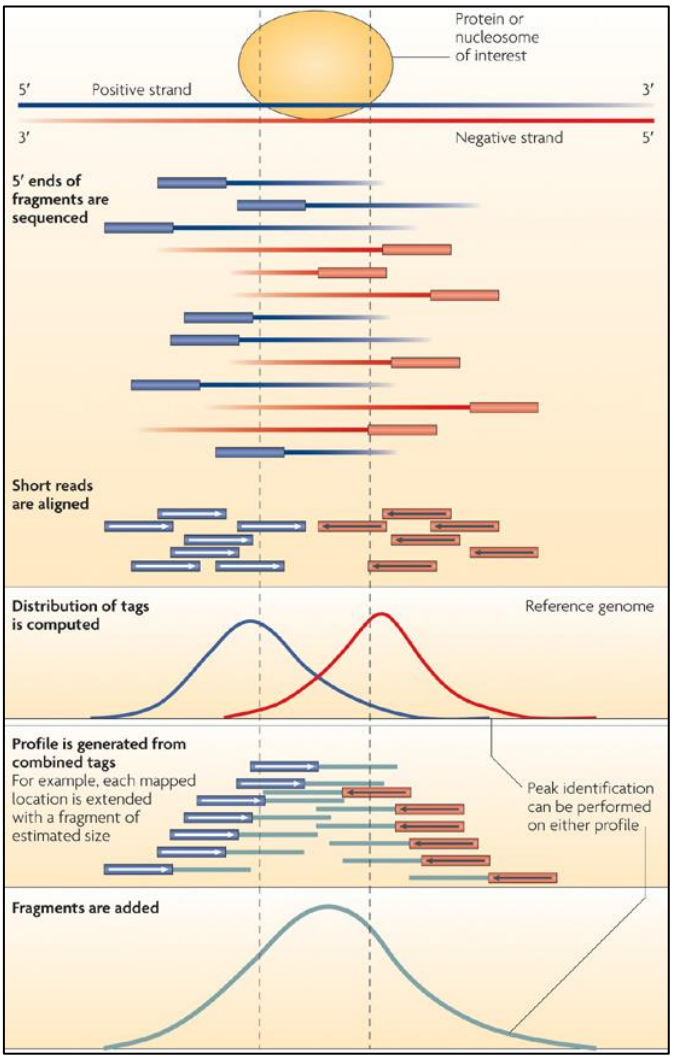



**Fig. 15: Aperçu du protocole expérimental de la technique du ChIP-seq. (Park, 2009)**

En général, le ChIP-seq peut se résumer aux étapes suivantes : (i) liaison des protéines à l'ADN de façon covalente (étape de *cross-linking*) ; (ii) puis une extraction de la fraction nucléaire contenant la chromatine purifié ; (iii) ensuite, la chromatine est fragmentée et immuno-précipitée en utilisant un anticorps spécifique dirigé contre la protéine d'intérêt ; (iv) puis les fragments d'ADN sont récupérés, purifiés, amplifiés par PCR et préparés pour le séquençage. Il faut noter que seules les extrémités des fragments vont être séquencées afin de générer les *reads*.

En parallèle, un autre échantillon dit de contrôle est également réalisé. Celui-ci correspond à une expérience de ChIP-seq classique mais sans l'étape d'immuno-précipitation (*input DNA*), ou dans d'autre situations réalisée avec un anticorps non spécifique. Ainsi, cet ensemble de l'ADN génomique de contrôle va permettre de définir le niveau du bruit de fond et limiter le nombre de faux positifs.

### c) Les analyses bioinformatiques :

Afin de localiser les sites de fixation des protéines, les *reads* (extrémités des fragments d'ADN séquencés) sont tout d'abord alignés sur le génome. Plusieurs outils ont été développés afin de réaliser cet alignement comme BWA (Li and Durbin, 2009) ou bien encore le logiciel commercial ELAND d'*Illumina*. Au cours de ma thèse, j'ai été amené à utiliser deux de ces outils : Bowtie (version 2.0) (Langmead et al., 2009) et BFast (Homer et al., 2009), qui possèdent des caractéristiques complémentaires. Le premier est caractérisé par sa rapidité d'exécution pour aligner les millions de *reads* séquencés sur le génome et par son paramétrage facile et qui permet de définir aisément le nombre de *mismatchs* autorisés (le nombre de variations autorisées entre la séquence d'un *read* et celle du génome de référence) lors de l'alignement. Alors que BFast beaucoup plus lent mais permet d'aligner les reads séquencés en code couleur (*color space*) généré par à la technologie *SOLiD*, alors que ce n'est pas possible avec Bowtie. Ainsi, après alignement, ces deux outils génèrent un fichier contenant la liste des *reads* et leur position dans le génome de référence.

Lorsque l'on visualise globalement la position des *reads* alignés sur le génome, on observe que certaines régions sont enrichies dans l'échantillon par rapport au contrôle (Fig. 16). Cet enrichissement correspond en fait à la localisation des fragments d'ADN séquencés sur lesquels la protéine d'intérêt était fixée. Ainsi, les séquences d'ADN à ces positions génomiques sont supposées contenir la modification épigénétique d'intérêt ou les sites de fixation de la protéine étudiée.

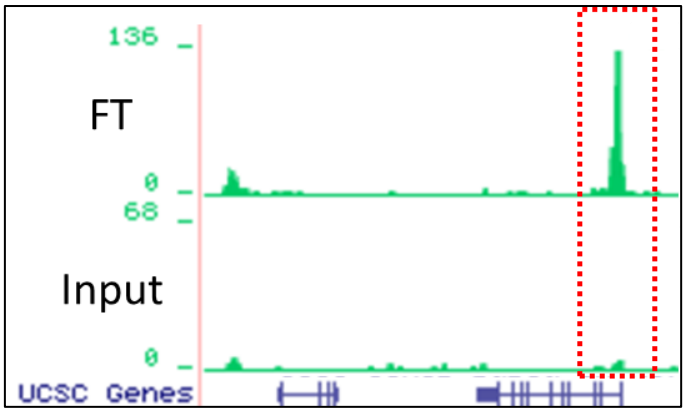



**Figure 16 : Distribution des *reads* alignés sur le génome pour une expérience ChIP-seq et son contrôle. La région enrichie en *reads* par rapport au contrôle est encadrée en rouge.**

Après alignement des *reads* sur le génome, une étape de recherche de pics (*peak calling*), est alors nécessaire. Celle-ci permet dans un premier temps d'identifier les régions enrichies en *reads* dans l'échantillon par rapport au contrôle, puis de déterminer précisément la position des sites de fixation en se basant sur la distribution des *reads* sur les deux brins d'ADN. De très nombreux outils ont été développés afin d'effectuer cette recherche de pics comme MACS (Zhang et al., 2008), SISSRS (Jothi et al., 2008) ou FindPeaks (Fejes et al., 2008). Au cours de ma thèse, j'ai été amené à utiliser l'outil MACS (*Model-based Analysis of ChIP-Seq*).

Finalement, après l'alignement des *reads* sur le génome et la recherche de pics, un fichier contenant les régions génomiques comportant les sites de fixation de la protéine d'intérêt est généré. Un fichier permettant de visualiser les pics sur le génome est également souvent créé. Celui-ci est généré après l'élongation des *reads* alignés sur le génome à la taille des fragments d'ADN, puis par le décompte du nombre de *reads* dans chaque *bin* (région de taille définie, par exemple 50 ou 100 bp) du génome. Une normalisation du signal par le nombre de *reads* alignés est ensuite réalisée. Les pics identifiés se situent typiquement au niveau des éléments régulateurs lors de l'étude d'un facteur de transcription ou d'un cofacteur, alors qu'ils se situent aux alentours des éléments régulateurs lors de l'étude d'une marque épigénétique.

### 4.1.2 Le RNA-seq :

#### a) Définition :

Le RNA-seq (*RNA sequencing*) est une technique qui a été développée afin d'étudier le transcriptome des cellules par séquençage à haut débit. Elle permet de quantifier les ARN dans les cellules.

### b) Le principe expérimental :

Brièvement, dans le cadre du RNA-seq, les ARN sont tout d'abord extraits des cellules, rétrotranscrits en cDNA, amplifiés puis fragmentés avant d'être finalement séquencés. Une sélection des ARN poly-adénylés (poly-A) est généralement réalisée avant l'étape de rétro-transcription afin de sélectionner principalement les transcrits matures. Une autre sélection peut se faire en fonction de la taille des ARN et non de la présence/absence d'une queue poly-A (par exemple short RNA-seq).

### c) Les analyses bioinformatiques :

L'analyse de données RNA-seq peut se faire principalement par l'utilisation de deux outils : TopHat (Kim et al., 2013), qui permet d'aligner les *reads* sur le génome et Cufflinks (Trapnell et al., 2012) qui permet de quantifier les *reads* par transcrit. Cette analyse peut simplement permettre de quantifier le niveau d'expression des gènes connus, mais donne également la possibilité de découvrir de nouveaux transcrits alternatifs ou bien d'annoter le génome avec la découverte de nouveaux gènes (introns et exons). En effet, le RNA-seq permet d'identifier efficacement les jonctions entre les exons dues à l'épissage alternatif. Grâce à cette même caractéristique, la découverte d'ARN fusionnés est également possible par RNA-seq (Tomazou et al., 2015).

## 4.1.3 L'accessibilité de la chromatine:

### a) DNase-seq :

La technique de DNase-seq utilise un principe connu des éléments actifs de l'ADN (promoteurs, éléments cis régulateurs) qui est que ces régions sont plus accessibles à l'enzyme DNase I que le reste du génome. DNase-seq consiste à mesurer la distribution des clivages de la DNase I pour identifier les régions décondensées de l'ADN contenant des éléments régulateurs de la transcription.

L'ADN, une fois extrait des cellules d'intérêt, est mis en présence d'une quantité optimale de DNase I afin qu'elle clive l'ADN aux endroits où celui-ci n'est pas condensé. Les fragments d'ADN obtenus sont ensuite purifiés afin d'éliminer les protéines de liaisons qui peuvent subsister. On a alors à ce stade récupéré toutes les portions d'ADN décondensé du génome. Vient ensuite l'étape du marquage de ces fragments qui va permettre leur amplification par PCR et leur identification, puis le séquençage haut débit. Les séquences obtenues sont enfin alignées sur le génome de référence et leur fréquence est analysée pour identifier les régions régulatrices.

### b) FAIRE-seq :

Le FAIRE-seq est en quelque sorte une version alternative du DNase-seq. Cette technique se base également sur le fait qu'une région de l'ADN non condensée correspond à une région régulatrice de la transcription. Le FAIRE, pour *Formaldehyde-Assisted Isolation of Regulatory Elements*, comme son nom l'indique, utilise le formaldéhyde pour fixer la chromatine et toute autre protéine de liaison *in vivo*. En effet, le formaldéhyde va lier les protéines à l'ADN par des liaisons covalentes, ce qui va en quelque sorte séquestrer cet ADN lors de la purification. On fragmente ensuite l'ADN par sonication (utilisation des ultrasons pour rompre des molécules), et on le purifie avec une solution de phénol-chloroforme. L'ADN non lié à des protéines se retrouve ainsi dans la phase aqueuse du mélange et peut ainsi être prélevé et préparé pour le séquençage, tandis que l'ADN lié aux protéines reste piégé dans la phase organique. Enfin, comme pour le DNase-seq, l'ADN est extrait, marqué, amplifié et séquencé.

Le protocole du FAIRE-seq est plus simple que celui du DNase-seq tout en ayant des résultats tout aussi précis. On notera tout de même que le FAIRE-seq détecte avec une meilleure sensibilité les éléments régulateurs distants mais manque des promoteurs qui sont eux détectés par le DNase-seq.

## 4.2. Les consortiums internationaux:

### 4.2.1 Les projets d'annotation du génome :

Cette dernière décennie, plusieurs consortiums et projets internationaux ont vu le jour dans le but de cataloguer tous les éléments fonctionnels du génome. Qu'il s'agisse de l'analyse de l'état de la chromatine, de l'identification des éléments régulateurs, de la localisation des gènes codants et non-codants. En conséquence, il y a eu une évolution considérable de l'annotation du génome à travers deux consortiums internationaux que sont ENCODE et FANTOM et qui ont largement contribués à ces nouvelles annotations.

#### a) Le projet ENCODE :

Le projet ENCODE (Consortium, 2004), pour *ENCyclopedia Of DNA Elements*, a été lancé en 2003 par le NHGRI (*National Human Genome Research Institute*) afin d'identifier tous les éléments fonctionnels présents dans le génome humain comme la position des gènes codants et non codants, la structure de la chromatine, la localisation des origines de réplication ou bien les éléments régulateurs. Ce consortium a rassemblé de nombreux organismes de recherche universitaires et privés afin de produire les données, les analyser, les stocker et finalement les rendre disponibles publiquement.

Le projet a subi trois phases de développement : une phase pilote, une phase de développement technologique et une phase de production. Les deux premières phases,

réalisées entre 2003 et 2007, consistaient à évaluer et à développer les techniques d'analyse du génome puis à établir les protocoles et les approches qui permettront d'identifier les différents éléments du génome. Cette phase reposait alors sur l'analyse d'une petite portion du génome, environ 1% (30Mb) (Birney et al., 2007). A partir de 2007, le projet est entré en phase de production avec une analyse à l'échelle du génome entier. Les résultats sont parus en 2014 et sont aujourd'hui accessibles publiquement sur de nombreux sites en téléchargement, ou visuellement sur le site de l'*UCSC Genome Browser*. Ce projet a également été étendu aux organismes modèles *Caenorhabditis elegans* (Gerstein et al., 2010) et *Drosophila melanogaster* (mod et al., 2010) via le projet modENCODE, puis à la souris via le projet mouseENCODE (Mouse et al., 2012).

Les approches utilisées par le consortium ENCODE sont nombreuses et variées. En effet, elles vont de l'analyse de l'état de la chromatine (DNaseI-seq, BS-seq par exemple) jusqu'à l'étude du transcriptome des cellules (RNA-seq) en passant par l'identification de sites de fixation de facteurs de transcription (ChIP-seq). Le consortium a également fait en sorte de varier les types cellulaires analysés en se basant sur divers critères comme leur disponibilité et la facilité de les conserver en culture. Une attention particulière a également été portée sur leur caryotype (normal ou cancer) afin de diversifier les études. Cependant, ENCODE recense aujourd'hui principalement des lignées cellulaires cancéreuses et peu de cellules primaires possédant un caryotype normal.

L'achèvement du projet ENCODE et la publication de nombreux articles ont permis de mettre en lumière de très nombreuses caractéristiques du génome humain qui n'étaient jusqu'à présent que partiellement explorées. Les différentes études réalisées à partir des données générées ont notamment mis en évidence la présence d'une association biochimique (marque épigénétique ou transcrit) pour près de 80% du génome et indiqué que 62-75% du génome était transcrit à travers l'ensemble des types cellulaires analysés (Bernstein et al., 2012, Thurman et al., 2012, Neph et al., 2012, Djebali et al., 2012). De plus, ces études ont permis d'identifier environ 500 000 régions régulatrices, 3 millions de sites hypersensibles à la DNaseI et 8,4 millions de sites de fixation de facteurs de transcription dans le génome humain. Enfin, ce projet a également permis d'améliorer considérablement l'annotation du génome à travers le projet GENCODE (Harrow et al., 2012).

### b) Le projet FANTOM

Le projet FANTOM (*Functional ANnoTation Of the Mammalian genome*) (Kawai et al., 2001), débuté en 2000 avant la sortie de la version finale du génome humain, a également pour ambition d'identifier les éléments fonctionnels dans le génome des mammifères. Il s'agit d'un projet d'envergure internationale impliquant environ 500 organismes différents.

Afin d'étudier les éléments du génome, le projet FANTOM se focalise sur l'étude du transcriptome des cellules en utilisant la technique CAGE (*Cap Analysis Gene Expression*) qui consiste à analyser uniquement l'extrémité 5' des ARN. Ce projet repose principalement sur l'étude du transcriptome de cellules primaires ou tissulaires, contrairement au projet ENCODE qui se base essentiellement sur l'analyse de lignées.
Le projet FANTOM fournit aujourd'hui la plus grosse collection de promoteurs identifiés chez l'homme à travers une très grande variété d'échantillons différents analysés (Consortium et al., 2014). Dans sa dernière version rendue publique en 2014, le projet FANTOM indique également pouvoir identifier les *enhancers* actifs dans les différents types cellulaires en étudiant la nature et la quantité des eRNA (*enhancers RNA*) (Andersson et al., 2014).

### 4.2.2 Les projets épigénomiques :

Divers autres projets internationaux ont été mis en place avec le développement des techniques de séquençages adaptées aux modifications épigénétiques. La plupart de ces projets sont coordonnés par l'IHEC (*International Human Epigenome Consortium*) qui s'est donné comme objectif de générer 1000 épigénomes en 7 à 10 ans avec une redondance minimale des types cellulaires analysés entre les différents projets supportés. Parmi ceux-ci, on retrouve les projets internationaux Roadmap et Blueprint, mais également des programmes nationaux de production d'épigénomes comme DEEP (*Deutsches Epigenom Programm*), CREST (*Core Research for Evolutional Science and Technology*) ou encore l'*Epigenomic platform program* de l'Allemagne, du Japon et du Canada respectivement.

#### a) Le projet Roadmap

Le projet Roadmap a été initié en 2008 par l'EDACC (*Epigenomics Data Analysis and Coordination Center*) en collaboration avec le NCBI, et fait intervenir plusieurs organismes américain et canadien (Bernstein et al., 2010). Son principal objectif est d'analyser de nombreuses modifications épigénétiques dans des types cellulaires d'une grande diversité chez l'homme. En effet, les analyses peuvent porter sur l'étude de la méthylation de l'ADN, sur les modifications d'histone ou sur l'accessibilité de la chromatine et peuvent être effectuées sur des tissus cellulaires chez le fœtus ou chez l'adulte, chez des individus sains ou malades. Le programme Roadmap a également pour objectifs annexes le développement de nouvelles technologies d'imagerie pour l'étude des modifications épigénétiques, l'investigation du rôle de l'épigénétique dans les pathologies, ainsi que l'étude de l'effet de l'environnement sur l'épigénome.

Le projet Roadmap propose actuellement la plus grosse collection d'épigénome réalisé chez l'Homme sur des cellules primaires ou des tissus. Par l'intégration de ces très nombreux échantillons, différentes études ont permis de mettre en évidence l'importance de l'épigénétique dans la régulation des gènes, la différenciation cellulaire

et les pathologies (Roadmap Epigenomics et al., 2015). Des études ont également été menées afin de comparer les marques épigénétiques avec des données de méthylation de l'ADN, d'accessibilité de la chromatine ou bien du niveau de conservation des régions régulatrices (Ernst et al., 2011).

### b) Le projet Blueprint

Le projet Blueprint, mis en place en 2011 au niveau européen, se focalise sur les modifications épigénétiques dans le système hématopoïétique. Son objectif est en fait d'étudier la différenciation cellulaire de ce système à travers la production d'une centaine d'épigénomes de cellules primaires, dont huit sont associées à des maladies auto-immunes, et de leur contrepartie leucémique. Pour cela, deux ensembles de modifications épigénétiques sont étudiés. Un premier ensemble consiste à analyser le transcriptome des cellules, la méthylation de l'ADN, l'accessibilité de la chromatine (DNaseI-seq) ainsi que six modifications d'histone dans les cellules. Un deuxième ensemble d'analyses additionnel repose quant à lui sur l'étude de six modifications d'histones supplémentaires et sur l'identification des sites de fixation de divers facteurs de transcription par ChIP-seq.

Finalement, une partie du projet Blueprint consiste à étudier l'impact des variations épigénétiques entre les individus et à étudier la relation qui existe entre la séquence d'ADN et l'épigénétique par l'utilisation de modèles murins.
Actuellement, Blueprint a généré plus de soixante épigénomes complets (six marques épigénétiques, expérience contrôle, BS-seq, RNA-seq) sur sept types cellulaires différents, et près de 28 épigénomes supplémentaires seront bientôt disponibles. Plus de 140 articles utilisant ces données afin d'étudier divers types cellulaires et pathologies ont aujourd'hui été publiés (Saeed et al., 2014, Shin et al., 2014, Chen et al., 2014).

# 5. Les longs ARNs non codant:

Historiquement les ARNm qui codent pour des protéines ont monopolisé nos projets de recherche. Cette dominance est due à leur abondance et la facilité avec

laquelle ils sont identifiés et caractérisés. Le développement rapide, ces dernières années, des nouvelles technologies et en particulier la technologie du séquençage à grande échelle a profondément modifié notre compréhension de l'organisation fonctionnelle du génome et dont la façon nous devrions l'interpréter.

Cependant, durant cette dernière décennie des études du transcriptome grâce à la technologie du RNA-seq ont mis en évidence qu'au moins 70% du génome peut être transcrit (Derrien et al., 2011). Ainsi, l'analyse plus fine de la nature des éléments constituant le transcriptome montre qu'il est composé d'une panoplie d'ARN non codant en addition des gènes codants pour des protéines.

Les ARN non codants peuvent être classés en plusieurs catégories selon leurs fonctions, tailles ou localisation. Ainsi, nous pouvons séparer les ARN non codant en deux importantes classes selon leurs tailles : les longs ARN non codant (LncRNAs) et les petits ARN non codant. Cette dernière classe est relativement bien définie et très bien caractérisée et comporte entre autre les miARN (Micro ARN) et les piRNA (*piwi-interacting RNA*). Tandis que, la classe des LncRNAs est mal explorée. Par contre, le peu des LncRNAs qui ont été étudiés montrent un rôle très important dans plusieurs processus physiologiques ou pathologique, d'où le nombre de publications en constante augmentation lié aux LncRNAs (Fig. 13) (Hu et al., 2012, Ulitsky and Bartel, 2013, Shi et al., 2013, Esteller, 2011, Koerner et al., 2009).

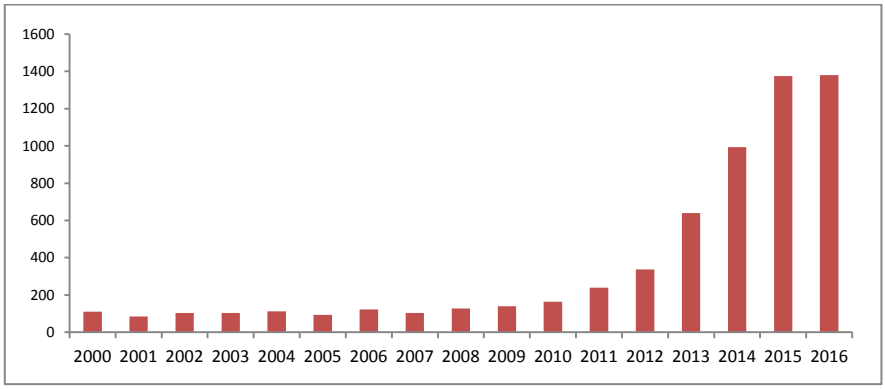



**Fig. 17 : Nombre de publication sur les LncRNAs (Source Pubmed)**

## 5.1. Définition des LncRNAs :

Les LncRNAs sont des transcrits de plus de 200 nucléotides et dépourvus de potentiel codant (Fig. 14A). Ils sont transcrits généralement par l'ARN polymérase II. D'une taille très variable qui peut s'étendre jusqu'à plusieurs dizaines de kilo-bases (Kb) (Fig. 14B) (Ulitsky and Bartel, 2013). Les LncRNAs partagent plusieurs caractéristiques avec les gènes codant pour des protéines, telles que des profils épigénétiques similaires, la présence de signaux d'épissage (Fig. 14C) et de polyadénylation, ainsi que la taille des exons et des introns (Fig. 14D). De multiples LncRNAs contiennent des éléments répétés,

comme des LINE (*long interspersed nuclear elements)* ou des SINE (*short interspersed nuclear elements*), avec des implications fonctionnelles potentielles (Kelley and Rinn, 2012, Kapusta et al., 2013).

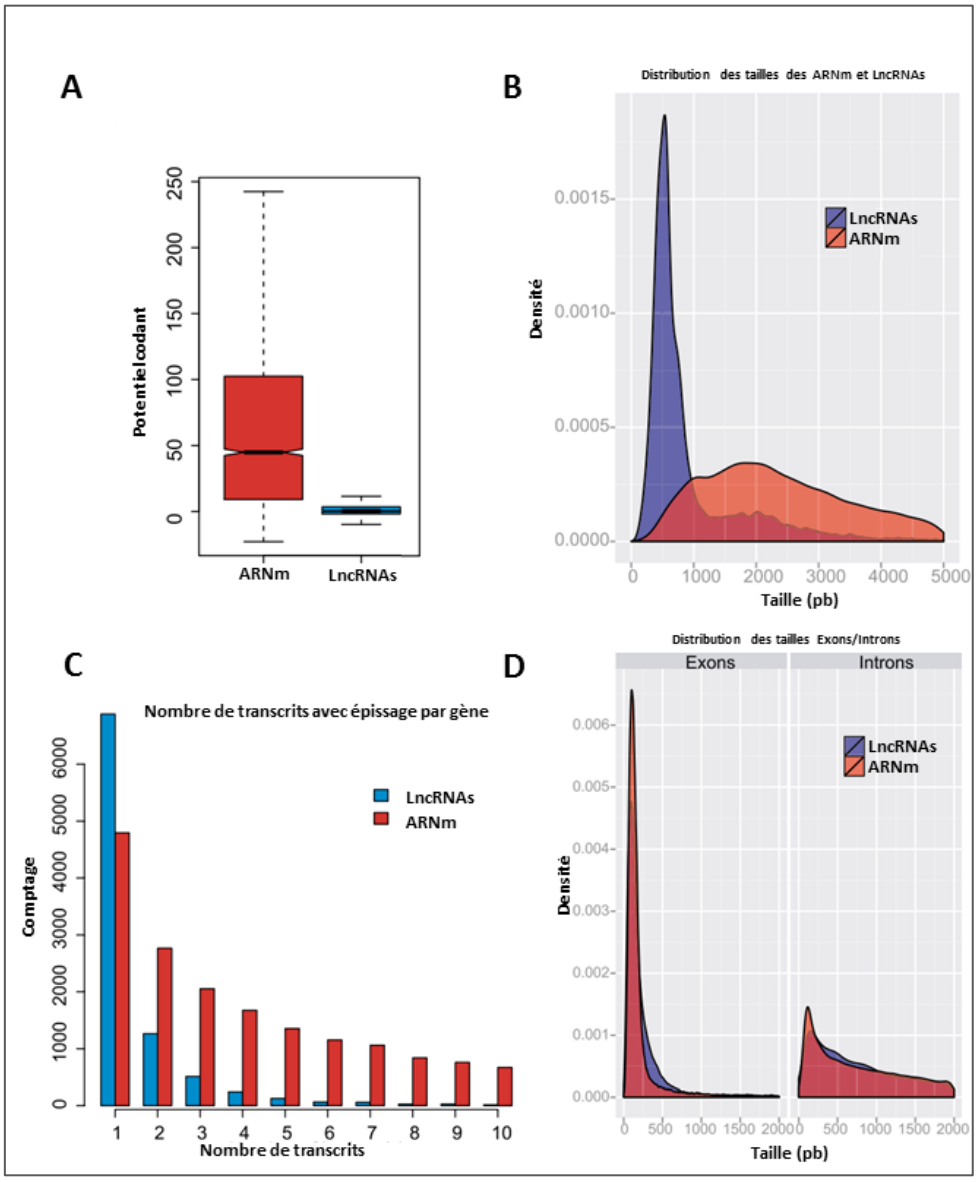



**Fig. 18 : Les caractéristiques des LncRNAs. (A) Le potentiel codant des transcrits. (B) Distribution des tailles des transcrits. (C) Histogramme du nombre de formes d'épissage alternatifs par gène. (D) distribution des tailles des Exons et Introns. Les transcrits codants pour des protéines sont représentés en rouge et les LncRNAs GENCODE v7 en bleu. (Derrien et al., 2012)**

Toutefois, et en comparaison aux ARNm, les LncRNAs sont généralement plus enrichis dans le noyau et montrent une plus faible conservation de séquence (Fig. 15A), bien que certains d'entre eux soient fortement conservés. De plus, les gènes de LncRNAs s'expriment plus faiblement que les gènes codants et leur expression est remarquablement tissu-spécifique (Fig. 15B) (Derrien et al., 2012).

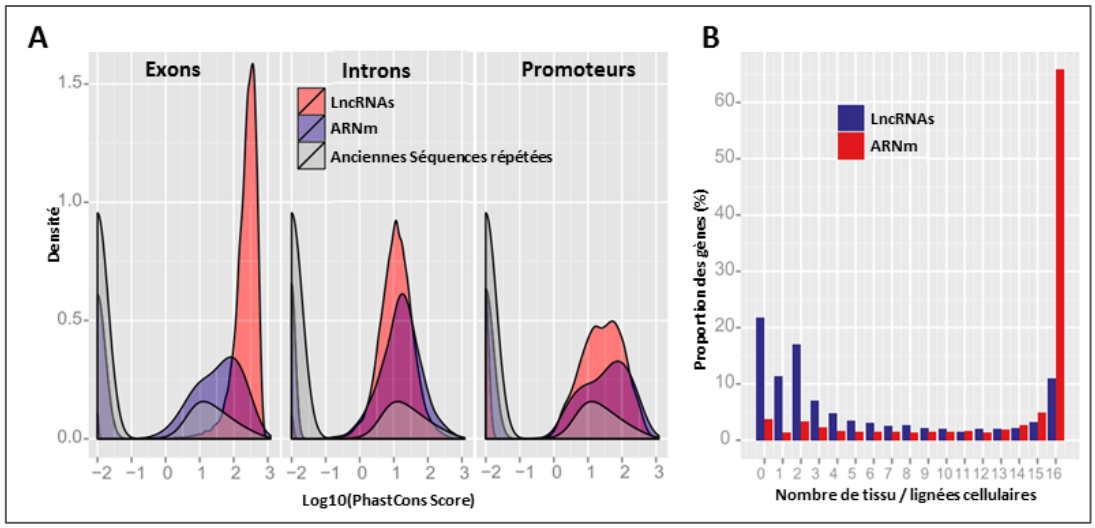



**Fig. 19 : Les niveaux d'expression des tissu-spécificité des LncRNAs. (A) Diagrammes de densité du score de l'outil *phastCons* pour les gènes codant pour des protéines (rouges), les LncRNAs (bleu) et les régions répétées (grise) ; dans les exons (gauche), introns (milieu) et les promoteurs (droite). (B) Histogramme de proportion des transcrits et le nombre de tissu les exprimant. (Derrien et al., 2012)**

## 5.2. Classification des LncRNAs :

Derrien et al en 2012 ont proposé une classification des LncRNAs basée sur leur position par rapport aux gènes codants. Ainsi, les LncRNAs peuvent être divisés en deux grandes catégories : les LncRNAs intergéniques et les LncRNAs intragéniques (Fig. 16). Les LncRNAs intergéniques, localisés par définition dans les régions non annotées du génome, sont généralement dénommés LincRNAs. Quant aux LncRNAs intragéniques, ils peuvent être subdivisés en sous classes en fonction de leur orientation et de la façon dont ils chevauchent les gènes codants (intronique, anti-sens). Il est à noter que de nombreux gènes de LncRNAs intergéniques peuvent être classés dans une sous classe séparée nommée les LncRNAs divergents. En effet, ces LncRNAs ont un site d'initiation de la transcription proche de celui d'un gène codant, la transcription se faisant sur le brin opposé (transcription divergente). Leur caractérisation génomique et fonctionnelle sera discutée en détail dans le chapitre III (voir aussi article II).

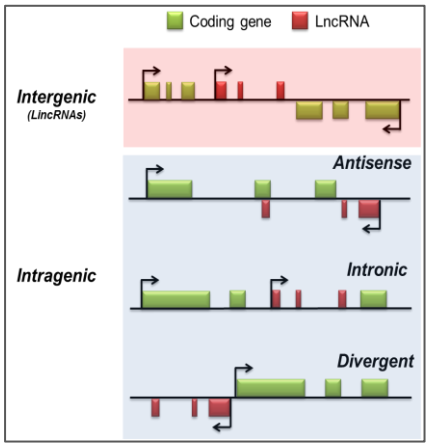



**Fig. 20 : Représentation schématique des principales catégories de LncRNAs selon leur**

**position par rapport aux gènes codant pour des protéines. (Mathieu et al., 2014)**

## 5.3. Bases de données et catalogues des LncRNAs :

Les nouvelles approches à grande échelle, associées à des analyses bio-informatiques poussées, ont permis l'identification d'un très grand nombre de LncRNAs exprimés dans une variété de tissus et de lignées cellulaires. Un des plus importants projets d'annotation du génome, le projet ENCODE (*encyclopedia of DNA elements*), maintient à jour la liste la plus exhaustive de LncRNAs exprimés chez l'homme via la base de donnée GENCODE (http://www.gencodegenes.org) (Derrien et al., 2012). Actuellement, la version V25 de GENCODE compte 15767 LncRNAs détectés chez l'homme, dont 7539 correspondent à des LincRNAs.

Plusieurs autres bases de données publiques regroupent les LncRNAs connus, tels que «*Lncrnadb*» et «*Noncode*» (voir Tableau 2), qui proposent des descriptions détaillées de leurs structures génomiques et leurs fonctions (si elles sont connues).

**Tableau 2 : les bases des données des LncRNAs**

| Base de données | Description | Site web | Références |
|---|---|---|---|
| CHIPBase | Base de données pour le décodage de la régulation transcriptionnelle de l'ARN non codant et des gènes de microRNA à partir de données ChIP-seq. | deepbase.sysu.edu.cn/chipbase | (Yang et al., 2013) |
| DIANA-LncBase | Les cibles microRNA expérimentalement vérifiées ou bioinformatiquement prédites pour LncRNAs. | diana.imis.athena-innovation.org | (Paraskevopoulou et al., 2013) |
| LNCipedia | Une base de données pour les séquences et les structures des LncRNAs transcrits humains annotés. | www.lncipedia.org | (Volders et al., 2012) |
| lncRNAdb | Base de données fournissant une annotation complète des LncRNAs chez les eucaryotes. | www.lncrnadb.org | (Amaral et al., 2011) |
| lncRNADisease | Base de données pour les associations LncRNAs - maladies expérimentalement valisées | cmbi.bjmu.edu.cn/lncrnadisease | (Chen et al., 2013) |
| lncRNAome | Base de données complète des longs ARN non codants chez l'homme. | genome.igib.res.in/lncRNome | (Nohata et al., 2016) |
| Noncode v3.0 | Annotation intégrative des longs ARN non codants. | noncode.org/NONCODERv3 | (Bu et al., 2012) |
| The Functional lncRNA database | Base de données de LncRNAs extrait manuellement de la littérature en plus d'une base de données parallèle contenant tous les ARN humains codant pour des protéines annotées | www.valadkhanlab.org | (Niazi and Valadkhan, 2012) |

## 5.4. Les LncRNAs dans le développement physiologique :

Contrairement aux protéines qui possèdent souvent des domaines fonctionnels bien définis, il est pour l'instant impossible de prédire la fonction des LncRNAs à partir de leur séquence. Toutefois, il apparaît que les LncRNAs agissent essentiellement en modulant l'expression des gènes (Tableau 3) (Yang et al., 2014a).

**Tableau 3. Exemples de LncRNAs impliqués dans des processus biologiques physiologiques :**

| Noms | Fonctions | Mécanismes | Références |
|---|---|---|---|
| HOTAIR | Répression des gènes HOXD. | Recrutement du complexe *POLYCOMB.* | *(Rinn et al., 2007) (Tsai et al., 2010)* |
| XIST | Répression mono-allélique du chromosome X. | Recrutement des complexes répresseurs. | *(Lee, 2011)* |
| HOTTIP et Mistral | Activation des gènes HOXA. | Recrutement du complexe activateur WDR5/MLL. | *(Bertani et al., 2011) (Wang et al., 2011)* |
| Linc-MD1 | Contrôle temporal de la différenciation musculaire. | Séquestration ces miR-133 et miR-135. | *(Cesana et al., 2011)* |
| UCHL1-as | Contrôle de la traduction du gène UCHL1 | Ciblage de l'ARNm vers le ribosome. | *(Carrieri et al., 2012)* |

Cette fonction peut s'exercer localement, lorsque les LncRNAs agissent en cis sur les gènes voisins, ou de façon distale, lorsque leurs fonctions se produisent indépendamment de la localisation des gènes cibles (Fig17).

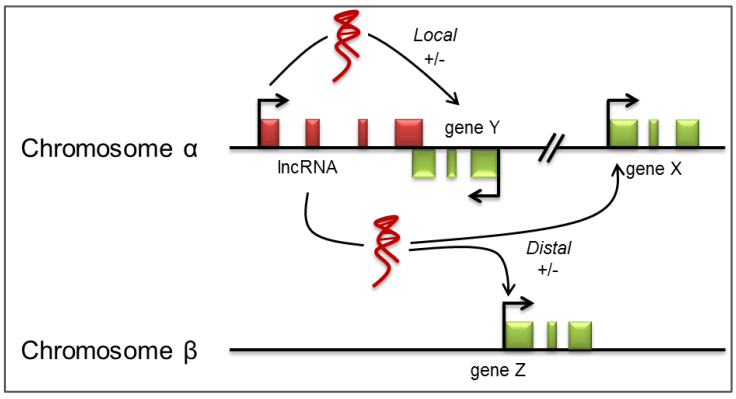



**Fig. 21 : Régulation positive ou négative de l'expression génique par les LncRNAs soit par action locale ou soit par action distale indépendamment de la localisation de leurs gènes cibles. (Mathieu et al., 2014)**

En particulier, il existe une classe de LncRNAs ayant une activité de type « *Enhancer* » et qui peuvent activer de façon transcriptionnelle des gènes voisins (Orom and Shiekhattar, 2013). Plus généralement, des études de gain/perte de fonction ont permis de mettre en évidence que les LncRNAs sont potentiellement impliqués dans divers processus biologiques chez les mammifères (Hu et al., 2012, Yang et al., 2014a). Ces processus incluent, par exemple, le maintien de la pluripotence des cellules souches embryonnaires, la différenciation cellulaire, la régulation du cycle cellulaire et la réponse immunitaire. Les LncRNAs peuvent réguler l'expression des gènes par divers mécanismes. Les aspects moléculaires de ces mécanismes ont été détaillés dans plusieurs revues récentes (Shi et al., 2013). Les LncRNAs peuvent potentiellement lier

l'ADN, des protéines ou d'autres ARN, formant des plateformes d'échafaudage et permettant ainsi l'interaction entre diverses molécules fonctionnelles (Fig. 22).

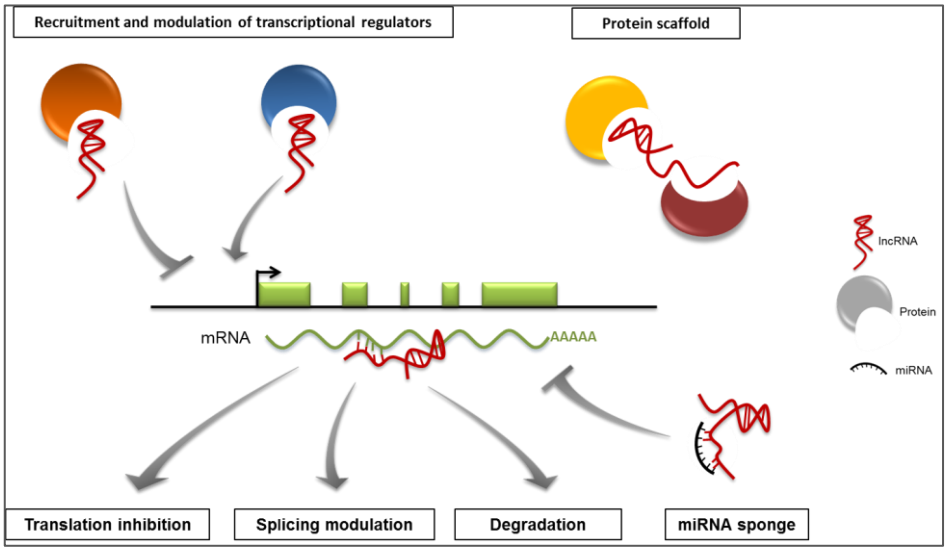



**Fig. 22 : Exemples de mécanismes d'action des LncRNA. (Mathieu et al., 2014)**

Certains LncRNAs peuvent modifier le contexte chromatinien près de leurs gènes cibles, en recrutant des facteurs de transcription ou des acteurs de la modification d'histones ou de la réorganisation de la chromatine, favorisant ou réprimant ainsi la transcription de gènes cibles.

L'un des exemples des LncRNAs le mieux caractérisé fonctionnellement est le gène XIST (*X-inactive specific transcript*). Le LncRNA XIST est localisé sur le chromosome X et transcrit à une taille d'environ 19-kb. XIST est directement impliqué dans l'inactivation du chromosome X chez les mammifères. En effet, XIST recouvre le chromosome X à inactiver et participe au recrutement du complexe répresseur Polycomb 2 (PRC2), permettant ainsi le maintien de l'inactivation de l'un des deux chromosomes X (Lee, 2011).

Il existe d'autres exemples de LncRNAs impliqués dans l'inactivation de l'expression génique, tels que, AIRN (*antisens of IGF2R RNA non-protein coding*), H19 ou KCNQ1OT1, ces LncRNAs assurent le recrutement des complexes répresseurs associés à la chromatine (Orom and Shiekhattar, 2013). Un autre exemple très intéressant est le LncRNAs HOTAIR (*Hox antisense intergenic RNA*), dont le gène est localisé dans le locus HOXC, qui servirait d'échafaudage à la liaison entre les complexes reliés à la répression transcriptionnelle : PRC2 et LSD1 (*lysine specific demethylase 1*)(Rinn et al., 2007, Tsai et al., 2010). Ce complexe favoriserait leur recrutement et ainsi la répression du locus HOXD. Les LncRNAs peuvent participer aussi à l'activation transcriptionnelle. Ainsi, nous avons plusieurs exemples tels que les LncRNAs Mistral et HOTTIP (*HOXA transcript at the distal Tip*) qui faciliteraient l'expression des gènes HOXA en recrutant le complexe

épigénétique activateur WDR5/MLL (*WD repeat-containing protein 5 / mixed lineage leukemia*) (Bertani et al., 2011, Wang et al., 2011).

Les LncRNAs sont aussi largement impliqués dans les processus post-transcriptionnels liés à la biogenèse des ARNm, tels que l'épissage, le transport, la traduction et la dégradation des ARNm (Fig18). Par exemple, UCHL1-as (*ubiquitin carboxyl-terminal hydrolase isozyme L1-antisens*), un LncRNA antisens qui chevauche le gène UCHL1 en extrémité 5', participe activement à la traduction de l'ARNm du gène UCHL1 (Carrieri et al., 2012).

Par ailleurs, Certains LncRNAs chevauchent de petits ARN, tels que des petits ARN nucléolaires (snoRNA) ou des miARN, avec des liens fonctionnels potentiels, comme dans le cas de régions soumises à l'empreinte génomique (Koerner et al., 2009). Les LncRNAs peuvent aussi agir comme des « éponges » pour prévenir la liaison de miARN à leurs ARNm cibles. Par exemple le LncRNA circulaire CDR1-as/ciRS-7 (*circular RNA sponge for miR- 7*) qui possède 70 sites de liaison pour miR-7 (Memczak et al., 2013).

Bien que le mécanisme d'action de nombreux LncRNAs soit maintenant établi, il reste néanmoins possible que l'expression de certains LncRNAs soit plutôt le reflet d'une structure particulière de la chromatine ou d'un sous-produit d'un mécanisme impliquant le processus transcriptionnel. Cela est le cas de l'expression des transcrits germinaux produits au cours du processus de réarrangement somatique des gènes codant les récepteurs pour l'antigène (Spicuglia et al., 2010).

De plus, certains ARN non codants, appelés *enhancer RNA* (eRNA), sont produits à partir d'éléments cis-régulateurs distaux (Orom and Shiekhattar, 2013). Le rôle de ces eRNA sur l'activité transcriptionnelle des gènes associés reste mal caractérisé ou pourraient être tout simplement des sous-produits d'activité de ces locus et sans aucune implication fonctionnelle.

## 5.5. Les LncRNAs dans le développement pathologique :

Etant donné leurs implications dans divers processus biologiques, une altération du niveau d'expression des LncRNAs pourrait mener au développement de plusieurs pathologies, comme les cancers. En effet, de nombreuses études démontrant une dérégulation des LncRNAs dans divers types de cellules cancéreuses suggèrent que les LncRNAs peuvent agir comme des suppresseurs de tumeurs ou des oncogènes potentiels (Shi et al., 2013).

Parmi les LncRNAs cités précédemment, Le LncRNA HOTAIR, qui joue le rôle d'échafaudage dans le complexe de la répression (PRC2 et LSD1), a été associé à

différents cancers. En effet, sa surexpression aurait une contribution dans la formation de métastases (Kogo et al., 2011, Gupta et al., 2010). Un autre exemple de LncRNA dont la surexpression est associée à l'état métastatique est le LncRNA MALAT1 (*metastasis associated lung adenocarcinoma transcript 1*). Son utilisation a été suggérée comme marqueur pronostique dans le cancer du poumon (Shi et al., 2013). De même, des fonctions oncogéniques ont été suggérées pour plusieurs autres LncRNAs (Tableau 4).

**Tableau 4. Exemples de LncRNAs impliqués dans des cancers**

| Types de Cancer | Symboles | Noms complets | Références |
|---|---|---|---|
| Foie | HULC | *highly up-regulated in liver cancer* | Panzitt K. et al. Gastroenterology 2007 |
| Prostate | PCA3 | PCA3, *prostate cancer antigen 3* | Fradet Y. et al. Urology 2004<br>Ferreira LB. BMC cancer 2012 |
| Rein | MVIH | *microvascular invasion in HCC* | Yuan SX. Et al. Hepatology 2012 |

La participation des LncRNAs au développement de pathologies ne se limite cependant pas aux cancers. Le nombre d'études suggérant l'implication de LncRNAs dans le développement de diverses maladies ne cesse d'augmenter (Shi et al., 2013). Ainsi, CDKN2B-as1 (ANRIL), un LncRNAs impliqué dans plusieurs cancers, est aussi associé à l'artériosclérose et à des risques cardiovasculaires accrus (Liu et al., 2009). De la même manière, le développement de diverses pathologies, comme la maladie d'Alzheimer, le diabète néonatal transitoire et la dystrophie facio-scapulo-humérale, pourrait découler de la dérégulation de l'expression d'un LncRNAs.

Enfin, outre la dérégulation de l'expression des LncRNAs, plusieurs études génétiques ont révélé la présence de mutations dans leurs séquences primaires (Shi et al., 2013). Étant donné que la plupart des variants génétiques identifiés par des études d'association génétique sont localisés dans les régions intergéniques, on peut s'attendre à ce qu'un grand nombre de ces mutations affectent des LncRNAs (Derrien et al., 2011), bien que les conséquences de ces mutations sur les fonctions des LncRNAs doivent être démontrées.

## 5.6. Les LncRNAs comme agents thérapeutiques et biomarqueurs:

Puisque plusieurs LncRNAs sont associés au développement pathologique tels que les tumeurs et les dérégulations métaboliques, ils pourraient servir comme des biomarqueurs de dépistage et de pronostic. En effet, des molécules d'ARN et d'ADN stables seraient relâchées dans les fluides corporels comme par exemple le sang par des cellules tumorales en nécrose ou en apoptose. Parmi ces ARN circulants, il y a des LncRNAs dont leur structure secondaire protectrice ou leur présence dans des

microparticules les protège de la dégradation. Ainsi, ils peuvent être détectés dans les fluides comme le sang ou l'urine. Actuellement, quelques LncRNAs sont déjà utilisés dans la clinique pour le diagnostic, comme PCA3 et HULC pour le cancer de la prostate et du carcinome hépatocellulaire (Panzitt et al., 2007, Fradet et al., 2004).

Les LncRNAs peuvent aussi servir comme des cibles thérapeutiques potentielles. En effet, certains essais cliniques en cours utilisent des oligonucléotides anti-sens (ASO) dont l'un de ses rôles est de provoquer la dégradation de leurs LncRNAs cibles ou prévenir la liaison entre le LncRNAs et ces cibles (protéine, ADN ou ARN). Cependant, l'un des obstacles majeur au développement de ces agents thérapeutiques est la distribution et la délivrance ciblée des ASO dans les organismes multicellulaires (Modarresi et al., 2012).

## 5.7. Les stratégies d'étude des LncRNAs :

Contraint par leurs faibles taux d'expression, et leurs tissu-spécificités et le manque de conservation de séquence (Ulitsky and Bartel, 2013), les LncRNAs demeuraient généralement négligés dans les projets de recherche. Mais suite aux avancées technologiques, le catalogue des LncRNAs grandit, et dépasse à présent les gènes codant pour des protéines chez l'homme (Shi et al., 2013). Cependant, seule une petite proportion - mais en croissance constante - est fonctionnellement caractérisée (Esteller, 2011).

Les LncRNAs ont d'abord été considérés comme étant impliqués principalement dans la régulation épigénétique de la transcription (Koerner et al., 2009), probablement en raison de XIST (Lepoivre et al., 2013, Kelley and Rinn, 2012), H19 (Kelley and Rinn, 2012), et HOTAIR (Kapusta et al., 2013) des exemples bien caractérisés. Des exemples plus récents illustrent leur potentiel pour réguler l'expression génique et la traduction protéiques à chaque étape de ces processus.

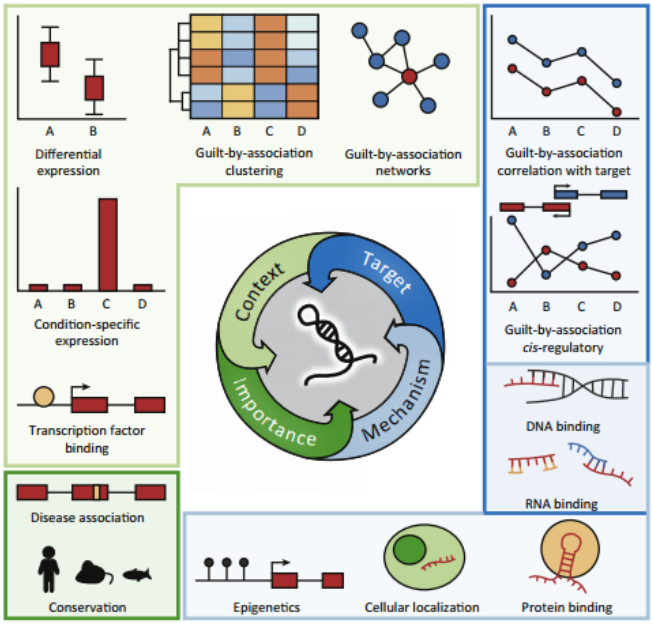



**Fig. 23 : Les approches computationnelles pour la prédiction fonctionnelle et la caractérisation des LncRNAs. Ceux prédisant des informations sur le contexte biologique sont mis en évidence en vert clair, de l'importance biologique en vert foncé, des cibles de la régulation en bleu foncé, et en fin des mécanismes fonctionnels en bleu clair. (Signal et al., 2016)**

Pratiquement, la fonctionnalité d'un LncRNA doit être testée et étudiée individuellement en utilisant des approches expérimentales. Cependant, les procédés classiques tels que le *knock-down*, la surexpression sont souvent pas adaptés à l'analyse de ce qui est typiquement un vaste champ de candidats.

Globalement, les caractéristiques mesurables des LncRNAs sont utiles pour priorisation des candidats pour guider la conception de ces expériences et pour un examen fonctionnel plus approfondi de certains candidats seulement. Récemment, plusieurs méthodes *in silico* ont été développées qui prédisent les rôles, les caractéristiques fonctionnelles, ou l'importance biologique de LncRNAs grâce à ses caractéristiques mesurables.

Parmi les caractéristiques mesurables, hors le niveau de l'expression, il y a les contextes spécifiques de leur expression, leur localisation cellulaires, leurs caractéristiques épigénétiques, les facteurs de transcription qui y sont associés, la capacité de fixation à des cibles (tels que l'ADN, des protéines ou avec d'autre ARNs) et bien sûr la corrélation de l'expression avec les gènes codants. Ainsi, par l'application d'une combinaison de techniques nous pouvons mettre en évidence la fonction ou la classe fonctionnelle d'un LncRNAs. Parmi ces stratégies nous pouvons citer les suivantes :

### a) L'identification des nouveaux transcrits et le potentiel codant :

La recherche des transcrits *De novo* est souvent effectuée avant toutes analyses fonctionnelles de LncRNAs. En raison de leurs profils d'expression très spécifiques et souvent variables, les annotations de LncRNAs sont probablement encore incomplètes dans de nombreux systèmes biologiques (Iyer et al., 2015, Kornienko et al., 2016, Bussotti et al., 2016). Par conséquent, lorsque des données de RNA-seq sont disponibles la recherche *De Novo* peut fournir une annotation plus complète et pertinente de notre modèle expérimental. Le potentiel codant des transcrits devrait être déterminé avant l'analyse fonctionnelle en aval afin d'exclure la possibilité que les nouveaux transcrits puissent coder pour des protéines.

### b) L'expression différentielle

Les profils d'expression des transcrits sont cruciaux pour préciser des hypothèses plus précises des fonctions des LncRNAs. La méthode la plus courante d'inférer une fonction aux LncRNAs dans un système est à travers l'analyse de leur expression différentielle. Bien que cette méthode largement utilisée et généralement acceptée est apte à prioriser les candidats à un examen plus approfondi, l'expression différentielle seule ne produit généralement pas des connaissances fonctionnelles. D'autres méthodes, telles que la culpabilité par association (voir ci-dessous), tire parti des caractéristiques générales de LncRNAs en exploitant d'autres contextes biologiques.

### c) La culpabilité par association

La culpabilité par association est fondée sur l'idée que les transcrits co-exprimés sont plus susceptibles d'être co-régulés, partagent des fonctions similaires, ou sont impliqués dans les processus biologiques similaires (Stuart et al., 2003). Ainsi, la culpabilité par association attribue aux transcrits les mêmes fonctions putatives que ceux avec lesquels sont co-exprimés. Contrairement à l'expression différentielle, les profils d'expression de multiples conditions biologiques connexes peuvent être utilisés, ce qui permet l'identification des relations distinctes entre les transcrits. La confiance de l'association dépend du nombre de conditions pour lesquelles les données d'expression sont disponibles. Les données cinétique peuvent être particulièrement précieuses, car la régulation dynamique d'expression peut être informative des voies particulières dans lesquelles LncRNA exerce (Dinger et al., 2008). La popularité et l'utilité de ce type d'approche ont donné lieu à plusieurs sous-types d'analyses - y compris l'utilisation des relations de cis-régulation et les approches de *clustering*.

Le ***clustering*** à l'échelle du génome des profils d'expression peut être utilisé pour identifier les groupes de transcrits qui sont Co-régulés et montrer l'enrichissement fonctionnel pour un processus. Les LncRNAs dans un groupe sont plus susceptibles

d'être impliqués dans, ou de réguler, les voies biologiques associés aux gènes codants de ce groupe (Stuart et al., 2003). Plusieurs méthodes existent pour le regroupement, chacune avec leurs propres avantages et inconvénients.

Il y a trois méthodes couramment utilisées qui sont (i) la classification hiérarchique (Ramos et al., 2013) ; (ii) k-means (Cabili et al., 2011) et *self-organising maps* (SOMs) (Kim et al., 2015). Toutes ces méthodes nécessitent une certaine forme de sélection du nombre et/ou taille des groupes, donc, des intentions particulières devraient donc être prises dans le choix et le soutien de ces paramètres afin d'assurer que les groupes obtenus sont informatives.

Compte tenu de la complexité des fonctions des LncRNA, plusieurs approches basées sur les réseaux (***network***) ont été utilisées pour les déchiffrer (Xiao et al., 2015, Zhou et al., 2015, Yao et al., 2015). De même pour le regroupement, les modules produits par les réseaux peuvent être utilisés pour assigner des associations fonctionnelles aux LncRNAs. Cependant, parce que les méthodes de mise en réseau peuvent utiliser de multiples couches d'information, en plus de modèles d'expression, des relations plus complexes peuvent être découverts (Guo et al., 2013, Liao et al., 2011, Sun et al., 2014, Yang et al., 2014b, Chen et al., 2015b). La construction du réseau peut être informatiquement difficile, et de nombreux algorithmes et approches statistiques existent à cet effet. Une approche plus simpliste consiste à utiliser une analyse de co-expression pondérée du gène (WGCNA) (Langfelder and Horvath, 2008); Cependant, cette approche est limitée à l'expression du gène seul.

Des groupes de transcrits identifiés par le regroupement doivent être soumis à une étape **d'enrichissement fonctionnel** pour interpréter les processus biologiques de ces gènes. Les voies potentielles et les enrichissements fonctionnels peuvent être déterminés en utilisant les données de *GO* (Ashburner et al., 2000), *KEGG* (Kanehisa et al., 2014), ou *Reactome* (Milacic et al., 2012). Plusieurs outils web et des logiciels sont disponibles qui effectuent des tests statistiques afin de déterminer si l'enrichissement des termes ou des voies particulières dans un groupe de gènes est plus important que prévu. De même, l'analyse d'enrichissement gène-set (GSEA) (Subramanian et al., 2005) peut être utilisée pour évaluer l'enrichissement des processus particuliers à l'aide d'une liste de classement des gènes. Il convient de noter que les prévisions fondées sur ces enrichissements ne sont aussi fiables que les annotations sous-jacentes, et il convient donc de veiller à ce qu'une base de données à jour soit utilisée.

Comme vu auparavant les LncRNAs peuvent agir soit en *cis* ou *trans*. Par conséquent, **la relation génomique** avec d'autres gènes peut être utilisée comme une étape de filtrage initiale utile pour rechercher des relations de régulation potentielles. Dans le cas des transcrits qui fonctionnent en cis, la localisation génomique peut être utilisée comme un guide pour les analyses d'association par culpabilité. Des corrélations positives significatives peuvent indiquer des relations de régulation positives (Orom et

al., 2010). Toutefois, cette corrélation peut vraisemblablement être due aux profils épigénétiques similaires des gènes proximaux (Ebisuya et al., 2008). Dans ce cas, la séparation des *loci* génomiques par des isolateurs ou différents états de la chromatine (Ernst et al., 2011) peut être utilisée comme preuve supplémentaire pour soutenir les relations cis-régulatrices positives. Des prédictions plus fiables peuvent résulter de corrélations négatives parce que ce ne serait pas dû à des états épigénétiques similaires. Pratiquement, une matrice de corrélation constituée de tous les scores de corrélations entre tous les transcrits est produite. Ainsi, des transcrits de fonction inconnue et qui sont significativement corrélés positivement ou négativement peuvent alors être identifiés.

### d) Condition spécifique d'expression

De nombreux LncRNAs montrent des profils d'expression temporaux et spatiaux spécifiques, qui peuvent nous orienter vers le contexte biologique dans lequel ils agissent (Mercer et al., 2008). Ainsi, plusieurs algorithmes sont disponibles pour la détection de l'expression spécifique dans un plus grand nombre de conditions et qui peuvent être utilisés à la place des analyses d'expression différentielle. Dans ce cas, les données d'expression à partir d'un large éventail de conditions - tels que les différents types tissulaires ou cellulaires - peuvent être interrogées. Bien que la méthode simple de comptage des conditions pour lesquelles un gène est exprimé au-dessus d'un niveau particulier puisse être utilisée, des approches plus sophistiquées sont disponibles. Par contre, en comparaison aux gènes codant pour des protéines, les LncRNAs ont des tendances à avoir plus de variabilité d'expression dans un même état physiologique, ce qui peut compliquer énormément l'annotation (Kornienko et al., 2016, St Laurent et al., 2012). En effet, la faible variabilité peut être utilisée comme un indicateur potentiel d'une fonction cellulaire normale, alors que la forte variabilité peut indiquer une fonction très environnement-indépendante et/ou une liaison à une maladie.

### e) Les associations à des maladies

Etant donné que l'expression des LncRNAs est hautement spécifique, il y a un intérêt croissant pour l'utilisation de ces molécules comme des biomarqueurs dans le diagnostic des maladies (Du et al., 2013). Les associations entre une maladie et des LncRNAs peuvent être identifiées par les analyses de l'expression différentielle entre sujet malade et sain. Des centaines d'associations LncRNA-maladie expérimentalement soutenues ont été identifiées et répertoriées au sein de la base de données *lncRNADisease* (Chen et al., 2013). Des SNP identifiés par des études d'association pan-génomiques (GWAS) et d'autres variations génomiques à proximité des LncRNAs associés à des phénotypes spécifiques peuvent également pointer vers des rôles fonctionnels. Ceux qui ont été impliqués avec différents phénotypes ont été catalogués dans de nombreuses bases de données GWAS. Bien que ces approches puissent révéler des biomarqueurs pour une

maladie ou une condition d'intérêt, il est important de noter que l'expression seule ne peut pas être considérée comme une preuve suffisante de la fonctionnalité. En effet, des maladies telles que les cancers peuvent fortement altérer les profils épigénétiques et la transcription en général, ainsi, il se peut que leur expression soit juste le produit de cette perturbation généralisée (Jones and Baylin, 2007).

### f) La conservation

La conservation peut être un outil puissant pour élucider l'importance fonctionnelle de séquences particulières. Bien que les séquences primaires des exons des LncRNAs, sont moins conservées que celle des gènes codant pour des protéines, ils restent plus conservés que les régions intergéniques et intronique aléatoires. Comme les LncRNAs n'ont pas les mêmes contraintes évolutives que les gènes codant pour des protéines (Hezroni et al., 2015, Necsulea et al., 2014), en raison de leur relation structure-fonction tout à fait différente, ainsi, l'utilisation des scores de conservation des LncRNAs devrait être abordée avec prudence. Au niveau de la structure secondaire, la conservation peut également pointer vers l'importance de l'évolution (Smith et al., 2013). Cependant, les relations structure-fonction sont encore à élucider. Les orthologues des LncRNAs peuvent être identifiés en utilisant une approche de synténie et d'alignement grâce à des outils comme BLAST. Je ne comprends pas « la conservation peut également pointer vers l'importance de l'évolution. De telles analyses sont possibles seulement si des données sont disponibles dans d'autres espèces.

### g) La localisation cellulaire

Les LncRNAs peuvent être classés selon le compartiment cellulaire dans lequel ils exercent leur fonction (Zhang et al., 2014). Les ARNs nucléaires sont plus susceptibles d'avoir des rôles dans la régulation de la transcription, la régulation post-transcriptionnelle des ARNm avant l'exportation vers le cytoplasme ou peut être comme un composant des complexes structuraux nucléaires. Cependant, les ARN cytoplasmiques ont plutôt un rôle dans la modification de la stabilité de l'ARNm et la traduction. Ainsi, la prédominance dans un compartiment cellulaire particulier peut être utilisée pour renforcer les hypothèses fonctionnelles (Carrieri et al., 2012). Cependant, ces localisations ne sont pas toujours exclusives, et peuvent être très contexte-dépendante. En effet, la localisation non spécifique ne peut pas être considérée comme une preuve suffisante pour déduire l'absence d'une fonction.

### h) Le statut épigénétique

Les LncRNAs intergéniques peuvent être divisés en *promotor-like* et *enhancer-like* en se basant sur les marques épigénétiques. Les LncRNAs *promotor-like*, ressemblent aux gènes codant pour des protéines. En effet, ils présentent un enrichissement de la marque H3K4me3 au niveau du promoteur et la marque H3K36me3 le long de la région

transcrite. Alors que, les LncRNAs *enhancer-like* arborent un enrichissement de la marque H3K4me1 plus élevé que H3K4me3 dans leur région promotrice. Ces deux sous-classes sont susceptibles d'avoir des fonctions distinctes, les *enhancer-like* affichant une corrélation significativement supérieure avec leurs gènes voisins codant pour des protéines (Marques et al., 2013). Cela donne à penser que l'utilisation des propriétés *enhancer-like* soit biologiquement pertinente et peut être utilisée pour déduire la fonction des LncRNAs. Durant ma thèse je me suis intéressé aussi à un autre aspect épigénétique, à savoir, l'aspect qualitatif de la marque épigénétique H3K4me3 dans la région promotrice et le corps du gène. Ce profil particulier marque les gènes tissu-spécifiques et les régulateurs principaux de la fonction de la cellule où il est exprimé. Nous allons voir dans les chapitres II et III comment nous pouvons utiliser cette propriété afin d'identifier les LncRNAs fonctionnels.

### i) Les interactions des facteurs de transcription avec les loci des LncRNAs

Les régions promotrices des LncRNAs peuvent détenir des indices sur leur régulation. En effet, les sites de liaison des facteurs de transcription (TFBS) dans les régions promotrices de LncRNAs peuvent être utilisés pour trouver les réseaux de régulation dans lesquels ils peuvent être impliqués (Sheik Mohamed et al., 2010). Ces TFBS peuvent être identifiés soit expérimentalement par la technique ChIP-seq ou trouvés dans des bases de données tel que ReMap (Griffon et al., 2015), soit par prédiction en analysant les séquences d'ADN des régions promotrices des LncRNAs par l'utilisation des outils d'enrichissement de motif comme RSAT (Thomas-Chollier et al., 2008), MEME (Bailey et al., 2015) et HOMER (Heinz et al., 2010). Ainsi, les motifs putatifs identifiés peuvent être annotés grâce à des bases de données comme JASPAR qui fournit une liste des motifs de liaison prédéfinie des facteurs de transcription à partir des données expérimentales (Mathelier et al., 2016).

### j) Les interactions LncRNA-Protéine

Identification de protéines qui interagissent avec LncRNAs fournit une autre possibilité de prédire la fonction. Par exemple, XIST, l'un des LncRNAs mieux étudiés, forme une liaison avec le complexe PRC2 et participe à déposer des marques répressives de la chromatine tout le long du chromosome X inactif chez les mammifères (Brockdorff, 2002). Plusieurs autre exemple plus récemment identifiés partagent ceci ou un autre mécanisme fonctionnellement similaire, via le recrutement ciblé de protéines pour réguler l'expression (O'Leary et al., 2015, Liu et al., 2016a, Huang et al., 2015). La régulation post-transcriptionnelle peut également être médiée par des interactions avec des protéines pour réguler la stabilité (Kretz et al., 2013) ou la modification des ARNs (Salameh et al., 2015).

La base de donnée *Npinter* offre un catalogue des liaisons ARN-protéine (Hao et al., 2016), elle utilise des données à haut débit, des données de *text mining* et aussi des

données expérimentale pour soutenir ces interactions. D'autre stratégies utilisant des méthodologies de *machine-learning* (Glazko et al., 2012, Ge et al., 2016, Muppirala et al., 2011) donnent des résultats de plus en plus prometteurs pour la prédiction du potentiel de liaison à des protéines, elles ont été appliquées avec succès dans une série de problèmes génomiques (Libbrecht and Noble, 2015).

### k) Les interactions LncRNA-ADN/ARN

La capacité à se lier à d'autres acides nucléiques d'une manière spécifique est une composante de base de la fonctionnalité des LncRNAs. Les liaisons potentielles aux cibles ADNs ou ARNs peuvent être utilisées pour identifier et renforcer les hypothèses fonctionnelles. En outre, des interactions spécifiques peuvent fournir une hypothèse du mécanisme d'action des LncRNAs. En effet, les liaisons spécifiques ARN/ARN (Gong and Maquat, 2011) pourraient avoir un rôle dans les processus comme l'épissage d'ARN, *RNA editing*, la stabilité, la traduction ou la localisation. Le potentiel de liaison entre un LncRNA et une cible peut être identifié en utilisant des méthodes telles que l'alignement avec BLASTn (Kielbasa et al., 2011). L'inhibition post-transcriptionnelle des ARNm par la liaison des miARN utilise également les liaisons ARN/ARN. Ainsi, les LncRNAs peuvent jouent un rôle de régulateur ou de protecteur en séquestrant ses *miARN.* En conséquence, nous pouvons évaluer le potentiel de liaison LncRNA/miARN grâce à des algorithmes de prédiction.

Des structures complexes peuvent se former par liaison à l'ADN ou à l'ARN, tels que les *triplexes* (Mondal et al., 2015), le *G-quadruplexes* et *R-loops*. En plus de la capacité à recruter des complexes, la formation de ces structures peut intrinsèquement réguler l'expression et la traduction (Buske et al., 2011). Enfin, des évidences pour la formation *R-loops* peuvent être obtenues expérimentalement, à partir de données DRIP-Seq, soit par l'utilisation d'algorithmes de prédiction des sites potentiels (Wongsurawat et al., 2012, Jenjaroenpun et al., 2015).

# 6. La lymphopoïèse T normale :

Les lymphocytes sont des leucocytes qui ont un rôle majeur dans le système immunitaire. En matière de structure et de fonction, on distingue deux lignées lymphocytaires différentes : les lymphocytes B et T. Le développement des lymphocytes T est un processus très bien étudié. En effet, depuis de nombreuses années des études ont bien caractérisé les différents stades de maturation des cellules T, les marqueurs phénotypiques de surface spécifique à chaque stade et les mécanismes de réarrangement des différents locus TCR. En plus, plusieurs études ont très bien décortiqués les points de contrôle de la maturation à travers les différents stades.

De même, l'analyse de l'expression des gènes durant la différenciation a permis également d'explorer de façon très spécifique le fonctionnement de la régulation des gènes. Cependant, le thymus - le foyer de la lymphopoïèse T - est l'un des organes possédant le plus grand nombre d'enhancers tissu-spécifiques (Andersson et al., 2014), ce qui suggère une régulation très fine et précise de l'expression des gènes. Ainsi, la compréhension des mécanismes de régulation durant la différenciation permettra l'identification et la caractérisation des mécanismes de la régulation de l'expression génique de façon plus globale.

## 6.1. L'hématopoïèse :

L'hématopoïèse est le processus assurant la production, le développement et la maturation des cellules hématopoïétiques du sang. Durant la vie fœtale, l'hématopoïèse se déroule dans le foie puis après la naissance dans la moelle osseuse (MO). Les travaux accumulés ces quarante dernières années ont progressivement mis en évidence l'existence des cellules souches hématopoïétiques (CSH). Les CSH sont au sommet de la hiérarchie du modèle de l'hématopoïèse, ce sont des cellules rares multipotentes et qui possèdent la capacité d'auto-renouvellement (**Fig. 24**).

Les progéniteurs multipotents (MPP) sont les descendants directs des CSH. Les MPP conservent un potentiel de reconstitution multi-lignée mais uniquement à court-terme (Adolfsson et al., 2001). Ensuite, ces MPP se différencient en LMPP (*lymphoid-primed multipotent progenitors*) ou ELP (*early lymphoid progenitors*) qui sont des populations progénitrices ayant perdu leur potentiel érythro-mégacaryocytique et exprimant des gènes lymphoïdes spécifiques (Kondo et al., 1997). Ces cellules contiennent les précurseurs T colonisant le thymus TSP (*thymic settling progenitors*). Puis après les LMPP/ELP, on distingue les progéniteurs lymphoïdes communs (CLP) qui présentent, entre autre, un potentiel lymphoïde (**Fig. 24**).

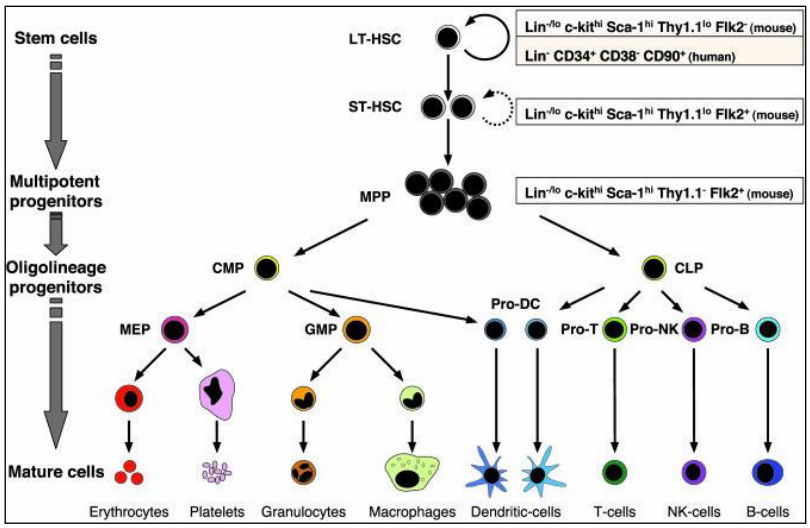



**Fig. 24 : Les différentes lignées de l'hématopoïèse et leurs progéniteurs. (Passegue et al., 2003)**

## 6.2. La thymopoïèse :

La thymopoïèse (ou lymphopoïèse T) désigne les étapes par lesquelles un précurseur hématopoïétique conduit à la production d'un lymphocyte T mature et fonctionnel. Ces étapes sont initiées et soutenues par un environnement unique dans un organe dédié, le thymus.

### a) Thymus:

Le thymus est un organe bilobé situé dans le médiastin antéro-supérieur et antéro-moyen au-dessus du cœur et derrière le sternum qui se divise en trois zones : la capsule, le cortex, et la médulla. Chez les mammifères, le thymus est l'organe spécialisé dans la différenciation lymphocytaire T (Miller and Mitchell, 1967). La partie externe du cortex contient des cellules épithéliales, appelées cellules nourricières, qui secrètent de nombreuses cytokines assurant l'adhésion, la survie et la prolifération des précurseurs. Le cortex représente 85 % du thymus et contient la majorité des thymocytes immatures qui prolifèrent activement. La médulla contient les thymocytes plus matures qui prolifèrent peu. Les cellules matures compétentes ressortent au niveau des vaisseaux sanguins à la jonction cortico-médullaire où se côtoient précurseurs et cellules matures. Les cellules matures rejoignent les ganglions et la rate par les vaisseaux sanguins.

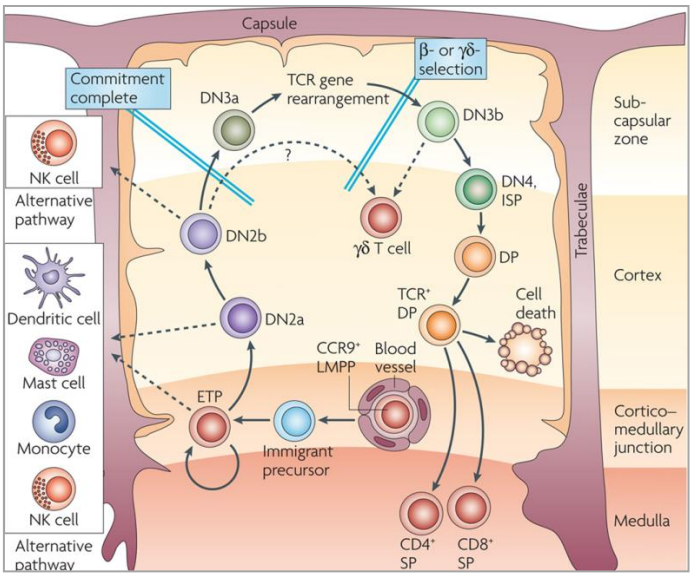



**Fig. 25 : Etapes de la maturation lymphoïde selon les compartiments thymiques (Love and Bhandoola, 2011).**

Chez la Souris, l'entrée des cellules dans le thymus s'effectue en deux phases : les CSH colonisent les niches pendant une semaine suivie d'une période réfractaire de trois semaines (Foss et al., 2001). La colonisation thymique est donc finement régulée par la disponibilité des niches (Goldschneider, 2006). Le thymus a besoin d'être colonisé en permanence car les cellules qui y rentrent ne s'auto-renouvellent pas (Fig. 25).

Chez l'Homme, l'ébauche thymique est capable de recevoir ces précurseurs à la 7ème semaine de gestation et le développement T y est détectable dès la 8ème semaine. A ce stade, les précurseurs thymiques proviennent du foie fœtal (Gale, 1987). A la 16ème semaine de gestation, les précurseurs du foie fœtal colonisent la MO et qui deviennent source de précurseurs thymiques à partir de la 22ème semaine et durant toute la vie post natale (Haynes and Heinly, 1995).

### b) La différenciation lymphocytaire thymique :

Le système immunitaire se caractérise par sa capacité à reconnaître spécifiquement une grande diversité de motifs protéiques (antigènes) et à les combattre lorsqu'ils n'appartiennent pas au soi. Cette reconnaissance se fait grâce aux récepteurs présents sur les lymphocytes: le récepteur des lymphocytes T (TCR) et le récepteur des lymphocytes B (BCR). Les lymphocytes T sont les médiateurs de l'immunité cellulaire. Spécialisés dans la défense contre les bactéries et les infections virales, ils médient également la surveillance immune contre les cellules tumorales et agissent contre les tissus étrangers.

La différenciation lymphocytaire thymique résulte en la génération des différentes sous-populations lymphocytaires fonctionnelles. Ces étapes se caractérisent à la fois par

des marqueurs phénotypiques membranaires et cytoplasmiques bien définis ainsi que par les réarrangements des loci des TCRs. Les réarrangements des différents loci du TCR s'effectuent dans un ordre précis lors de la maturation thymique : premièrement le locus δ suivi du γ et β et du locus α (Blom et al., 1998). Ainsi, chez l'homme en fonction de l'expression ou non des marqueurs et des réarrangements, différents stades sont définis (Galy et al., 1993): le stade Triple Négatif CD3-CD4-CD8- (TN), le Stade Immature Simple Positif (ISP), le stade Early Double Positive CD3-CD4+CD8α+β- (EDP), le stade Double Positif CD4+CD8α+β+ (DP) et le stade Simple Positif mature CD4+ ou CD8+ (SP4 ou SP8).

Par contre, il faut noter l'existence des différences entre la différenciation T humaine et murine, portant notamment sur l'expression de marqueurs de surface ou sur l'ontogénie (Fig. 26).

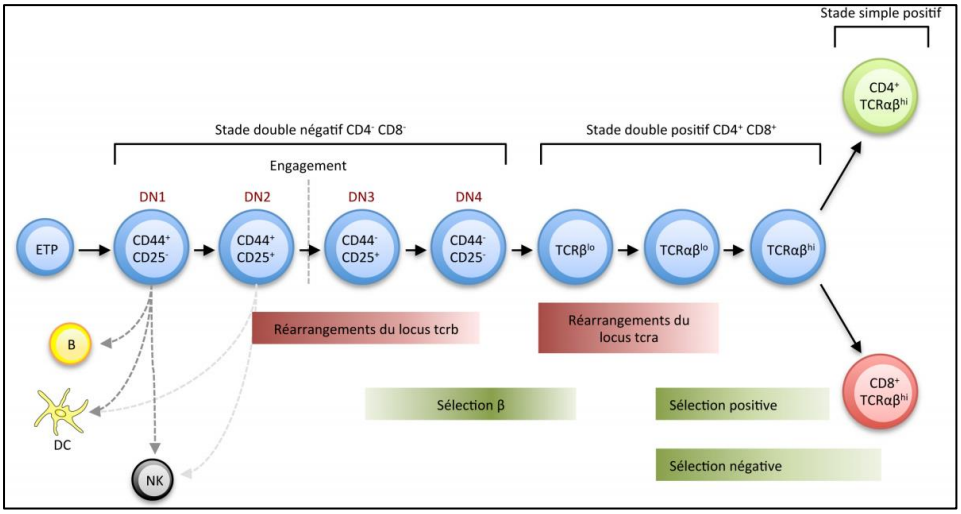



**Fig. 26 : Schéma simplifié du développement murin des lymphocytes T dans le thymus. Un progéniteur hématopoïétique précoce(ETP) entre dans le thymus est subit des réarrangements génétiques dans le but d'exprimer un TCR fonctionnel. Après plusieurs étapes de sélection et de maturation, un lymphocyte T CD4+ ou CD8+ simple positif sort du thymus pour coloniser les organes lymphoïdes périphériques. Adapté de (Sebzda et al., 1999).**

En effet, les thymocytes murins immatures DN peuvent être décomposés en 4 stades selon l'expression des marqueurs CD44 et CD25 (Godfrey et al., 1993) (Fig. 26 et 27), de DN1 (CD44+, CD25-), DN2 (CD44+CD25+) et DN3 (CD44-CD25+) à DN4 (CD44-CD25-). Chez l'homme, les DN sont caractérisés par des marqueurs différents (CD34, CD38 et CD1a) (Dik et al., 2005, Spits, 2002). Alors que les thymocytes murins continuent leur maturation en CD8 ISP pour devenir DP (Paterson et al., 1987). Les thymocytes humains passent par une étape CD4ISP après l'acquisition du CD1a (Galy et al., 1993). Avant l'expression du CD3 de surface, les thymocytes humains DP passent par 2 stades : EDP CD3-CD4+CD8α+β- (Hori et al., 1991) puis DP CD3-CD4+CD8α+β+. La durée de la β-sélection est aussi différente : chez la souris, elle est plus précoce et

synchronisée à la transition DN3 (CD44-CD25+) à DN4 (CD44-CD25-) (Dudley et al., 1994, Mallick et al., 1993) tandis que chez l'homme la durée est plus étendue allant du stade 4ISP au EDP. Ensuite pour les 2 espèces, la β-sélection entraîne le réarrangement du TCRα et l'expression d'un TCRαβ de surface (Fehling and Von Boehmer, 1997) mais avec apparition de CD27 chez la souris et non de CD28 (stade DN3b) (Taghon et al., 2006). Puis les étapes de sélections négative et positive permettent d'obtenir des lymphocytes T matures CD4 et CD8 SP (Singer et al., 2004).

Il existe également des différences dans les réarrangements des loci des TCR (Fig. 27). Chez la souris, ils débutent au stade DN2 par les réarrangements du TCRδ, puis γ, et enfin β qui finissent au stade DN3 alors que chez l'homme une fraction des réarrangements complets du TCRβ ne se produit pas avant le stade DP. Ainsi la détermination de la lignée γδ se fait beaucoup plus tôt chez la souris (Joachims et al., 2006).

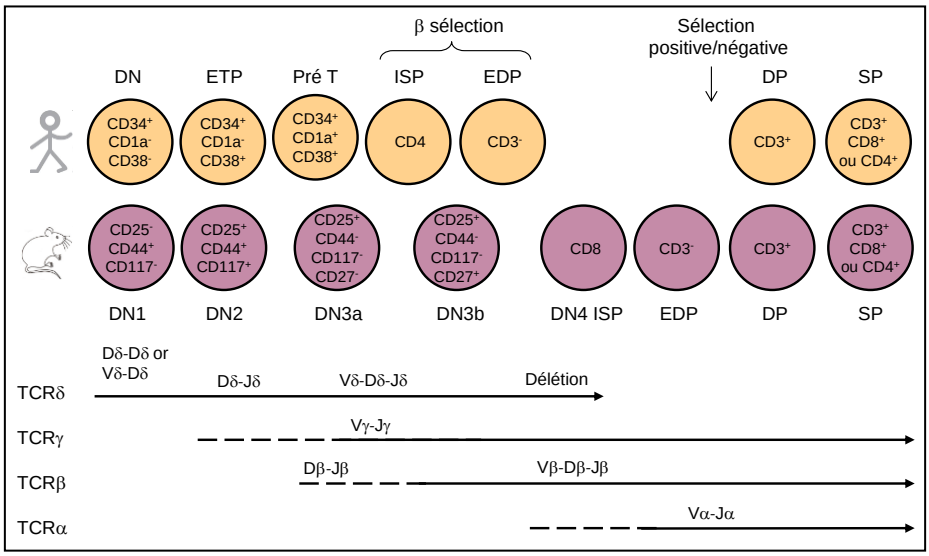



**Fig. 27 : Principales différences phénotypiques et génotypiques entre l'Homme et la Souris.**

## 6.3. La régulation épigénétique des réarrangements des TCRs:

Durant la sélection des thymocytes dans le thymus, les cellules T réarrangent les locus des chaines α et β, permettant l'expression d'un TCR spécifique pour un antigène. Ce réarrangement génique est catalysé par un complexe enzymatique appelé recombinase V(D)J. Ce complexe est composé des protéines RAG1, RAG2 et de protéines impliquées dans la réparation de l'ADN ainsi que de l'ADN ligase IV, chacune ayant un rôle bien spécifique dans la recombinaison V(D)J. Ces enzymes réorganisent tout d'abord les segments VDJ de la chaine β puis les segments VJ de la chaine α. Les locus

réarrangés vont être transcrits puis traduits en chaines α et β pour composer le TCR. (Bassing et al., 2002).

Une régulation stricte de ce processus garantit la spécificité des lignées T, dicte l'ordre temporel des réarrangements TCR et permet l'exclusion allélique. Les spécificités des lignées et l'organisation temporelles de la recombinaison V(D)J sont réglées principalement par l'accessibilité chromosomique. Au cours de la dernière décennie, les gènes des récepteurs antigénique (RA) ont servi de modèles capables d'étudier la régulation de l'expression des gènes avec des loci génomiques complexes (Bonnet et al., 2009, Perlot and Alt, 2008). Ces études ont notamment conduit à une meilleure appréciation de la fonction hiérarchique des éléments cis-régulateurs (les enhancers et les promoteurs germinaux) et de leur impact sur le contrôle de la recombinaison V(D)J grâce à la modulation de la structure de la chromatine (Thomas et al., 2009).

Dans un locus RA donné, l'accessibilité chromosomique de la recombinase V(D)J est généralement corrélée à la présence de plusieurs marqueurs épigénétiques signe d'une euchromatine (comme la déméthylation des CpG et l'enrichissement pour les marques d'histones actives) et à une délocalisation de ces RA à l'extérieur des régions hétérochromatiques du noyau (Jhunjhunwala et al., 2009).

Bien que nous sachions que la présence de marques épigénétiques actives corrélée avec la recombinaison V(D)J, son rôle dans le ciblage de la recombinase sur les loci demeurait incertain (Jhunjhunwala et al., 2009). Une des réponses est venue de l'étude qui montre que le doigt PHD (*plant homeodomain*) de RAG2 se lie à l'histone H3K4me2 ou me3, avec une préférence pour H3K4me3 (Liu et al., 2007, Matthews et al., 2007, Ramon-Maiques et al., 2007).

Dans une autre étude, Lieber et al. (Shimazaki et al., 2009) ont rapporté que H3K4me3 stimule le *cis*-clivage de l'ADN médié par l'enzyme RAG in vitro, un effet qui a également été obtenu en ajoutant simplement le peptide H3K4me3 en *trans*. Ainsi, ceci soutient la notion d'un impact direct de l'environnement épigénétique sur l'activité catalytique de la recombinase. D'un autre côté, les auteurs ont également observé que les signaux de recombinaison cryptique les plus communs (RSS), connus pour être utilisés de façon aberrante lors de la transformation leucémique T, sont localisés à proximité des domaines enrichis en H3K4me3 chez les cellules T normales. Ceci pourrait relier les translocations des RSS avec le paysage épigénétique dans les cellules subissant une recombinaison V(D)J.

La liaison RAG2-H3K4me3 a soulevé la possibilité qu'un code histone spécifique puisse contrôler la recombinaison V(D)J *in vivo*. Cependant, la marque H3K4me3 ne se limite pas aux loci RA, et à ce jour aucune combinaison particulière de marques d'histones n'a été exclusivement associée à ces loci (Morshead et al., 2003, Xu and Feeney, 2009).

Bien que les modifications de la chromatine représentent le processus majeur de régulation de la recombinaison V(D)J, le complexe RAG ainsi que les RSS peuvent influencer l'efficacité des réarrangements. En effet, la protéine RAG2, par son domaine C-terminal, lie les histones et favorise ainsi les réarrangements V-DJ et non les réarrangements D-J (West et al., 2005). La séquence des RSS, quant à elle, peut réprimer l'accessibilité en stimulant le recouvrement par la structure nucléosomale (Baumann et al., 2003). Ce résultat suggère que les RSS contrôlent le positionnement des nucléosomes et qu'elles seraient donc directement impliquées dans la régulation de la recombinaison V(D)J.

# 7. Les leucémies aigües lymphoblastiques T:

Les Leucémies Aigues Lymphoblastiques T (LAL-T) sont des hémopathies malignes causées par la prolifération et l'accumulation de cellules T bloquées à un stade donné de leur différenciation. Il s'agit d'une hémopathie rare, représentant 15 % des LAL de l'enfant et 25 % des LAL de l'adulte (Pui et al., 2004), soit environ 150 nouveaux cas/an en France. Leur pronostic a été nettement amélioré grâce aux progrès thérapeutiques récents, mais reste médiocre avec des taux de survie de 50-60 % à 5 ans chez l'adulte et environ 70-80 % chez l'enfant (Huguet et al., 2009, Pui and Evans, 2006, Silverman et al., 2001).

## 7.1. Classification des LALT:

Les LAL-T sont un groupe hétérogène d'hémopathies, tant sur le plan phénotypique, immunologique ou transcriptionnel. On distingue différentes classifications.

### 7.1.1 Classification immunologique:

La première classification des LAL-T est phénotypique. La plus communément utilisée est la classification du GEIL (Groupe Européen de caractérisation Immunologique des Leucémies (Bene et al., 1995)). Cette classification se veut globalement superposable à la maturation phénotypique du thymocyte normal chez l'Homme.

Elle repose sur i) l'identification de marqueurs pan-T CD3 (dont l'expression est nécessaire et suffisante pour définir la lignée T selon la classification WHO2008), habituellement associée à l'expression du CD2, CD5, et/ou CD7, ainsi que sur ii) le démembrement d'après l'expression du TCR, du CD1a et d'autres marqueurs (CD2, CD5, CD8). Les LAL-T exprimant un TCR de surface, sont matures et définissent le stade IV, lui-même subdivisé en deux stades selon qu'il s'agit d'un TCRαβ (stade IVa) ou TCRγδ (stade IVb). Les LAL-T exprimant le CD1a mais n'exprimant pas de TCR de surface sont dites corticales et définissent le stade III. Enfin, les LAL-T n'exprimant ni le TCR de surface, ni le CD1a sont considérées immatures et peuvent correspondre à deux stades distincts : soit le stade I (CD2-, CD5-, CD8-), soit le stade II (CD2+ et/ou CD5+, et/ou CD8+).

### 7.1.2 Classifications immunogénétique:

Les stades d'arrêt de maturation des LAL-T reproduisent fidèlement les différentes étapes de maturation thymique humaine (Asnafi et al., 2003a). De ce fait, l'analyse des réarrangements des loci des TCRs, l'expression de la chaîne du TCRβ en intra-cytoplasmique (cTCRβ) et l'expression du TCRγδ ou αβ à la surface des blastes T ont

permis d'affiner de manière sensible le stade d'arrêt de maturation des blastes. Les LAL-T peuvent ainsi être subdivisées en trois grands groupes : les LAL-T immatures, pré-αβ, et matures (Fig. 28) :

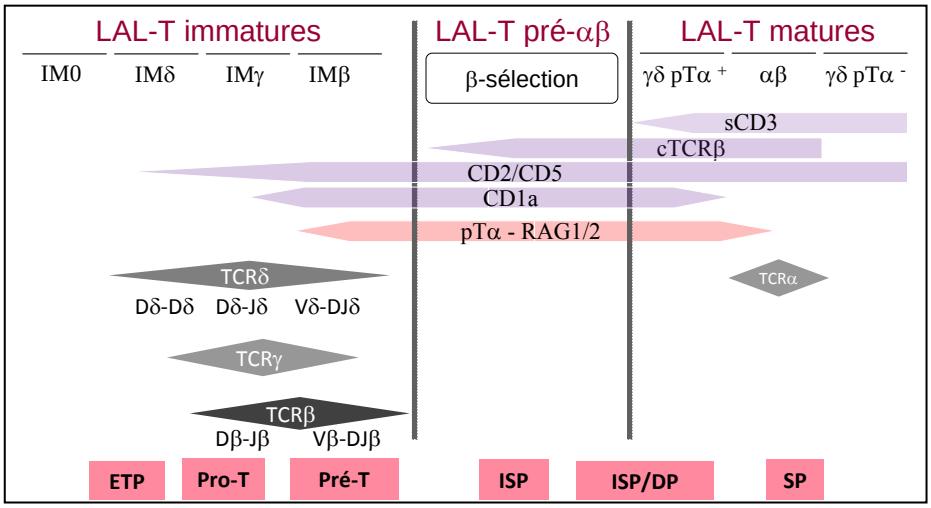



**Fig. 28 : Classification immunogénétique des LAL-T (d'après (Asnafi et al., 2003a)). IM : Immature, ISP : Immature Simple Positif, DP : Double Positif, SP : Simple Positif, pTα : Pré T alpha, TCR : T Cell Receptor.**

### a) LAL-T immatures

Les LAL-T sont généralement de phénotype CD1a-, CD4/8 DN, pTα. Elles expriment fréquemment des marqueurs non-T restreints à la lymphopoïèse (CD34, CD13, CD33, CD56), et ont un niveau bas sinon absent de RAG1, suggérant une origine à partir de précurseurs non-T restreints. Malgré la dérégulation leucémique, les réarrangements des loci du TCR restent parfaitement ordonnés dans les LAL-T, de manière strictement identique à la séquence physiologique. La configuration des différents loci du TCR permet le démembrement des LAL-T immatures selon que les loci sont en configuration germinale (IM0) ou réarrangés, avec d'abord, le TCRδ (IMδ), puis TCRγ (IMγ) et enfin le TCRβ (IMβ). Les stades d'arrêt de maturation observés dans les LAL-T reproduisent ainsi fidèlement les différentes étapes de la maturation thymique humaine.

### b) LAL-T pré-αβ (en cours de β-sélection)

Elles sont caractérisées par l'expression de la chaîne TCRβ en intracytoplasmique (cTCRβ), suggérant un engagement vers la lignée T et probablement Tαβ. Environ 50% des LAL-T de la lignée αβ expriment ainsi RAG1, le pré-TCR, et cTCRβ, et présentent un profil CD4/CD8 DP>SP>DN, sCD3-, CD1a+. De manière caractéristique elles n'ont pas de délétion du locus TCRδ. Ce type de LAL-T correspond à l'expansion clonale de cellules exprimant un pré-TCR, probablement aux stades ISP, DP et SP du développement thymique, et reflète une transformation aux dépens de thymocytes en cours de β−sélection.

### c) LAL-T matures sCD3+

Les LAL-T matures sont isolées à partir de l'expression du sCD3 associée à celle du TCR, soit TCRαβ (la moitié), soit TCRγδ (la moitié). A noter que la fréquence des LAL-T au stade immature est plus importante chez l'adulte (> 20 ans) que chez l'enfant (<10 ans); en effet, pour ces derniers, la moitié des cas sont matures et expriment un TCR de surface. Les LAL-T des adolescents et de l'adulte jeune (11-20 ans) présentent des stades d'arrêt de maturation intermédiaire permettant de visualiser une évolution progressive entre ces deux catégories immatures et matures (Asnafi et al., 2004). L'involution thymique est à l'origine de l'incidence croissante des formes immatures en fonction de l'âge.

## 7.1.3 Classification transcriptomique:

L'oncogenèse des hémopathies T immatures est multigénique, avec coexistence de plusieurs dérégulations oncogéniques chez le même patient (Aifantis et al., 2008a). Une approche moléculaire est donc indispensable à la bonne identification des lésions oncogéniques des LAL-T. Ainsi, plusieurs travaux ont décortiqué les LAL-T à partir d'études pan-génomiques de leurs profils d'expression génique et établi des signatures transcriptionnelles spécifiques des oncogènes dérégulés (Ferrando et al., 2002, Homminga et al., 2011, Soulier et al., 2005). Comme l'illustre très bien le *heatmap* réalisé par Homminga et al. 2011 dont ils observent une bonne séparation entre les différentes classes des LAL-T grâce à une signature d'expression spécifique (Fig. 29).

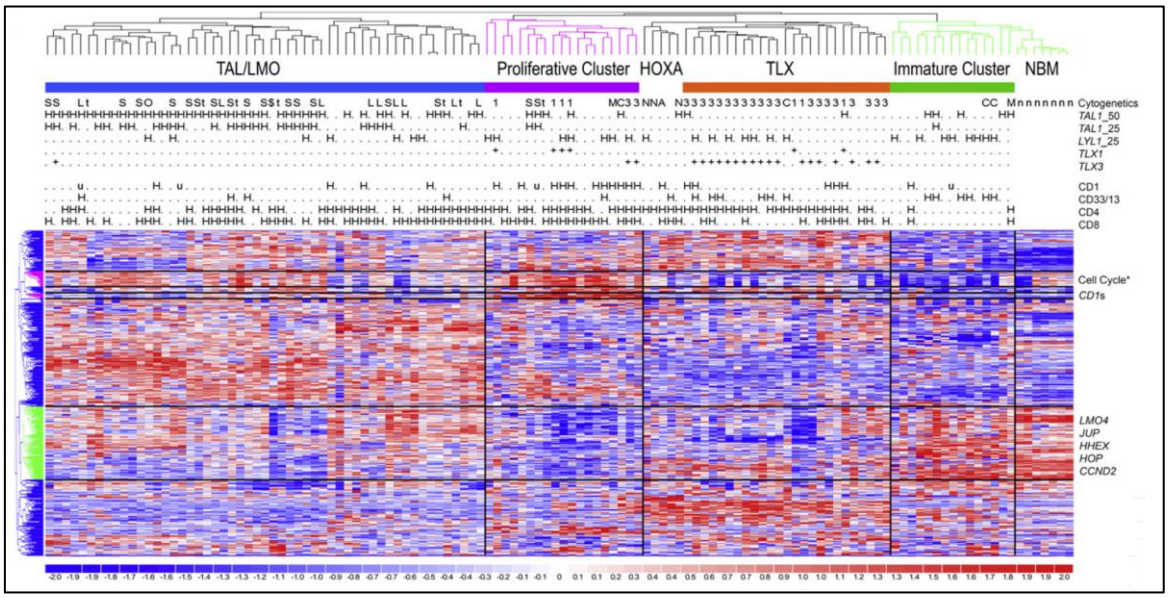



**Fig. 29: Analyses de *clustering* hiérarchiques non supervisées basées sur des données d'expression de transcriptome (*Microarrays)* de 117 échantillons T-ALL pédiatriques et de sept contrôles normaux de moelle osseuse (Homminga et al., 2011).**

Et donc, parmi les sous-groupes identifiés figurent la signature LYL1/MEF2C au stade «double négatif», TLX/HOXA/NKX2.1 au stade «cortical précoce», et TAL1 au stade «cortical tardif». Les LAL-T TLX1 et TLX3 partagent ainsi la même signature transcriptionnelle (Fig. 30).

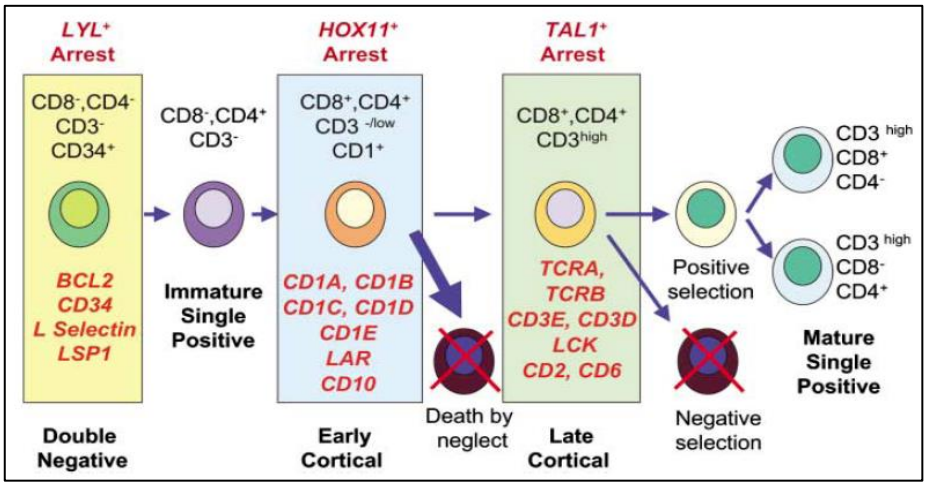



**Fig. 30 : Classification transcriptionnelle des LAL-T. Représentation schématique des clusters transcriptionnels des LAL-T identifiés, superposés au schéma du développement thymique normal. En haut en lettres noires les marqueurs de surface associés aux stades développementaux physiologiques. En bas en lettres rouges les signatures transcriptionnelles caractéristiques des LAL-T, d'après (Ferrando et al., 2002).**

## 7.2. Dérégulation oncogénique dans les LAL-T:

L'oncogenèse des LAL-T est multi-étape, ce qui implique une grande variété d'anomalies moléculaires incluant des translocations, amplifications, délétions, duplications et mutations. Ces anomalies oncogéniques peuvent être classées selon 2 types : les anomalies du groupe A, mutuellement exclusives et responsables du blocage de maturation et les anomalies du groupe B qui peuvent être redondantes avec des anomalies du groupe A et/ou du groupe B (Van Vlierberghe and Ferrando, 2012, Van Vlierberghe et al., 2008).

### 7.2.1 Oncogène de type A:

Ces anomalies sont fondatrices, mutuellement exclusives et impliquées dans le blocage de différenciation. Ces oncogènes déterminent une signature d'expression génique et délimitent des sous-types spécifiques de LAL-T (Ferrando et al., 2002, Soulier et al., 2005, Van Vlierberghe and Ferrando, 2012, Van Vlierberghe et al., 2008).

Les anomalies de type A comprennent entre autres les dérégulations des gènes HOX avec notamment la surexpression de HOXA9, TLX1 (HOX11) et TLX3 (HOX11L), les

dérégulations de LYL1, les anomalies de type CALM-AF10, MLL et SIL-TAL. Ces anomalies sont listées dans le tableau 5 selon (Van Vlierberghe and Ferrando, 2012).

Les anomalies MLL sont associées à un arrêt au stade immature, la surexpression de TLX1 à un arrêt autour de la β-sélection et les microdélétions cryptiques *STIL-TAL1* à un arrêt tardif dans la maturation (Asnafi et al., 2004). Les anomalies CALM-AF10 peuvent quant à elles être associées tantôt à des LAL-T immatures, tantôt à des LAL-T TCRγδ, selon le transcrit qui est engagé (transcrit 5' et 3' respectivement) (Asnafi et al., 2003b, Ben Abdelali et al., 2013).

Les principaux sous-groupes sont :

- *Le sous-groupe TAL/LMO*
- *Le sous-groupe HOXA*
- *Le sous-groupe TLX1 et TLX3*
- *Le sous-groupe c-MYB*

**Tableau 5 : Classification des anomalies génétiques du groupe A dans les LAL-T Adapté de (Van Vlierberghe and Ferrando, 2012, Van Vlierberghe et al., 2008)**

| Catégorie | Gène | Réarrangement | Blocage de maturation | Incidence | Pronostique |
|---|---|---|---|---|---|
| Famille bHLH | ***TAL1*** | t(1;14)(p32;q11) | Pré-αβ/TCRαβ | **3%** | **Bon** |
| | | t(1;7)(p32;q34) | | **3%** | **Bon** |
| | ***STIL/TAL1*** | 1p32 délétion | Pré-αβ/TCRαβ | **16-30%** | **Bon** |
| | ***TAL2*** | t(7;9)(q34;q32) | | **1%** | **Inconnu** |
| | ***LYL1*** | t(7;19)(q34;p13) | | **1%** | **Inconnu** |
| | ***BHLHB1*** | t(14;21)(q11.2;q22) | | **1%** | **Inconnu** |
| Famille LMO | ***LMO1*** | t(11;14)(p15;q11) | | **1%** | **Inconnu** |
| | | t(7;11)(q34;p15) | | **1%** | **Inconnu** |
| | | t(11;14)(p13;q11) | | **6%** | **Inconnu** |
| | ***LMO2*** | t(7;11)(q34;p13) | Immature/ Pré-αβ/TCRαβ | **6%** | **Inconnu** |
| | | 11p13 délétion | | **3%** | **Inconnu** |
| | ***LMO3*** | t(7;12)(q34;p12) | | **<1%** | **Inconnu** |
| Famille Homeobox | ***TLX1*** | t(11;14)(p15;q11 | Immature/ Pré-αβ | **5-30%** | **Bon** |
| | ***TLX3*** | t(11;14)(p15;q11) | Immature/ Pré-αβ/lignéeγδ | **5-20%** | **Mauvais** |
| | ***HOXA*** | Inv(7)(p15q34) | TCRαβ ou TCRγδ | **3%** | **Inconnu** |
| | | t(7;7)(p15;q34 | | **3%** | **Inconnu** |
| | ***HOXA (CALM-AF10)*** | t(10;11)(p13;q14) | | **5-10%** | **Mauvais** |
| | ***HOXA (MLL-ENL)*** | (11;19)(q23;p13) | Immature/lignée γδ | **1%** | **Inconnu** |
| | ***HOXA (SET-NUP214)*** | 9q34 délétion | Immature/lignée γδ | **3%** | **Pas d'impact** |
| | | inv(14)(q11.2q13) | Immature/lignée γδ | | **Inconnu** |
| | ***NKX2.1*** | inv(14)(q13q32.33) | | **5%** | **Inconnu** |
| | | t(7;14)(q34;q13) | | | **Inconnu** |
| | ***NKX2.2*** | t(14;20)(q11;p11) | | **1%** | **Inconnu** |
| **Proto-oncogènes** | ***C-MYB*** | t(6;7)(q23 ;q34) | | **3%** | **Inconnu** |

## 7.2.2 Oncogène de type B:

Les oncogènes de classe B sont partagés par les différents sous-types de LAL-T et synergisent avec les événements de classe A dans le développement leucémique (Van Vlierberghe and Ferrando, 2012). Ce sont généralement des gènes impliqués dans le contrôle du cycle cellulaire, de l'auto-renouvèlement, de l'engagement T, de la

signalisation pré-TCR, de la différenciation T, ou encore l'activation de tyrosines kinases: par exemple, les anomalies de *NOTCH1/FBXW7*, de *CDKN2A/CDKN2B*, de *RAS*, ou de *PTEN*. Le tableau 6 récapitule ces anomalies oncogénétiques.

**Tableau 6 : Classification des anomalies oncogénétique de type B dans les LAL-T Adapté de (Van Vlierberghe and Ferrando, 2012, Van Vlierberghe et al., 2008)**

| Catégorie | Gène | Réarrangement | Incidence | Pronostic |
|---|---|---|---|---|
| Défaut du cycle cellulaire (B1) | ***CDKN2A/2B*** | 9p21deletion méthylation | **70%** | **Bon** |
| | ***CCND2*** | t(7;12)(q34;p13)<br>t(12;14)(p13;q11) | **1%** | |
| | ***RB1*** | 13q14 délétion | **4%** | **Pas d'impact** |
| | ***CDKN1B*** | 12p13 délétion | **2%** | |
| NOTCH activation (B2) | ***NOTCH1*** | t(7;9)(q34;p13) | **<1%** | |
| | | Mutations activatrices | **>60%** | **Bon** |
| | ***FBXW7*** | Mutations inactivatrices | **8-30%** | |
| Signalisation pré-TCR/TCR (B3) | ***LCK*** | t(1;7)(p34;q34) | **<1%** | |
| | ***NRAS*** | Mutations inactivatrices | **5-10%** | **Pas d'impact** |
| | ***NF1*** | Mutations inactivatrices/délétions | **3%** | **Pas d'impact** |
| | ***PTEN*** | Mutations inactivatrices | **10%** | **Pas d'impact** |
| | | Délétions | **10%** | **Mauvais** |
| Dérégulation additionnelle de la différenciation T (B4) | ***MYB*** | Duplication | **8-15%** | |
| Activation d'autres tyrosine kinases (B5)/transduction du signal | ***NUP214-ABL1*** | Episomal 9q34 amplification | **4%** | **Mauvais ou pas d'impact** |
| | ***EML1-ABL1*** | t(9;14)(q34;q32) | **<1%** | |
| | ***ETV6-ABL1*** | t(9;12)(q34;p13) | **<1%** | |
| | ***BCR-ABL1*** | t(9;22)(q34;q11) | **<1%** | **Mauvais** |
| | ***FLT3*** | Mutations | **3%** | |
| | ***JAK1*** | Mutations activatrices | **4-18%** | **Pas d'impact** |
| | ***ETV6-JAK2*** | t(9;12)(p24;p13) | **<1%** | |
| | ***JAK3*** | Mutations activatrices | **5%** | **Pas d'impact** |
| | ***IL7R*** | Mutations activatrices | **10%** | **Pas d'impact** |
| Remodelage de la chromatine | ***EZH2*** | Mutations inactivatrices/délétion | **10-15%** | **Mauvais** |
| | ***EED*** | Mutations inactivatrices/délétion | **10%** | **Pas d'impact** |
| | ***SUZ12*** | Mutations inactivatrices/délétion | **10%** | **Pas d'impact** |
| | ***PHF6*** | Mutations inactivatrices/délétion | **20-40%** | **Pas d'impact** |
| Facteur de transcription, tumeurs suppresseurs | ***MYC*** | t(8;14)(q24;q11) | **1%** | |
| | ***WT1*** | Mutations inactivatrices/délétion | **10%** | |
| | ***LEF1*** | Mutations inactivatrices/délétion | **10-15%** | |
| | ***ETV6*** | Mutations inactivatrices/délétion | **13%** | |
| | ***BCL11B*** | Mutations inactivatrices/délétion | **10%** | |
| | ***RUNX1*** | Mutations inactivatrices/délétion | **10-20%** | |

## 7.3. Dérégulation épigénétique dans les LAL-T:

Aifantis et coll. Ont suggéré, en 2008, que la dérégulation des processus épigénétiques durant la différentiation des cellules T pourrait être à l'origine de la transformation oncogénique des précurseurs de cellules T en cellules cancéreuses (Aifantis et al., 2008b). En particulier, les mutations dites de « type A » qui sont principalement des facteurs de transcription et donc ils impactent directement le transcriptome et l'épigénome en bloquant la différenciation des cellules T à des stages spécifiques de leur développement et définissent des sous-groupes ayant des profils spécifiques d'expression génétique (Ferrando et al., 2002, Homminga et al., 2011). La surexpression des gènes HOX, comme le NKK et l'HOXA, ainsi que la dérégulation des TLX1/3 sont fréquentes dans les mutations de « type A ». Ces mutations mènent à l'arrêt prématuré de la maturation des cellules corticales (Ferrando et al., 2002).

Des travaux réalisés par Asnafi et coll. à l'hôpital Necker ont démontré récemment que l'arrêt de la maturation corticale observé avec la surexpression de TLX1/3 est dû à la liaison directe des TLX1/3 à ETS1 (Dadi et al., 2012). Ceci cause la répression épigénétique de l'enhanceosome ETS1/LEF1/RUNX1 du TCRα. En conséquence, le réarrangement des TCR-φα est bloqué. D'un autre côté, l'abolition des TLX1/3 restaure le réarrangement de TCRα et la différentiation cellulaire, mettant en lumière une possible thérapie ciblée dans le traitement des LAL-T.

Un autre exemple très intéressant d'une dérégulation épigénétique chez les LAL-T est celui d'une mutation hétérozygote (micro-insertion) en amont de l'oncogène TAL1. En effet, cette insertion intervient dans une région non codante en introduisent un motif de liaison pour le facteur de transcription MYB. En se liant à ce nouveau site, MYB, recrute ces partenaires tels que l'enzyme H3K27 acétylase (CBP) et le complexe RUNX1, GATA3 et TAL1, ce qui induit un échange de H3K27me3 à H3K27ac et la création d'un super enhancer dans cette région (Fig. 31). Ainsi, ce super enhancer provoque l'expression de l'oncogène TAL1 (Navarro et al., 2015, Mansour et al., 2014)

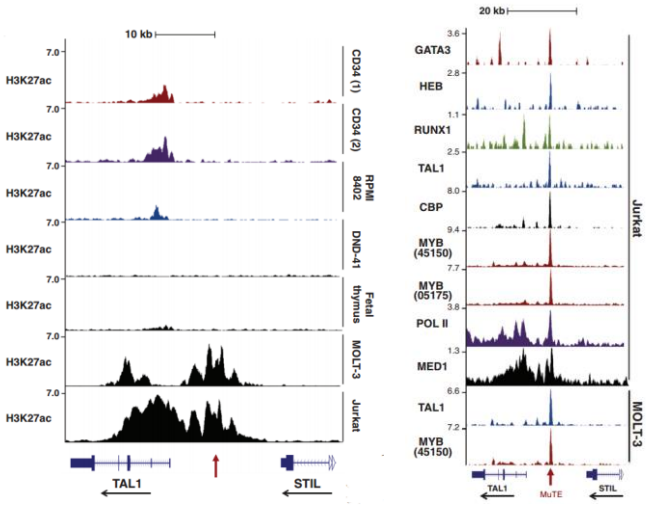



**Fig. 31 : Les données de ChIP-seq normalisées pour : [à Gauche] H3K27ac au locus STIL-TALl dans deux échantillons de cellules souches hématopoïétiques normales humaines (CD34) ; la lignée cellulaire RPMI-8402 qui surexprime TAL1 à la suite d'une délétion (cellules RPMI-8402) ; la lignée cellulaire DND-41 qui n'expriment pas TAL1 ; thymus fœtal humain, et les lignées cellulaires MOLT-3 et Jurkat qui ont des mutations sur un site non codant à environ 7,5 kb du TSS de TAL1 (flèche rouge). Les flèches noires indiquent la direction de la transcription. [à droite] GATA3, HEB, RUNX1, TAL1, MYB, CBP et MED1 dans les lignées cellulaires MOLT-3 et Jurkat.(Mansour et al., 2014)**

Les LncRNAs sont aussi impliqués dans la modulation de l'épigénome cancéreux et facilitent la transformation des cellules normales en cellules cancéreuses. Parmi les divers exemples de l'implication des LncRNAs dans le cancer, le rôle de HOTAIR dans les néoplasies humaines est le plus étudié (Esteller, 2011). Dans les cellules épithéliales

cancéreuses, la surexpression de HOTAIR cause la redistribution des complexes Polycomb dans le génome. En conséquence, les complexes Polycomb ne répriment plus leurs gènes cibles usuels, comme les gènes du locus Hox, et mènent donc à l'augmentation de l'invasion cellulaire et à la propension à métastaser. Une fonction similaire a été suggérée pour d'autres LncRNAs, comme le LincRNA-p21. En effet, ce dernier est un répresseur dans la réponse transcriptionnelle dépendante de p53 (Huarte et al., 2010). Le LncRNA antisens p15, p15AS, est un autre exemple de LncRNA identifié dans les leucémies humaines. Le LincRNA p15AS réprime le locus du suppresseur de tumeur p15, en induisant la formation d'hétérochromatine (Yu et al., 2008). Aucun LncRNA n'a été identifié jusqu'à ce jour dans les LAL-T. Par contre, une inactivation génétique par les complexes répressifs Polycomb (PRC2) y a récemment été décrite. Ceci suggère un nouveau mécanisme potentiel dans la dérégulation des gènes et en absence d'anormalités oncogéniques (Ntziachristos et al., 2012). Ainsi, durant ma thèse je me suis intéressé à la découverte de nouveau LncRNA fonctionnel dans les leucémies (voir Chapitre III article 4).

# CHAPITRE II :
# Les domaines larges de la méthylation Histone 3 lysine 4

## 1. Introduction:

La caractérisation du paysage épigénétique contribue au décryptage des mécanismes de régulation gouvernant l'expression des gènes. Ainsi, le développement des technologies de séquençage à haut débit a permis l'étude des différentes combinaisons des modifications chromatiniennes dans le cadre des différentes régions génomiques. Ainsi, des signatures chromatiniennes ont été associées à chaque élément régulateur, tel que les enhancers et les promoteurs.

D'autres études ont pu identifier des signatures spécifiques qui s'associent aux différents statuts fonctionnels. Par exemple, les promoteurs actifs sont marqués par la marque H3K4me3, alors que les promoteurs réprimés sont quant à eux enrichi par la marque H3K27me3 (Barski et al., 2007).

Nous avons analysé, dans une publication parue en 2011, la dynamique chromatinienne de la marque H3K4me3 dans les enhancers (Pekowska et al., 2011). Nous avons identifiés une signature spécifique des enhancers actif. Ainsi, nous avons amené une évidence de l'existence d'un lien direct entre l'enrichissement en H3K4me3/Pol II et l'activité enhancer. En effet, il a été suggéré auparavant que les enhancers en général sont enrichis en H3K4me1 et non H3K4me3. Alors que dans notre étude, nous suggérons que la dynamique de la marque H3K4me3 dans les enhancers corrèle avec les niveaux de régulation transcriptionnelle (Fig. 32).

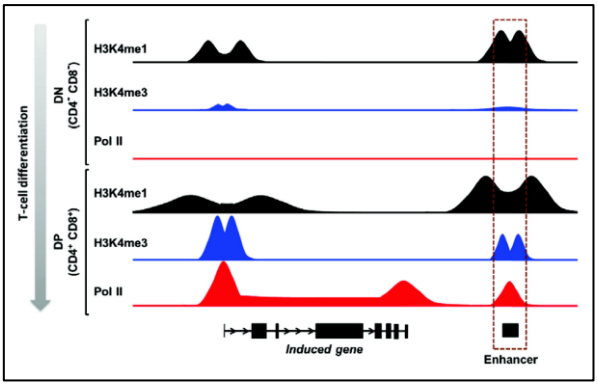



**Fig. 32: La dynamique épigénétique des gènes induits lors de la différenciation T. (Spicuglia and Vanhille, 2012)**

Cependant, très peu d'études se sont focalisées sur l'aspect qualitatif de ces profils épigénétiques. En effet, dans les profils reportés dans Barski et al. 2007, il a été observé des disparités dans la forme globale des profils dans certains locus. En général, les marques activatrices sont restreintes à des *loci* génomique très focalisés. Mais ils peuvent présenter aussi une distribution élargie qui s'étend sur des kilo-bases dans des régions régulatrices distales ou proximales. Par exemple, les super enhancers comme vue dans le chapitre I section 2.2.g sont des régions très larges (jusqu'à 50kb) composés de plusieurs enhancers, très enrichis en modifications épigénétiques H3K27ac et H3K4me1. Ces plateformes étendues sont associées aux gènes tissu-spécifiques et hautement exprimés en comparaison aux gènes classiques de ménage.

Quant à Koch et coll, ils ont identifié des régions génomiques appelées plateformes d'initiation transcriptionnelles (TIP) qui sont caractérisées par de vastes zones de recrutement de Pol II et des GTF sur des promoteurs, des régions intergéniques et intragéniques. Les TIPs montrent des largeurs variables (de 0,4 à 10 kb) et corrèlent avec un enrichissement élevée en CpG. Ce qui laisse à suggérer de nouvelles caractéristiques importantes du génome (Koch et al., 2011).

Dans une autre étude réalisée par mon équipe avant mon arrivé au laboratoire, il a été montré l'existence d'une sous classe de gènes présentant un niveau élevé des méthylation des histones H3K4 dans le corps du gène en comparaison à une autre classe qui présente un profil très focalisé autour du TSS (Fig. 33) (Pekowska et al., 2010).

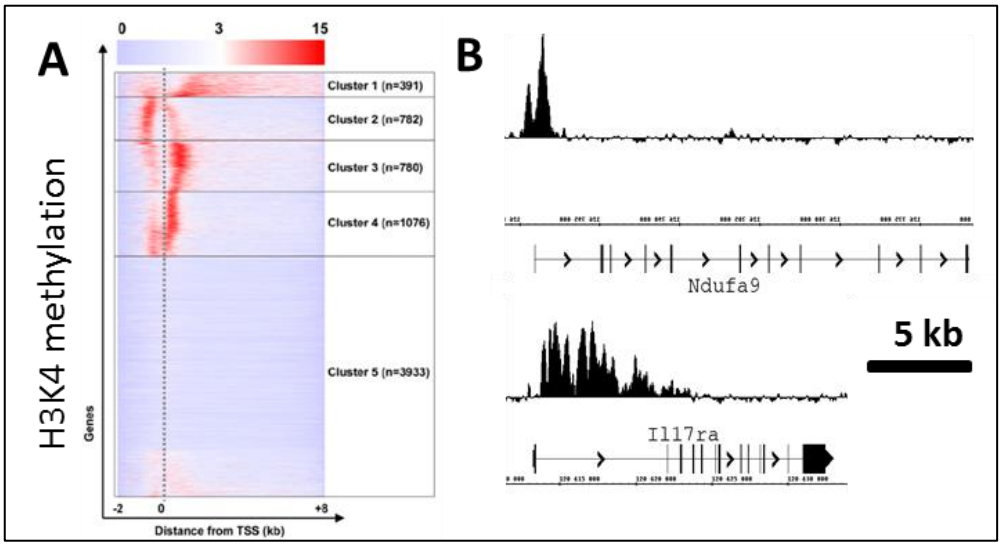



**Fig. 33: (A) *Clustering* du signal CHIP-seq H3K4me2 montre certains gènes avec des domaines larges (cluster1) de la méthylation et des gènes avec des profils localisés. (B) Des exemples de profil étroit localisé au TSS (haut à droite) et un profil étendu au corps du gène (bas à droite)**

Comme pour les super enhancers, ces profils de méthylation larges et les TIP sont une caractéristique des gènes tissu-spécifiques. Ils sont également liés à une déposition accrue de la PolII indépendamment de leur niveau d'expression (Pekowska et al., 2010, Koch et al., 2011).

Ces observations ont été confirmées par Benyoun et coll. ; Ils ont suggéré que cette classe des domaines larges de la H3K4me3 marque les gènes de l'identité cellulaire et la consistance du taux d'expression (Fig. 34). En effet, les gènes associés à ces domaines présentent un niveau de variabilité d'expression très faible à travers les différentes cellules du tissu d'intérêt (Benayoun et al., 2014).

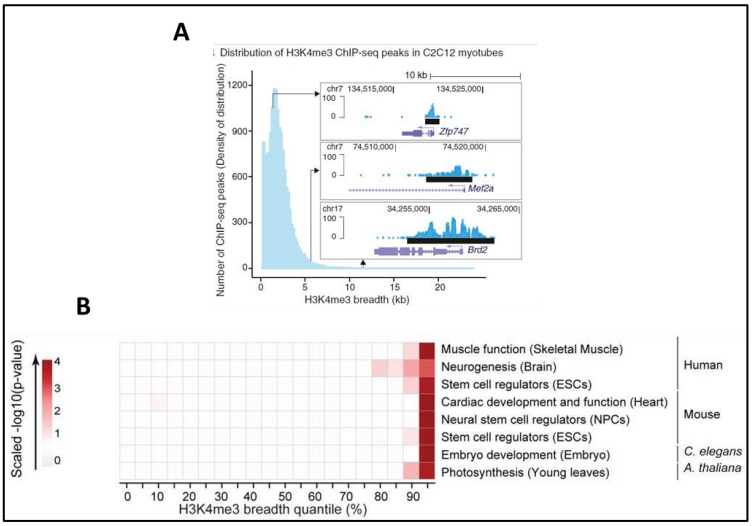



**Fig. 34: (A) distribution des pics large dans les cellules musculaires C2C12. (B) enrichissement des gènes associés aux pics larges en gènes tissu-spécifiques (Benayoun et al., 2014).**

Dans une analyse plus fine, publiée dans *Nature Genetics* par Chen et al. 2015, il a été observé que les domaines élargis présentent un enrichissement non seulement pour les gènes de l'identité cellulaire mais aussi aux gènes suppresseurs de tumeurs ; à l'inverse, des supers enhancers qui sont associés plutôt aux principaux oncogènes (Fig. 35). Ainsi dans cette étude les auteurs suggèrent l'utilisation de l'association et la conservation des domaines larges de la méthylation des histones afin d'identifier, de façon mutation-indépendante, des nouveaux gènes suppresseurs de tumeurs (Chen et al., 2015a) .

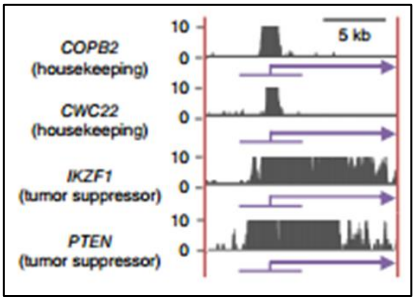



**Fig. 35: profils de H3K4me3 aux gènes de ménage COPB2 et CWC22 et les suppresseurs de tumeurs IKZF1 et PTEN (Chen et al., 2015a).**

Enfin, quatre articles parus dans un numéro spécial de *Nature* décrivent les domaines larges de la H3K4me3 dans les oocytes (Liu et al., 2016b, Dahl et al., 2016, Zhang et al., 2016, Wu et al., 2016). Toutefois, ces domaines ne sont pas associées à des gènes tissues-spécifiques mais plutôt une régulation épigénétique globale.

Ainsi, durant ma thèse je me suis intéressé a deux parties : (i) la caractérisation d'un cas extrême de ces domaines dans le locus $TCR_{\beta}$ ; (ii) l'étude de la dynamique des domaines larges de la H3K4me3 dans un modèle de différentiation T et un modèle leucémique.

# Article 1:

**Mise en évidence d'une régulation épigénétique spécifique du locus Tcrb**

Les gènes codant pour les récepteurs antigéniques Ig et TCR constituent de bons modèles d'étude de la différenciation cellulaire, en relation avec les processus épigénétiques de modulation de l'expression génique au sein de locus complexes. Dans ce projet, nous nous sommes demandé s'il existait un mécanisme spécifique responsable de l'ouverture de la chromatine au locus TCRβ. Nous avons mis en évidence un marquage épigénétique spécialisé caractérisé par des niveaux très élevés et étendus de H3K4me3 à travers les régions transcrites du locus TCRβ. Nous montrons que ce marquage spécifique dépend d'un mécanisme lié à la transcription par l'ARN polymérase II. Nous proposons que l'ensemble des régions transcrites du TCRβ se comportent comme des plates-formes "d'initiation transcriptionnelle", reliant ainsi un mécanisme spécialisé de la transcription avec l'établissement d'une structure chromatinienne caractérisée par un profil étendu de l'H3K4me3. Enfin, nous montrons aussi qu'un nombre réduit des gènes très tissus-spécifiques et hautement régulés au cours de la différenciation des lymphocytes T partage des propriétés similaires à celle décrite pour le TCRβ (voir Article2).

**Contributions dans l'article :**

- Réalisation de toutes les analyses bioinformatiques ;
- Participation à l'écriture de l'article ;
- Réalisation des figures :
    - Fig. 1D, E et F ;
    - Fig. 2D et E ;
    - Fig. 4A, B, C, D, E et F ;
    - Fig. 5A, B, C et D.

# Transcription-Dependent Generation of a Specialized Chromatin Structure at the TCRβ Locus

**Joaquin Zacarías-Cabeza,[*,†,‡,1,2] Mohamed Belhocine,[§,¶,2] Laurent Vanhille,[§,¶] Pierre Cauchy,[*,†,‡,§,¶] Frederic Koch,[*,†,‡] Aleksandra Pekowska,[*,†,‡] Romain Fenouil,[*,†,‡] Aurélie Bergon,[§,¶,‖] Marta Gut,[#] Ivo Gut,[#] Dirk Eick,[**] Jean Imbert,[§,¶,‖] Pierre Ferrier,[*,†,‡] Jean-Christophe Andrau,[*,†,‡,††] and Salvatore Spicuglia[*,†,‡,§,¶]**

**V(D)J recombination assembles Ag receptor genes during lymphocyte development. Enhancers at AR loci are known to control V(D)J recombination at associated alleles, in part by increasing chromatin accessibility of the locus, to allow the recombination machinery to gain access to its chromosomal substrates. However, whether there is a specific mechanism to induce chromatin accessibility at AR loci is still unclear. In this article, we highlight a specialized epigenetic marking characterized by high and extended H3K4me3 levels throughout the Dβ-Jβ-Cβ gene segments. We show that extended H3K4 trimethylation at the *Tcrb* locus depends on RNA polymerase II (Pol II)–mediated transcription. Furthermore, we found that the genomic regions encompassing the two DJCβ clusters are highly enriched for Ser[5]-phosphorylated Pol II and short-RNA transcripts, two hallmarks of transcription initiation and early transcription. Of interest, these features are shared with few other tissue-specific genes. We propose that the entire DJCβ regions behave as transcription "initiation" platforms, therefore linking a specialized mechanism of Pol II transcription with extended H3K4 trimethylation and highly accessible Dβ and Jβ gene segments.** ***The Journal of Immunology*****, 2015, 194: 3432–3443.**

V(D)J recombination assembles Ag receptor genes from germline V, D, and J segments during lymphocyte development (1). In αβT cells, this leads to the subsequent expression of TCR β- and α-chains. For V(D)J recombination to occur, the presence of the lymphoid-specific proteins RAG1 and RAG2 and the ubiquitously expressed DNA repair factors from the nonhomologous end joining pathway are required (2). Control of V(D)J recombination is required to ensure cell lineage specificity, dictate the temporal order of rearrangements, and allow allelic exclusion at certain AR genes (3). This regulation mainly relies on the modulation of chromatin accessibility at the AR-associated recombination sequences (RSs)to the recombination machinery.

The accessibility model was initially based on the observation that transcription of AR germline gene segments correlated developmentally with their recombination during lymphoid cell differentiation (4). Subsequently, this model has been strengthened by findings that link V(D)J recombination to transcriptional control elements, such as AR-associated enhancers and promoters, and to several molecular parameters related to open chromatin (including association with active histone marks, DNA hypomethylation, and nuclease hypersensitivity) (3, 5). Robust germline transcription at

*Centre d'Immunologie de Marseille-Luminy, Aix-Marseille University, UM2, 13288 Marseille, France; [†]INSERM, U1104, 13288 Marseille, France; [‡]Centre National de la Recherche Scientifique, UMR7280, F-13009 Marseille, France; [§]INSERM U1090, Technological Advances for Genomics and Clinics, F-13009 Marseille, France; [¶]Aix-Marseille University, UMR-S 1090, Technological Advances for Genomics and Clinics, F-13009 Marseille, France; [‖]Transcriptomic and Genomic Marseille-Luminy, Infrastructures en Biologie, Santé et Agronomie, 13288 Marseille, France; [#]Centre Nacional D'Anàlisi Genòmica, Parc Científic de Barcelona, Baldiri i Reixac, 08028 Barcelona, Spain; **Department of Molecular Epigenetics, Helmholtz Center Munich, Center for Integrated Protein Science, 80336 Munich, Germany; and [††]Institut de Génétique Moléculaire de Montpellier, Centre National de la Recherche Scientifique, UMR5535, 34293 Montpellier, France

[1]Current address: Institute of Biomedical Research, University of Birmingham, Birmingham, UK.

[2]J.Z.-C. and M.B. contributed equally to this work.

Received for publication March 27, 2014. Accepted for publication February 2, 2015.

This work was supported by institutional grants from INSERM, the Centre National de la Recherche Scientifique, Aix-Marseille University, and by specific grants from the Fondation Princesse Grace de Monaco (to P.F.), the Fondation de France (to P.F.), the Association pour la Recherche sur le Cancer (to S.S., Project SFI20111203756), the Fondation pour la Recherche Médicale (to P.F.), the Agence Nationale de la Recherche (to P.F.), the Institut National du Cancer (to P.F.), the European Union's Seventh Framework Program (FP7) (to S.S., Agreement 282510-BLUEPRINT), and Initiative D'Excellence Aix-Marseille Project ANR-11-IDEX-0001-02 (to S.S.) funded by the Investissements d'Avenir French Government program. Sequencing costs for this work were supported by a European Study Group with Industry Consortium grant of the European Union (to J.-C.A., program T-DynRegSeq) from the FP7 (FP7/2007-2013) under Grant Agreement 262055. The Transcriptomic and Genomic Marseille-Luminy sequencing platform is supported by grants from Infrastructures en Biologie, Santé et Agronomie and the France Génomique National infrastructure, funded as part of the Investissements d'Avenir program managed by the Agence Nationale pour la Recherche (Contract ANR-10-INBS-09). J.Z.-C. was supported by Grant R07116AS from the Agence Nationale de la Recherche MIMe program (to P.F.). Work in J.-C.A.'s laboratory is also supported by a grant of the Fondation pour la Recherche Médicale (AJE20130728183).

The sequences presented in this article have been submitted to the National Center for Biotechnology Information's Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) under accession numbers GSE63416, GSE64709, GSM1360722–GSM1360727, and GSM1359828.

Address correspondence and reprint requests to Dr. Pierre Ferrier, Dr. Jean-Christophe Andrau, or Dr. Salvatore Spicuglia, Centre d'Immunologie de Marseille-Luminy, Aix-Marseille University, UM2, 13288 Marseille, France (P.F.), Institut de Génétique Moléculaire de Montpellier, Centre National de la Recherche Scientifique, UMR5535, 34293 Montpellier, France (J.-C.A.), or Technological Advances for Genomics and Clinics, INSERM UMR-S 1090, Parc Scientifique de Luminy, Cedex 9, F-13009 Marseille, France (S.S.). E-mail addresses: ferrier@ciml.univ-mrs.fr (P.F.), jean-christophe.andrau@igmm.cnrs.fr (J.-C.A.), or salvatore.spicuglia@inserm.fr (S.S.)

The online version of this article contains supplemental material.

Abbreviations used in this article: ChIP, chromatin immunoprecipitation; CTD, C-terminal domain; DN, double negative; DP, double positive; FAIRE, formaldehyde-assisted isolation of regulatory elements; GEO, Gene Expression Omnibus; Pol II, polymerase II; qPCR, quantitative PCR; RS, recombination sequence; seq, sequencing; Ser[5P], phosphorylated serine 5; TIP, transcription initiation platform; TSS, transcriptional start site.



www.jimmunol.org/cgi/doi/10.4049/jimmunol.1400789

(D)J clusters is an initial activation event at all AR loci that generates a focal zone of RAG1/2 binding, termed the recombination center (6, 7). More insight into the accessibility model was provided by recent studies demonstrating that the PHD finger domain of RAG2 binds with high affinity to histone H3 trimethylated at K4 (H3K4me3) and that RAG2 is recruited to H3K4me3 domains genome-wide (6–8).

A central prediction of the accessibility model is, therefore, that transcriptional control elements and/or transcription itself are critical for allowing the recombination machinery to gain access to RSs (9). However, in most mammals' genes, highly open chromatin structure is mainly confined to the *cis*-regulatory sequences themselves (10). In particular, H3K4me3 is highly enriched at promoter regions of expressed genes but is not generally found in the body of the genes (11). Thus, the question still remains as to how chromatin accessibility is established at the recombining gene segments and associated RSs, which are often located distant from the *cis*-regulatory elements. We and others have recently shown that a subset of tissue-specific genes might display broad epigenetic marking, including extended H3K4me2 and H3K4me3, along with elevated loading of polymerase II (Pol II) (12–14). This raises the possibility that transcriptional activity throughout V(D)J rearranging loci might play a more elaborate role in the remodeling of chromatin structure and targeting of the recombinase machinery.

Genetic studies at the *Tcrb* locus have shed light on the complex cooperation between enhancer- and promoter-bound transcription factors to control V(D)J recombination during T lymphocyte development (15, 16). The mouse *Tcrb* locus spreads over ~670 kb, including a ~390-kb 5′ domain containing 21 Vβ gene segments and a 26-kb 3′ domain comprising a duplicated cluster of Dβ-Jβ-Cβ gene segments, followed by a single Vβ gene segment, Vβ31. *Tcrb* gene recombination is restricted to the T cell lineage and is activated along with locus expression. In $CD4^-CD8^-$ double negative (DN) thymocytes, V(D)J recombination proceeds in a stepwise manner (Dβ-to-Jβ joining occurring first, before Vβ-to-DJβ assembly), triggering, if productive, allelic exclusion at the *Tcrb* locus and further development into the $CD4^+CD8^+$ double positive (DP) cell stage in the αβT cell lineage, an intricate process also known as β-selection (17).

The 560-bp *Tcrb* gene enhancer (Eβ) lies at the center of the ~10-kb Cβ2-Vβ31 intervening region (18, 19). Knockout mouse models have revealed a critical function of Eβ in the efficient onset of *cis* recombination, with homozygous Eβ-deleted ($E\beta^{-/-}$) mice displaying impaired TCRβ-chain production and αβT cell development (20, 21). Further analysis implied that this element, working together with Dβ-associated promoters of germline transcripts, directs transcription, along with histone marking and chromatin opening, throughout the adjacent DJCβ clusters (22–26). Although Eβ-dependent activity is clearly required to initiate V(D)J recombination at the *Tcrb* locus, the precise mechanism or mechanisms inducing long-range histone marking and chromatin remodeling along the DJCβ regions are still poorly understood. A key feature is the Eβ-dependent transcription activity across the Dβ-Jβ recombination center, which is thought to mediate H3K4 trimethylation at this site, followed by RAG1/2 deposition (3, 16, 22, 23, 26–28).

In the current study, we used chromatin immunoprecipitation (ChIP)–sequencing (seq) technology to comprehensively map H3K4 methylation in germline *Tcrb* alleles from $Rag2^{-/-}$ thymocytes. We found that the DJCβ transcription units were highly enriched for H3K4me3 and linked to local accessibility of the Dβ and Jβ gene segments, highlighting a distinctive epigenetic marking at the *Tcrb* locus. This property was dependent on an unusual regulation of Pol II–mediated transcription in which features of transcription initiation and early elongation, including high levels of phosphorylated serine 5 ($Ser^{5P}$) Pol II and short-RNA transcripts, were found throughout the entire DJCβ regions. Of interest, these features are shared with a small subset of tissue-specific genes, including other *Tcr* loci. Overall, our study revealed a specialized role for Pol II transcription in the establishment of a highly accessible chromatin domain at the *Tcrb* locus.

## Materials and Methods

### *Mice*

Homozygous *Rag2*-deficient (ΔRag) (29) and Eβ-deleted (ΔEβ) (21) mice were housed under specific pathogen–free conditions and handled in accordance with European directives. Mice were bred on a C57BL/6J background and sacrificed for analysis between 4 and 6 wk of age.

### *KM05283 treatment*

A total of $15 \times 10^6$ exponentially growing P5424 cells (30) was incubated with either 50 μM KM05283 (Maybridge, Cornwall, U.K.) or control DMSO (Sigma-Aldrich, St. Louis, MO) in RPMI 1640 medium for 16–18 h at 37°C. After incubation, cells were washed two times with 1× Dulbecco's PBS and processed for ChIPs as indicated below. Inhibition of Pol II $Ser^2$ phosphorylation was confirmed by Western blot, as described previously (31).

### *ChIP*

ChIP experiments were performed as described previously (14). For histone modification marks, we used $2 \times 10^6$ cells along with 3 μg of the following Abs: anti-H3K4me1 (ab8895; Abcam, Cambridge, U.K.), anti-H3K4me2 (ab32356; Abcam), anti-H3K4me3 (ab8580; Abcam), and anti-H3K36me3 (ab9050; Abcam). For Pol II ChIPs, the following Abs and cell numbers were used: anti–total-Pol II (Santa Cruz Biotechnology, Dallas, TX; sc-899×, 10 μg and $10 \times 10^6$ cells), anti-$Ser^{2P}$ Pol II [rat monoclonal, clone 3E10 (32), 10 μg and $60 \times 10^6$ cells], and anti–$Ser^{5P}$ Pol II [rat monoclonal, clone 3E8 (32); 10 μg and $30 \times 10^6$ cells]. The DNA fragments were purified and recovered using the QIAquick PCR Purification Kit (QIAGEN, Hilden, Germany). The quality of individual ChIP samples was checked at known target sites by quantitative PCR (qPCR), and DNA size was verified on a 2100 Bioanalyzer (Agilent, Santa Clara, CA). Primer sets used for qPCR are available upon request.

### *Formaldehyde-assisted isolation of regulatory elements*

Formaldehyde-assisted isolation of regulatory elements (FAIRE) was performed as previously described (33), with slight modifications. Briefly, $20 \times 10^6$ thymocytes from ΔRag or ΔRag;ΔEβ mice were cross-linked with 1% formaldehyde for 10 min at room temperature and sonicated 14 times on an S-4000 Sonifier (Misonix, Farmingdale, NY) with 30-s pulses to give DNA fragments of length between 200 and 500 bp. The soluble chromatin of $2 \times 10^6$ thymocytes was isolated and subjected to three consecutive phenol-chloroform extractions. Samples were then incubated overnight at 65°C to reverse cross-linking. DNA was finally purified using the MinElute PCR Purification Kit (QIAGEN). DNA concentration was measured using a Nanodrop 1000 (Thermo Scientific, Illkirch, France).

### *ChIP-seq data generation*

Sequencing of ChIP samples was performed according to the Illumina Genome Analyzer ChIP-seq protocol and aligned against the mouse mm9 genome using integrated Eland software. As prefiltering steps, only uniquely mapped tags were used for further processing, and all duplicate tags (those with identical coordinates) were filtered out to remove possible sequencing and/or alignment artifacts. Remaining tags were processed using a custom R pipeline, employing the ShortRead library3 (14). Read-count intensity profiles (wiggle files) were constructed by elongating each mapped read to the estimated fragment size, and counting the elongated read overlaps within a window of 50 nucleotides after normalization of the profile by the number of mapped reads. ChIP-seq data from total-Pol II and from micrococcal nuclease–treated H3K4me1 and H3K4me3 from ΔRag thymocytes were published previously (Ref. 34; GSE55635). Mapped reads, estimated fragment size, and Gene Expression Omnibus (GEO) accession numbers are listed in Supplemental Table I.

### *RNA extraction and RNA-seq experiments*

Total RNA from $10 \times 10^6$ thymocytes of ΔRag mice was extracted as previously described (14). Strand-specific preparation, sequencing, and processing of short-RNA samples were carried out as explained earlier (14). RNA quantity and quality were verified using RNA Pico chips on a 2100 Bioanalyzer (Agilent). Mapped reads and GEO accession numbers are listed in Supplemental Table I. Total and polyA RNA-seq data from ΔRag thymocytes were published previously (Ref. 35; GSE44578).

### *ChIP-seq and RNA-seq data analyses*

We first selected non-overlapping genes, harboring a single transcript annotated in the RefSeq database and longer than 8 kb (Supplemental Table II). From this set, the 300 highest expressed genes (Top-300) were selected, based on gene expression data in ΔRag thymocytes (34). To quantify the enrichment levels in H3K4me3, $Ser^5$ Pol II, and short-RNAs within the gene body, the ChIP-seq signal from wiggle files was quantified within the region from the transcriptional start site (TSS) to +8 kb. In the case of the DJCβ1, DJCβ2, Dδ2Jδ1, Jγ1Cγ1, and Jγ4Cγ4 clusters, the region from the Dβ1, Dβ2, Dδ2, Jγ1, and Jγ2 gene segments to +8 kb, respectively, was used to quantify H3K4me3, $Ser^5$ Pol II, and short-RNA levels. To directly compare expression levels between the selected RefSeq genes and the different *Tcr* gene clusters, we used polyA RNA-seq data. PolyA RNA level was estimated by counting the average number of tags at the exons of RefSeq genes and at the different *Tcr* gene clusters. All quantifications are shown in Supplemental Table II. The pausing index (also called traveling ratio) was calculated as previously described (36), using the selection of the Top-300 genes.

### *Analysis of ChIP-on-chip and FAIRE experiments*

Enriched DNA fragments from ChIP or FAIRE experiments were hybridized together with input DNA to a previously described 15K array (Agilent) containing the whole *Tcrb* locus at 100-bp resolution (34), following the manufacturer's instructions. The results obtained with two biological replicates were averaged and converted into SGR files using CoCAS software (37). Data from ΔRag and ΔRag;ΔEβ thymocytes were normalized using the overall signal on the entire microarray (excluding the probes within the *Tcrb* locus). Normalized data were displayed in the form of log2 ratio using IGB software (http://bioviz.org/igb/).

### *Inflection point*

We first calculated the average signal of H3K4me3 in the gene body (TSS to +8kb) for each gene. The broad H3K4me3 genes were then determined by identifying an inflection point of the average signal versus gene rank. The inflection point was computed by determining the diagonal line of the curve from endpoints, and by sliding this diagonal line to find where it is tangential. We identified 58 broad H3K4me3 genes (Supplemental Table II).

### *Transcription initiation platform selection*

We selected promoter-associated transcription initiation platforms (TIPs), defined previously in DP thymocytes (14) and expressed in the P5424 cell line (671 TIPs). The TIPs were separated according to their size into three categories: <2 kb (557), between 2 and 2.5 kb (47), and >2.5 kb (67).

### *Average and boxplot profiles*

Average profiles were generated by extracting the ChIP-seq signal from wiggle files around the TSS (from −2kb to +8kb), using a custom R script. Rescaled average profiles were performed by dividing the region from the TSS to the transcriptional termination site into 200 bins. To test whether the differences between different gene sets were statistically significant, we first extracted the average signal of the region of interest and plotted them in boxplot representation and performed a Student *t* test.

### *Gene expression analyses*

Gene expression data of αβ T cells were downloaded from the Immunological Genome Project Web site (www.immgen.org) (38). A quantile normalization was then applied on gene expression of stages ETP (early thymic progenitor), DN1 (DN stage 1), DN2, DN3, DN4, ISP (immature single positive), and DPbl (DP blast). The raw expression data for 74 mouse tissues were downloaded from the National Center for Biotechnology Information GEO (accession number: GSE10246). The raw expression data were normalized by the variance stabilization and normalization method (39), and probe annotation to the NCBI37/mm9 was used for subsequent analyses. To compare the level of expression of genes between T cells and other tissues, we calculated the mean level of expression of genes in T cell samples (including T cell $CD4^+$, T cell $CD8^+$, T cell $Foxp3^+$, thymocytes DP $CD4^+/CD8^+$, thymocyte SP $CD4^+$, and thymocyte SP $CD8^+$) and in the remaining 69 samples. Statistical significance was calculated using a paired Student *t* test.

### *Gene ontology terms enrichment*

Enrichments in Gene Ontology Terms for Biological Process were calculated using the DAVID tool (40), with default settings (count threshold: 2; EASE threshold: 0.1; multiple testing correction by the Benjamini procedure) and *Mus musculus* as background model. We selected the top 10 terms retrieved for each gene set with the lowest *p* values.

### *CpG density*

The genomic sequences ±500 bp around the TSS of each set of genes were recovered. The total number of CpG was counted for each sequence. Statistical significance between the Top-300 and the other set of genes was calculated using a Student *t* test.

## Results

### *A highly open chromatin structure at the DJCβ region*

To assess epigenetic features associated with chromatin remodeling of the *Tcrb* locus, we analyzed the three levels of histone H3K4 methylation by ChIP of thymocytes purified from *Rag2*-deficient mice (hereafter ΔRag), followed by high-throughput sequencing (ChIP-seq). The use of the ΔRag mouse model ensures the germline configuration of *Tcrb* alleles while providing an enriched and homogeneous source of T cell precursors. We concentrated our analyses on the Eβ-proximal region, including the two DJCβ clusters (Fig. 1A). We observed that H3K4 methylation marks were not exclusively localized to the known regulatory regions (i.e., the pDβ promoters and Eβ) but, instead, extended throughout the Jβ and Cβ regions. For instance, H3K4me1 and H3K4me2 covered the entire Eβ-proximal region spanning 30 kb from ∼3 kb upstream of Dβ1 to ∼3 kb downstream of Vβ31, thus defining a domain of open chromatin that roughly corresponds to the previously described Eβ-regulated domain (22–26, 41). Intriguingly, however, H3K4me3, which has been shown to be highly enriched at promoter regions (11), but is not generally found in the body of the genes, was broadly distributed throughout the two DJCβ germline transcription units. To exclude any potential bias owing to cross-linked chromatin, we confirmed the extended profile observed for H3K4me1 and H3K4me3 at the DJCβ regions by analyzing ChIP-seq data performed with mononucleosome preparations of native chromatin from ΔRag thymocytes (34) (Fig. 1B). Moreover, Eβ-deleted alleles displayed an almost complete loss of H3K4me3 at the DJCβ regions, suggesting that this epigenetic marking depends on Eβ-mediated transcriptional activation of the locus (Fig. 1C).

We then asked whether the extended H3K4me3 profile observed at the DJCβ regions in ΔRag thymocytes could be reminiscent of a highly open chromatin. To directly determine the accessibility of the chromatin, we performed a FAIRE assay, which allowed the recovery of the soluble (i.e., nucleosome-free) fraction of the chromatin (33). As expected, FAIRE signals were highly enriched at the Eβ region in ΔRag thymocytes (Fig. 1C). In addition, we observed that regions overlapping the Dβ and Jβ gene segments also display high levels of FAIRE signal in ΔRag. We confirmed that the highly open chromatin revealed by FAIRE at the Dβ and Jβ gene segments was largely dependent on Eβ-mediated chromatin remodeling (Fig. 1C). These results confirm and extend previous observations describing extensive Eβ-dependent remodeling of the DJCβ clusters (3, 16, 22, 23, 25–27). Note, however, that residual levels of chromatin accessibility are still observed around the Dβ

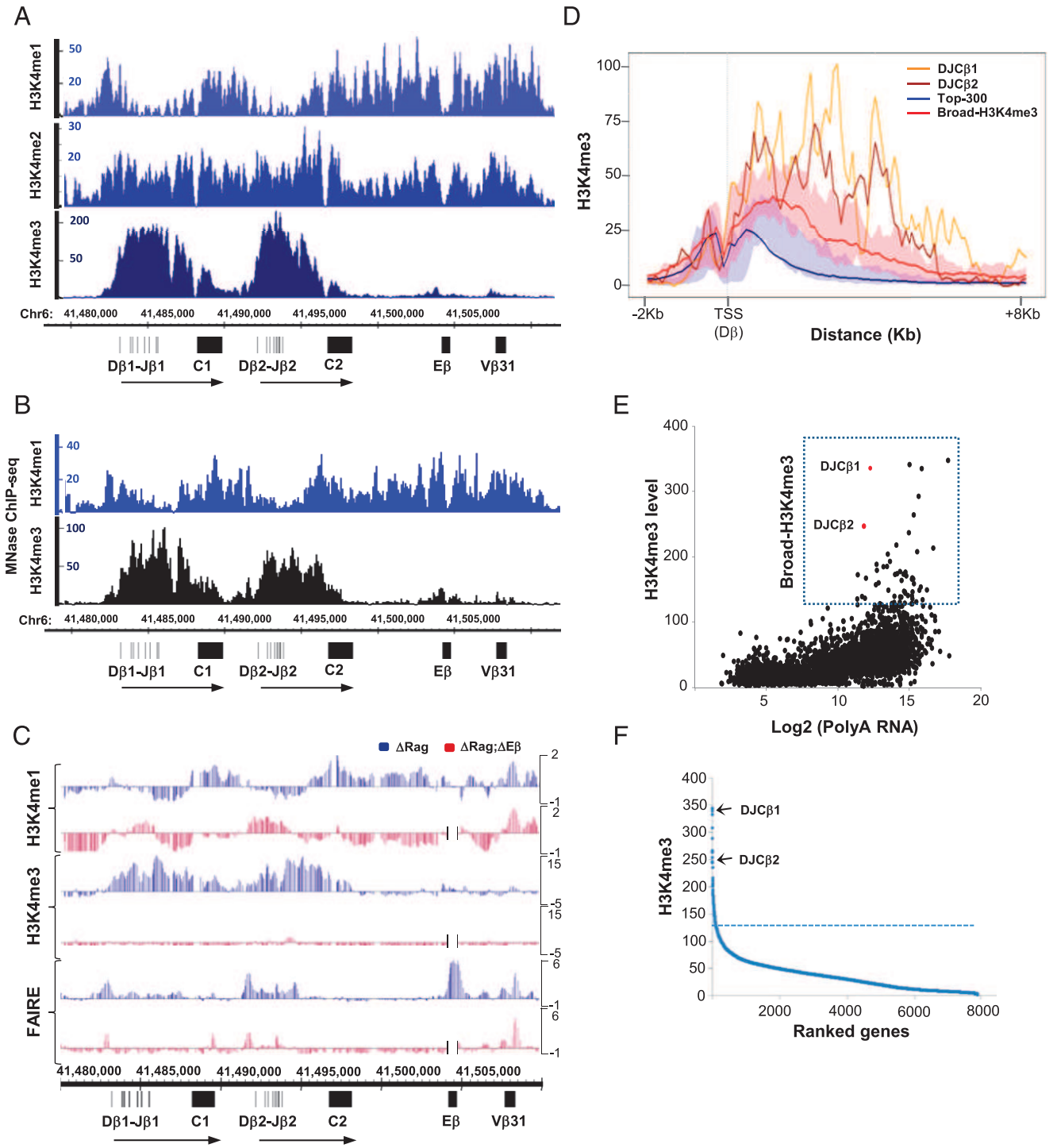



**FIGURE 1.** Extended H3K4me3 profiles at the DJCβ clusters. (**A**) ChIP-seq profiles of H3K4me1, H3K4me2, and H3K4me3 from cross-linked chromatin from ΔRag thymocytes at the 3′ side of the *Tcrb* locus. The genomic coordinates (mm9) and the location of the *Tcrb* gene segments and the Eβ enhancer (vertical dashes and rectangles) are indicated at the *bottom* of the panel. Horizontal arrows indicate the orientation of the two DJCβ transcription units. (**B**) ChIP-seq profiles of H3K4me1 and H3K4me3 from native chromatin from ΔRag thymocytes at the 3′ side of the *Tcrb* locus. Symbols are as in (A). (**C**) ChIP-on-chip and FAIRE profiles from ΔRag and ΔRag;ΔEβ thymocytes at the 3′ side of the *Tcrb* locus. ChIP-on-chip and FAIRE experiments were performed in duplicate and the merged ChIP(FAIRE)-over-input signals were normalized between the ΔRag and ΔRag;Eβ samples (see *Materials and Methods* for details). The log2 scale is shown. Deletion of the Eβ region in ΔRag;Eβ thymocytes is represented as a gap. (**D**) Average profiling of H3K4me3 ChIP-seq data in ΔRag thymocytes at genomic regions from −2 to +8 kb around the TSS of the set of highly expressed genes (Top-300) and the set of Broad-H3K4me3 associated genes, compared with the profiles of the two DJCβ clusters (the coordinates of the Dβ gene segments were set to 0). The shadow represents the border of the 25th and 75th percentiles. (**E**) Scatter plot showing the H3K4me3 density in the gene body of Refseq genes and in the two DJCβ clusters related to the polyA RNA-seq level. Each dot represents one gene. The two DJCβ clusters are highlighted in red. (**F**) Graph showing the ranked Refseq genes and the two DJCβ clusters related to H3K4me3 density. The two DJCβ clusters are highlighted by arrows. The threshold defining the set of Broad-H3K4me3 genes is indicated by a dotted line.

gene segments in the absence of Eβ, in agreement with an Eβ-independent role of Dβ-associated promoters (23, 25). Overall, in ΔRag thymocytes, highly accessible chromatin domains at the 3′ proximal region of the *Tcrb* locus are not restricted to the enhancer and promoter elements, but are spread over the Dβ and Jβ gene segments, thus providing a unique chomatin signature.

### *The extended H3K4me3 profile is a specific feature of the* Tcrb *locus*

To determine whether this extended profile was a general feature of highly expressed genes, we compared the H3K4me3 profiles at the two DJCβ clusters with the average H3K4me3 profiles of a set of highly expressed genes (Top-300; Fig. 1D). As predicted, expressed genes displayed an H3K4me3 enrichment around the TSS (peaks at −0.5 and +1 kb from the TSS). In comparison, the H3K4me3 profiles at the DJCβ regions extended throughout the transcribed regions with no particular enrichment at the 5′ sides. Moreover, we found that H3K4me3 levels were 3- to-4 fold higher at the DJCβ regions than the level observed around the TSS of highly expressed genes (Fig. 1D). To directly compare the H3K4me3 enrichment within the gene body of individual genes, we calculated the density of H3K4me3 at the two DJCβ clusters and within the genomic regions from the TSS to +8 kb of mRNA genes. We next plotted the H3K4me3 values in the function of mRNA levels, obtained by polyA RNA-seq (see *Materials and Methods*). As shown in Fig. 1E, the two DJCβ clusters displayed very high levels of H3K4me3 as compared with the rest of the genes, whereas the mRNA level of the two clusters was relatively modest. We observed that a relatively small subset of genes also displayed elevated H3K4me3 enrichment (Fig. 1E). Genes ranked in function of H3K4me3 level identified 59 genes harboring substantially higher levels of H3K4me3 (Fig. 1F; see *Materials and Methods* for details). These genes displayed a broad distribution of H3K4me3 within the 5′ regions of the gene body (Fig. 1D; hereafter named Broad-H3K4me3 genes), as observed for the *Tcrb* locus, and reminiscent of previous findings of genes associated with extended H3K4 methylation (12, 13). However, the two DJCβ clusters ranked within the top 10 of the highest H3K4me3-enriched genes in ΔRag thymocytes (Fig. 1F). Thus, the active DJCβ clusters display an unusual H3K4me3-extended chromatin structure that is larger and stronger than the one observed at the vast majority of expresssed genes, without being associated with a high level of polyadenylated RNA.

### *Pol II–dependent chromatin remodeling*

The above results raise the question of whether a specialized transcription mechanism plays a key role at the *Tcrb* locus, which ultimately leads to a highly accessible chromatin structure at the Dβ and Jβ gene segments. Chromatin accessibility at the AR loci has been generally associated with the presence of germline transcription (9). Moreover, H3K4me3 marking across the Jα segments of the *Tcra* locus has been shown to directly depend on germline transcription (42). More generally, functional links have been described between Pol II binding and H3K4 trimethylation at promoter regions (43, 44). Thus, we asked whether the atypical H3K4me3 profiles observed at the DJCβ regions may depend on Pol II–mediated transcription. To this end, elongating Pol II was blocked by inhibiting the CDK9 kinase with the KM05283 chemical compound (31, 34). We reasoned that if H3K4 trimethylation depends on local Pol II transcription, then its level was likely to decrease following the KM05283 treatment. In these experiments we used the pro–T cell line P5424, which is derived from ΔRag thymocytes and harbors a recombination-competent *Tcrb* locus (45). Efficient blocking of Pol II elongation upon KM05283 treatment was validated by global loss of phosphorylated serine 2 of the C-terminal domain (CTD) of Pol II ($Ser^{2P}$ Pol II), as assessed by Western blot (Fig. 2A), as well as complete loss at the *Tcrb* region of H3K36me3, a mark of transcription elongation (46; Fig. 2B). Next, we performed ChIP-seq experiments for H3K4me3 and total-Pol II, using both KM05283- and DMSO-treated chromatin. Interestingly, we observed a strong decrease in H3K4me3 within the two DJCβ transcription units in KM05283-treated P5424 cells (Fig. 2C; note that the H3K4me3 profile at the *Tcrb* locus was consistent between the P5424 cell line and ΔRag thymocytes). Thus, H3K4 trimethylation at the DJCβ regions is largely dependent on Pol II transcription.

In most expressed genes, transcription initiation and elongation are regulated independently. Indeed, inhibition of transcriptional elongation normally results in the loss of Pol II within the gene body and its accumulation at promoter regions (36). This result was confirmed by increased Pol-pausing index in the KM05283-treated cells (Fig. 2D) and exemplified by the average profiles of the set of 300 highly expressed genes (Fig. 2E, *left panel*), as well as visual inspection of several expressed genes (Fig. 2F). Strikingly, however, we observed a complete loss of Pol II binding at the two DJCβ regions after KM05283 treatment (Fig. 2C, 2E, *right panel*). The specific loss of Pol II binding at Dβ promoters, but not at control genes, upon inhibition of transcription elongation was further confirmed by independent ChIP-qPCR experiments (Fig. 2G). Thus, in the absence of transcription elongation, Pol II was unable to remain stably associated to the Dβ promoters, as is the case for the vast majority of expressed genes. These results suggest that, at the *Tcrb* locus, recruitment of Pol II is directly coupled to the elongation phase of transcription.

### *TIPs cover the DJCβ transcription units*

The above results suggest that 1) the broad distribution of H3K4me3 (and likely chromatin accessibility) is linked to Pol II–mediated transcription and 2) the regulation of the transcription process might differ between *Tcrb* locus and canonical mRNA coding genes. Pol II transcriptional activity is regulated via phosphorylation of the CTD (47). At expressed genes, phosphorylation of $Ser^{5}$ of the CTD, which is associated with transcription initiation and early elongation, is found at the 5′ end of genes, whereas $Ser^{2P}$, which is required for productive elongation, is found to be enriched at the 3′ end of genes (43, 44). Moreover, Pol II phosphorylation at $Ser^{5}$ has been shown to be required for H3K4me3 trimethylation (43, 44). To explore whether the *Tcrb* locus displays a distinctive Pol II profile, we first analyzed the distribution of phosphorylated and total forms of Pol II at the DJCβ1 cluster by ChIP-qPCR from ΔRag thymocytes. The expected patterns of Pol II phosphorylation were fully reproduced at control active genes *Actb* and *Sfrs*3: we found high levels of $Ser^{5P}$ at the TSS, low levels of phosphorylated Pol II within the gene body, and high levels of $Ser^{2P}$ at the 3′ end of these genes (Fig. 3A). However, in the case of the *Tcrb* locus, we found relatively high levels of $Ser^{5P}$ Pol II throughout the DJCβ1 region, whereas the $Ser^{2P}$ Pol II accumulated at the 3′ end of the DJCβ1 transcription unit (Fig. 3A). Indeed, although $Ser^{5P}$ Pol II downstream of the TSS of control genes is reduced to background levels, the enrichment at equivalent regions of the DJCβ1 cluster remains elevated.

To have a more comprehensive view of Pol II profiles at the *Tcrb* locus, we performed ChIP-seq experiments for both total- and $Ser^{5P}$ Pol II in ΔRag thymocytes. Again, we observed an accumulation of total- and $Ser^{5P}$ Pol II across the two DJCβ regions (Fig. 3B), whereas control genes displayed the expected patterns (Fig. 3C). Note that total- and $Ser^{5P}$ Pol II profiles were consistent between the ChIP-qPCR and ChIP-seq data (compare Figs. 3A with 3B, 3C). A more thorough analysis revealed that the $Ser^{5P}$ Pol II profiles were quantitatively and qualitatively different between the set of highly expressed genes and the DJCβ regions (Fig. 4A). Indeed, the level of $Ser^{5P}$ Pol II at the DJCβ regions was higher than the majority of expressed genes (Fig. 4B; the DJCβ clusters ranked in the top three of the highest $Ser^{5P}$ Pol II–enriched genes in ΔRag thymocytes). Thus, Pol II is found in its initiating/early elongating form throughout the entire DJCβ transcription units.

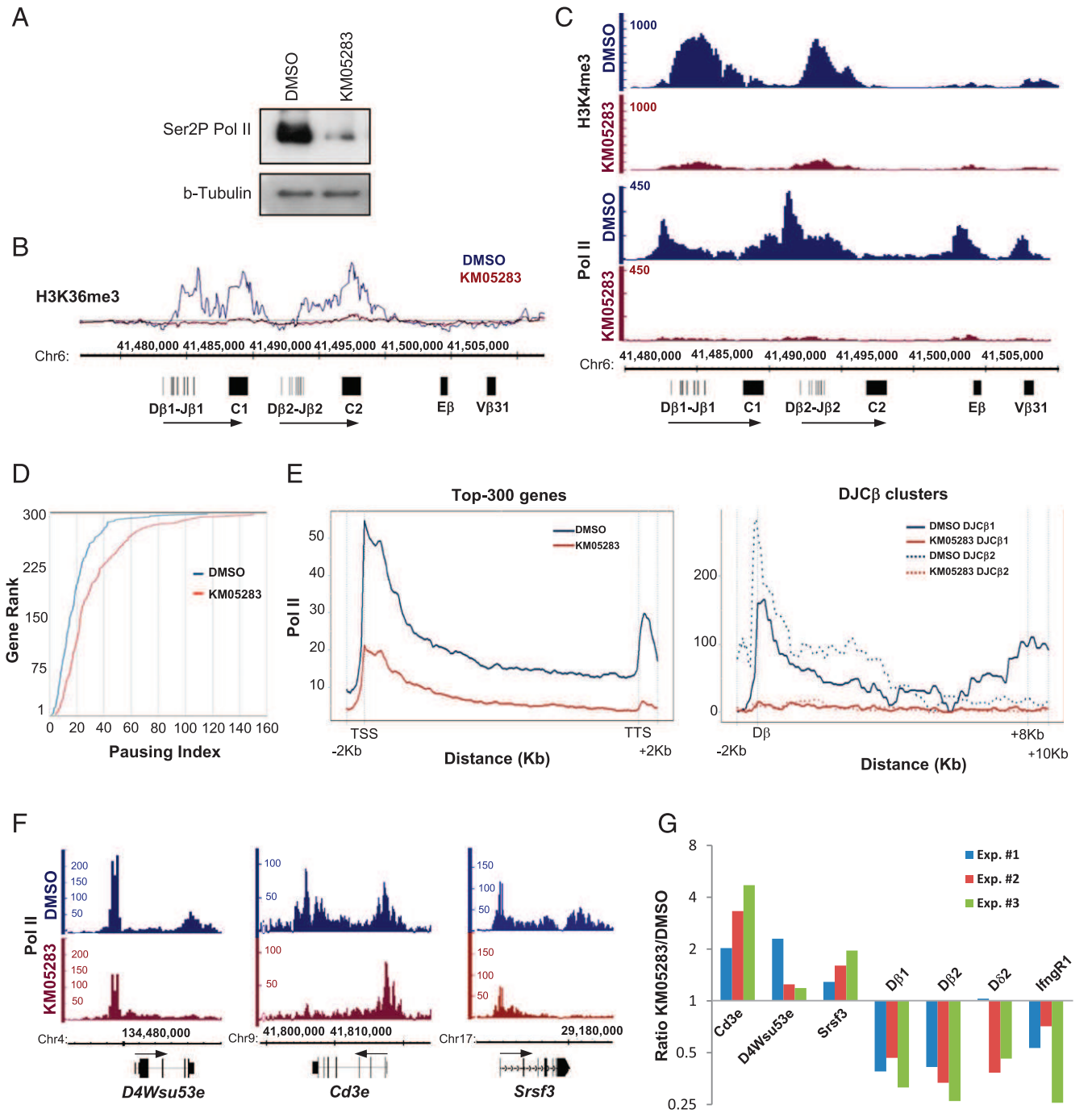



**FIGURE 2.** Pol II–mediated trimethylation at the DJCβ regions. (**A**) The level of $Ser^{2P}$ Pol II was assessed by Western blot in P5424 cells treated with either DMSO or KM05283 for 18 h. (**B**) ChIP-on-chip profiles after inhibiting the elongation activity of Pol II showed a significant reduction in H3K36me3 signals at the DJCβ regions compared with the cells treated with DMSO. (**C**) Profiles of H3K4me3 and Pol II ChIP-seq from the P5424 cell line treated with KM05283 or DMSO at the 3′ region of the *Tcrb* locus. (**D**) The pausing index was calculated as previously described (32) for a selection of highly expressed genes (Top-300) in P5424 cells treated with either DMSO or KM05283. (**E**) Average profiles of Pol II ChIP-seq from the P5424 cell line treated with KM05283 or DMSO centered on the Top-300 of expressed genes (*left panel*) or the DJCβ regions (*right panel*). (**F**) Profiles of H3K4me3 and Pol II ChIP-seq from the P5424 cell line treated with KM05283 or DMSO at three control genes. Arrows indicate the sense of transcription. (**G**) ChIP-qPCR analyses of Pol II binding in the P5424 cell line treated with KM05283 or DMSO at the promoters of the indicated genes. The KM05283/DMSO ratio is shown in a log2 scale. TTS, transcriptional termination site.

A hallmark of transcription initiation in higher eukaryotes is the presence of bidirectional short-RNAs around the TSS (hereafter short-RNA), a feature related to Pol II pausing (48). Given the above results, we hypothesized that the DJCβ regions might be enriched in initiating short transcripts. To explore this possibility, we performed short-RNA–seq experiments from ΔRag thymocytes and compared them with strand-specific total (ribosomal-depleted) and polyA RNA-seq profiles previously generated (35). As expected, total and polyA RNA-seq signals overlapped with the DJCβ regions and were oriented in the sense of defined transcription units (Fig. 3C). The continuous RNA-seq signal observed at the DJCβ regions probably reflects a low splicing efficiency at this locus. Analysis of short-RNA–seq data revealed the presence of several discrete peaks of short transcripts, along with an overall enrichment of this RNA population throughout the entire DJCβ regions, suggesting that

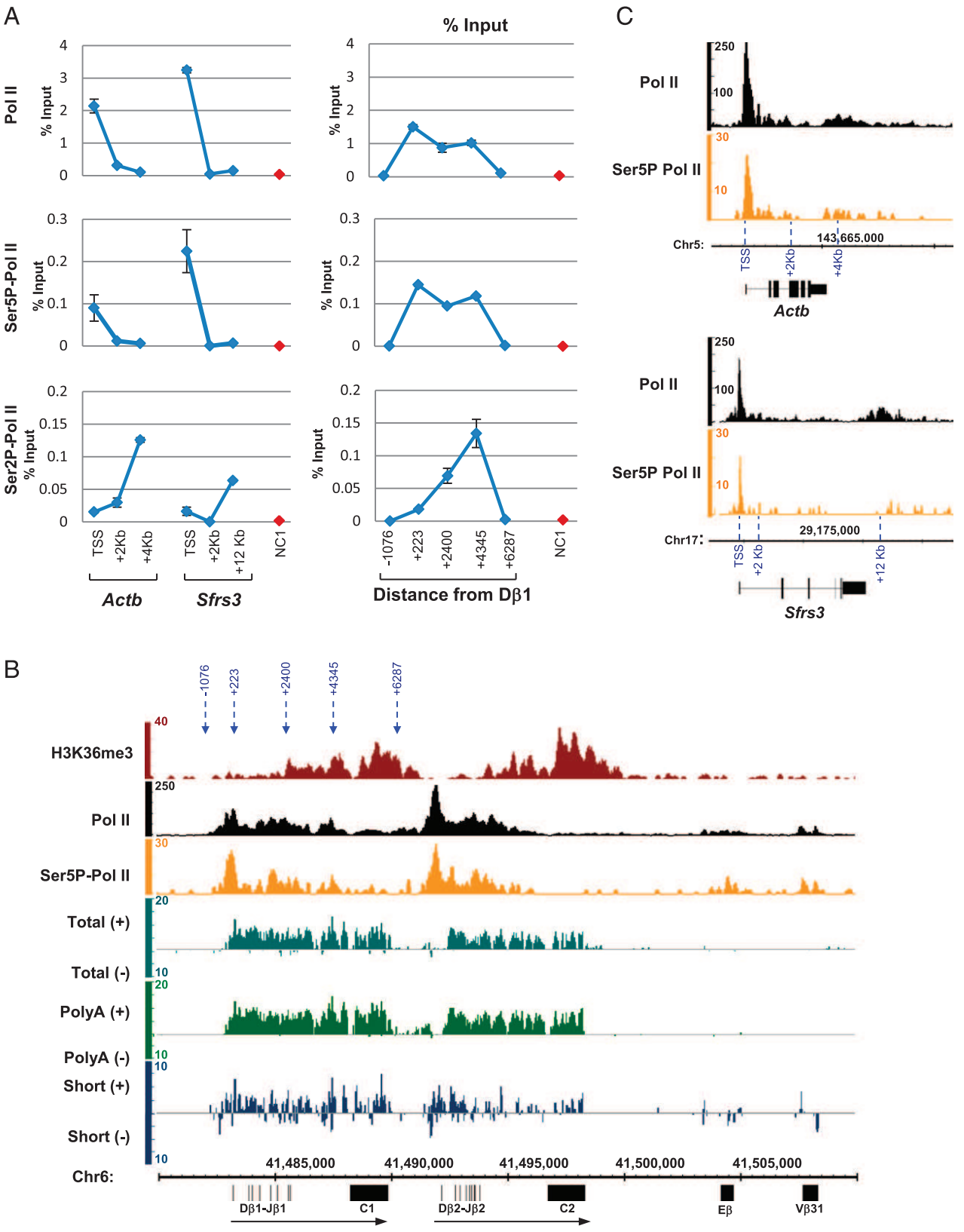



**FIGURE 3.** The DJCβ regions are enriched for transcription initiation features. (**A**) ChIP-qPCR assays showing the relative enrichment of total (N20), Ser$^{2}$ and Ser$^{5}$ phosphorylated Pol II at two active genes, at the indicated locations of the DJCβ1 region and at a negative control region (NC1). The genomic location of primer sets with respect to the Dβ1 gene segment or control genes is highlighted in (**B**) and (**C**). (B) Profiles of total and initiating (Ser$^{5P}$) Pol II ChIP-seq experiments, as well as total polyA and short-RNAs from directional RNA-seq experiments in ΔRag thymocytes, are shown at the 3′ region of the *Tcrb* locus. For RNA-seq the profiles are log2 scaled, and strand orientation is indicated at the *left* of each panel. Other data are as in Fig. 1. (C) The ChIP-Seq profiles of total and Ser$^{5P}$ Pol II in ΔRag thymocytes are shown at three active genes.

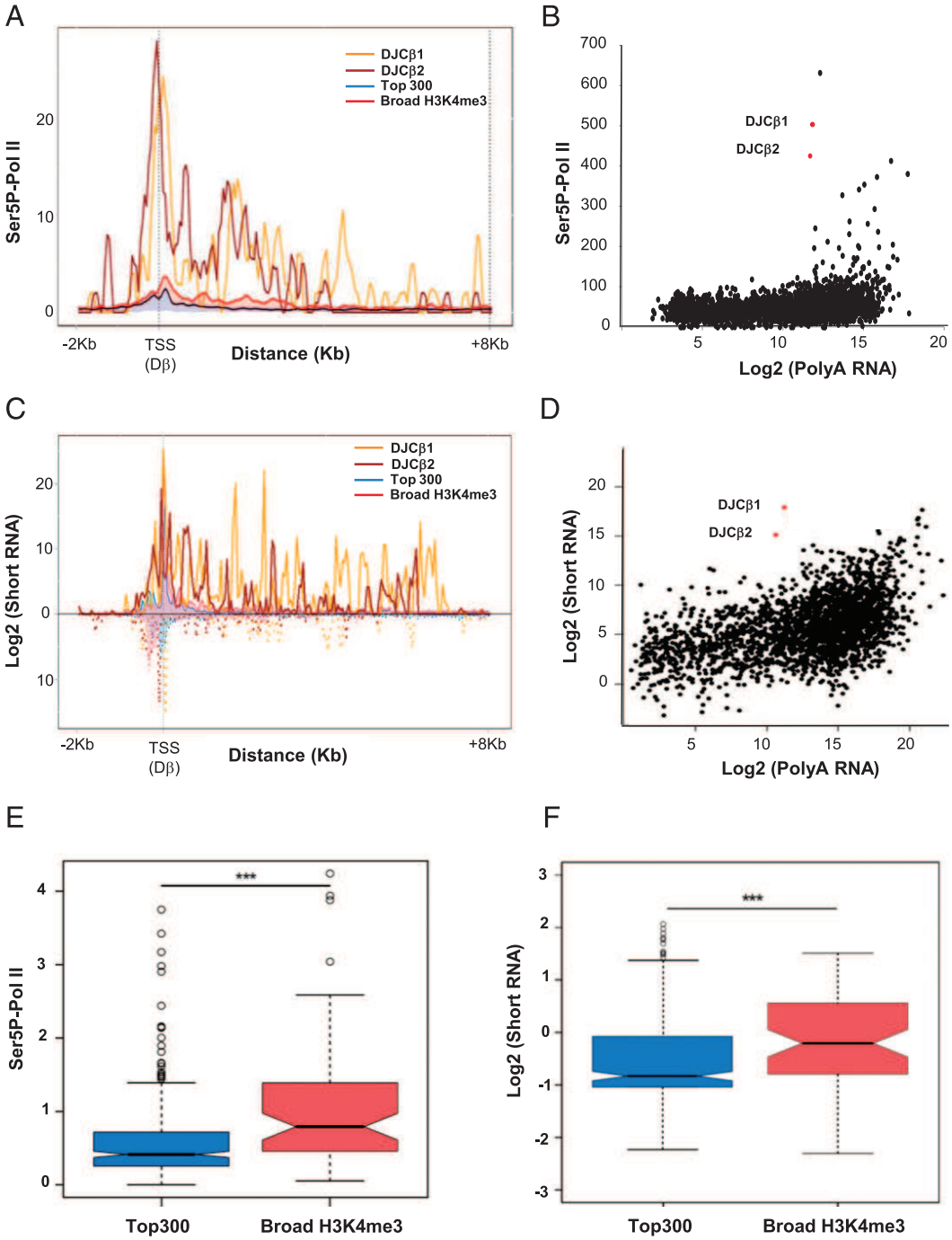



**FIGURE 4.** Global analyses of transcription initiation features. (**A**) Average profiling of Ser$^{5P}$ Pol II ChIP-seq for the set of highly expressed genes (Top-300) and the set of Broad-H3K4me3 genes, compared with the profiles of the two DJCβ regions. The shadow represents the border of the 25th and 75th percentiles. (**B**) Scatter plot showing the Ser$^{5P}$ Pol II density in the gene body of Refseq genes and in the two DJCβ clusters in the function of polyA RNA levels. (**C**) Average profiling of short-RNA–seq for the set of highly expressed genes (Top-300) and the set of Broad-H3K4me3 genes, compared with the profiles of the two DJCβ regions. (**D**) Scatter plot showing the short-RNA–seq density in the gene body of Refseq genes and in the two DJCβ clusters in the function of polyA RNA levels. (**E** and **F**) Boxplot showing the distribution of Ser$^{5P}$ Pol II and short-RNA densities (between +1 and +4 kb from the TSS) of the Top-300 and Broad-H3K4me3 gene sets. Statistical significance was assessed by Student *t* test. ***$p < 10^{-3}$.

Pol II pausing occurs at different places downstream of the Dβ promoters (Fig. 4C, 4D). This was a specific feature of the *Tcrb* locus, as the overall distribution of short-RNAs was clearly different between the DJCβ regions and the set of highly expressed genes, for which bidirectional short-RNAs accumulate around the TSS (Fig. 4C, 4D). Previously we have identified TIPs (14), which are large genomic regions associated with Ser$^{5P}$ Pol II and TBP. TIPs were also associated with high levels of H3K4me3. The *Tcrb* might represent an extreme example of these genomic features. We concluded that the entire DJCβ regions behave as transcription initiating and early elongating platforms, thus providing a direct link between Pol II–mediated chromatin remodeling and H3K4 trimethylation at the Dβ/Jβ recombination segments.

### *Shared features between* Tcrb *and Broad-H3K4me3 genes*

As mentioned above, a small subset of genes was found to be associated with broad H4K4me3 marking (Fig. 1D–F). These

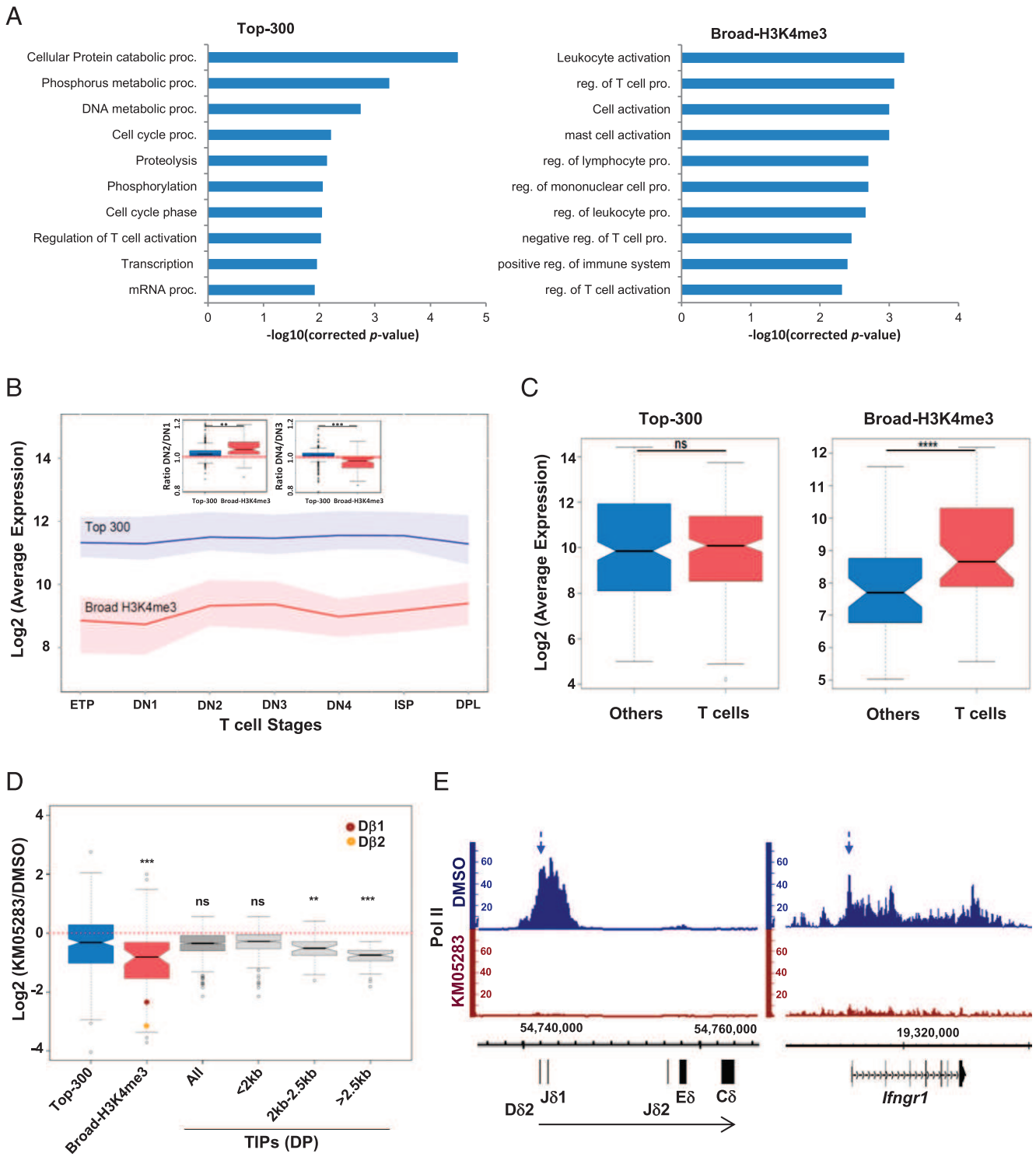



**FIGURE 5.** Characterization of Broad-H3K4me3 genes. (**A**) Enrichment scores of top 10 biological processes found enriched in the indicated gene sets. (**B**) Expression dynamic of Top-300 and Broad-H3K4me3 genes during T cell differentiation. Cell populations are as defined by the Immunological Genome Project and are ordered from less to most differentiated stages: ETP, early thymic progenitors; DN1–4, DN 1–4; ISP, immature single positive; and DPL, DP Blast, respectively. The *inset* shows the differential expression between the DN1-to-DN2 and DN3-to-DN4 transitions. (**C**) Boxplot showing the mean expression distribution in T cell samples and 69 other samples (see *Materials and Methods* for details). (**D**) Boxplot showing the distribution of the KM05283/DMSO ratio of Pol II binding at the promoters (±2 kb around the TSS) of the Top-300, Broad-H3K4me3, and TIPs-associated genes, as well as size-selected subsets of TIPs. The KM05283/DMSO ratios of the two DJCβ regions are also indicated. Statistical significance in (B) through (D) was calculated by the Student *t* test. (**E**) Profiles of Pol II ChIP-seq from the P5424 cell line treated with KM05283 or DMSO at the *Tcrd* locus and one Broad-H3K4me3 gene. The genomic location of primer sets used in Fig. 2G are indicated by dashed arrows. The *p* values are as follows: $^{**}p < 10^{-2}$, $^{***}p < 10^{-3}$, $^{****}p < 10^{-4}$. Proc., process; reg., regulation.

genes also displayed significant enrichment of $Ser^{5P}$ Pol II and, to a lesser extent, short initiating transcripts (Fig. 4E, 4F, Supplemental Fig. 1A, 1B). In general, genes with high levels of H3K4me3 also displayed high levels of $Ser^{5}$ Pol II (Supplemental Fig. 1C). Therefore, a small subset of genes with broad H3K4me3 marking also displays features of transcriptional initiation in ΔRag

thymocytes (although *Tcrb* might represent an extreme example of this phenomenon).

To gain further insight into the function of Broad-H3K4me3 genes, we analyzed the functional enrichment of the biological process and found that they were specifically enriched on T cell– and immune-related functions, whereas the set of Top-300 genes were enriched for metabolic processes (Fig. 5A). Indeed, the list of Broad-H3K4me3 genes include many genes known to be involved in T cell differentiation and signaling, such as *Lef1, Il2ra, Themis, Ifngr1, Fyb, RhoH,* and *Cd274* (Supplemental Table II). Accordingly, the set of Broad-H3K4me3 genes was highly tissue specific (Fig. 5B). Although these genes were expressed at relatively low levels in primary thymocytes, their expression was highly regulated during early T cell differentiation (Fig. 5C), namely, between DN1-to-DN2 and DN3-to-DN4 cell transitions (Fig. 4C, *insets*). Thus, the subset of Broad-H3K4me3 genes is reminiscent of the *Tcrb* locus, as they represent highly regulated genes involved in T cell function. They might represent extreme examples of genes with broad H3K4 methylation patterns described previously by us and others (12–14). To assess whether other AR genes could share the same features as the *Trcb*, we analyzed, in a similar way, gene segments of the *Tcrd* and *Tcrg* locus, which are the two other AR loci in an open chromatin configuration in ΔRag thymocytes (see *Materials and Methods* for details). We found that gene segments from *Tcrd* (spanning Dδ2-Jδ1 gene segments) and *Tcrg* (Jγ1-Cγ1 and Jγ4-Cγ4 gene segments) loci also displayed high levels of H3K4me3, Ser$^{5P}$ Pol II, and short initiating transcripts to a similar extent as those observed for the *Tcrb* locus (Supplemental Figs. 1C, 2), thus suggesting that large initiating platforms might be a general feature of AR loci.

Finally, we asked whether Pol II binding at Broad-H3K4me3 genes was also highly sensitive to transcription elongation, as observed for the *Tcrb* locus. Quantification of Pol II levels around the TSS of the Top-300 and Broad H3K4me3 genes in P5424 cells treated with either DMSO or KM05283 demonstrated that Pol II binding is specifically lost at BroadH4K4me3 genes, although not to the same extent as observed around the Dβ gene segments. This finding was evidenced at several genes, including the *Tcrd* and *Infgr1* loci (Fig. 5E), and validated by independent ChIP-qPCR (Fig. 2G; note that the *Tcrg* locus could not be analyzed, as this gene was found to be inactive in the P5424 cell line; data not shown). To determine whether this phenomenon was a general property of TIPs, we analyzed our previously defined selection of TIPs-associated genes in DP thymocytes (14), excluding the genes that were not expressed in the P5424 cell line (see *Materials and Methods* for details). As a group, the TIPs-associated genes did not display a loss of Pol II binding at their promoters after inhibition of Pol II elongation (Fig. 5D). However, when TIPs were classified according to their size, we found that genes associated with large TIPs (>2.5 kb) significantly lost Pol II binding at their promoters (Fig. 5D). We concluded that a subset of Broad-H3K4me3 genes and large TIPs-associated genes display regulatory features similar to those of the *Tcrb* locus, including tissue-specific gene expression, the presence of a TIP, and coupled Pol II recruitment and elongation (Fig. 6).

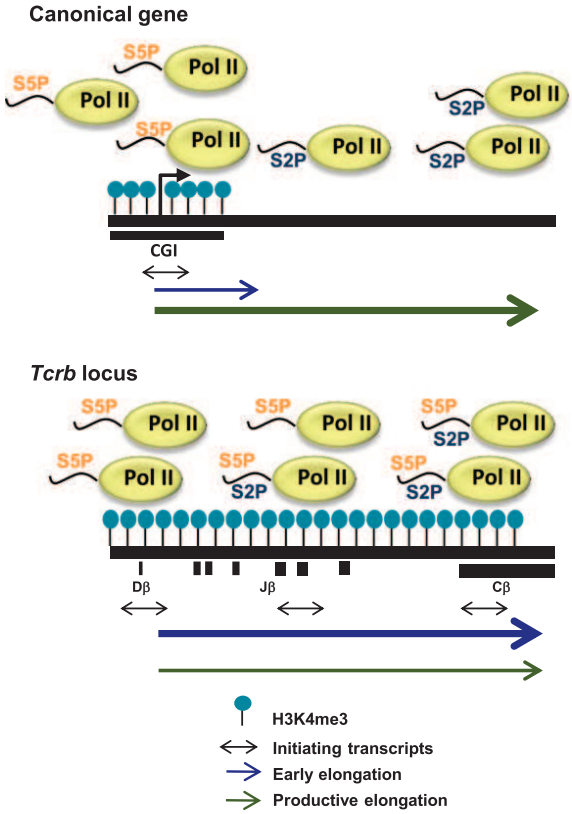



**FIGURE 6.** Model summarizing Pol II transcription at canonical expressed genes and DJCβ clusters. Promoters of active genes (*top panel*) are generally linked to CpG islands (CGI) and associated with high levels of initiating (Ser$^{5P}$) Pol II, bidirectional short transcripts, and H3K4me3. In contrast, at the *Tcrb* locus, transcription initiation hallmarks and H3K4me3 are found throughout the transcribed regions spanning the Dβ and Jβ gene segments. This feature might explain the unusual high chromatin accessibility observed at these regions.

## Discussion

Previous work from our laboratory and other laboratories has shown a remarkable open chromatin structure encompassing the Dβ-Jβ recombination center, including chromatin accessibility and histone marking (3, 16, 22, 26–28). More specifically, H3K4me3 was found to be enriched at Dβ and Jβ segments using ChIP-qPCR (7). Similar extended H3K4me3 patterns have been shown across the Jα segments of the *Tcra* locus (42). In the current study, we extend these finding by showing that the distribution of H3K4 trimethylation over the DβJβ regions is both quantitatively and qualitatively different from that in the vast majority of expressed genes (Fig. 1). Although H3K4me3 generally accumulates within 2kb around the TSS of genes (11), we observed that H3K4me3 enrichment at the DJCβ clusters is much broader, spanning ≤8 kb downstream of the germline Dβ promoters, and including all Jβ gene segments. Moreover, the level of H3K4 trimethylation found at the DJCβ clusters was exceptionally high, representing one of the most enriched domains in developing thymocytes. We show that this extended profile depends on an unusual Pol II regulation process. In the case of canonical genes, Pol II accumulates around the TSS in its initiating form (high Ser$^{5P}$ Pol II), which correlates with high enrichment of H3K4me3 and the presence of short-RNA transcripts. However, in the case of the *Tcrb* locus, the entire DJCβ regions display features of transcription initiation and Pol II pausing, including high levels of Ser$^{5P}$ Pol II and short-RNAs. Unexpectedly, inhibition of Pol II elongation resulted in complete loss of Pol II across the DJCβ clusters (Fig. 2). To our knowledge, this is the first example in mammals whereby Pol II accumulation at the promoter is strictly dependent on transcription elongation. Remarkably, this phenomenon was also observed at the *Tcrd* locus (Figs. 2G, 5E). We propose that a high level of initiating Pol II throughout the entire DJCβ regions targets the H3K4 histone methyltransferases, resulting in an unusual extended H3K4me3 profile, and ultimately

leads to a highly accessible chromatin structure around the Dβ and Jβ gene segments (Fig. 6).

We have previously shown that tissue-specific genes expressed in T cells generally display high levels of H3K4 methylation within the 5′ region of the gene body (12). Along the same line, a recent study has shown that H3K4me3 domains that spread more broadly over genes in a given cell type preferentially mark genes that are essential for the identity and function of that cell type (13). Besides, we also described TIPs at proximal and distal sites, which were characterized by the presence of $Ser^{5P}$ Pol II, TBP, and epigenetic marks H3K4me1 and H3K4me3 (14). In this article, we show that genes with broad H3K4me3 domains display features related to large initiation platforms (including accumulation of $Ser^{5P}$ Pol II and short initiating transcripts) similar to the TIP genomic domains. However, TIP domains as defined previously in DP thymocytes (14) display a wide range of size, varying from 0.45 kb to 10 kb (80% of TIPs are <2 kb). Whether broad H3K4me3 and TIPs define the same type of genes remains to be precisely investigated, but our results suggest that common features are shared by both types of structures. Genes marked by the broadest H3K4me3 domains exhibit enhanced transcriptional consistency rather than increased transcriptional levels (13). Moreover, Pol II accumulation at the promoter of Broad-H3K4me3 genes tends to be dependent on transcription elongation, a phenomenon also observed at the promoters of genes associated with large TIPs (Fig. 5D). Thus, it is likely that the broad H3K4me3 domains defined in this article (in particular, those found at the *Tcr* loci) might represent a subset of larger TIPs. Indeed, larger TIPs also have a tendency to be more tissue specific (14). All in all, our results suggest the existence of a specialized transcriptional regulation mechanism restricted to a subset of tissue-specific genes. In this context, the *Tcrb* locus might represent an extreme example of this phenomenon. Our finding has implications not only for regulatory strategies used by AR loci but also for the epigenetic mechanisms that control gene expression of cell identity genes.

Are intrinsic genomic features responsible for the highly open and H3K4me3-enriched chromatin structure observed at the *Tcrb* locus? In mammals, Pol II accumulation and enrichment for active histone marks at promoters are generally linked to their high CpG content (49). We have previously shown that TIPs overlapped with CpG density, although larger TIPs displayed lower or more disperse CpG content (14). Consistently, we found that promoters of Broad-H3K4me3 and large TIPs-associated genes display significantly lower CpG density as compared with the set of Top-300 genes ($p < 0.01$ and $p < 0.0001$, respectively; Student *t* test; see *Materials and Methods*). The DJCβ regions do not contain any CpG island and also display relatively low G and C nucleotide content (data not shown). It is, therefore, plausible that in the absence of CpG islands, the Pol II molecules recruited at the Dβ associated promoters are immediately engaged in the elongation process while still harboring the transcription initiation mark (i.e., $Ser^{5P}$) and therefore remain associated with H3K4 methyltransferases (43, 44)(Fig. 6). This hypothesis would be consistent with the complete loss of Pol II at the *Tcrb* locus after inhibition of transcription elongation (Fig. 2). Another intriguing, but not mutually exclusive, possibility is that the extended H3K4me3 profile is related to the unusual structure of the *Tcrb* locus, which contains several J segments, each harboring a 5′ splicing site. A recent study has shown that H3K4 trimethylation at the 5′ border of mammalian genes is directly linked to the length of the first exon of genes (average size is 250 nt) (50). However, in the case of the DJCβ transcription units, the first splicing donors are located at the end of each Jβ segment, ranging between 641 nt and 2.5 kb from the Dβ segments, which make the first exons considerably longer than the average size. Moreover, the Jβ-associated splicing sites appear to be relatively inefficient, as judged by the high level of RNA-seq signal observed downstream of the Jβ gene segments (Fig. 3B). As described previously (50), the first exon length >500 nt results in a flat H3K4me3 profile extending to the 3′ end of the first exon, as well as increasing Pol II pausing, both features reminiscent of what is observed at the *Tcrb* locus. Thus, it is plausible that the location of Jβ gene segments, each behaving as a first exon, will result in the distinctive chromatin structure observed at the DJCβ clusters.

It has been recently demonstrated that RAG1 and RAG2 bind in vivo to focal regions, termed "recombination centers," covering mainly the J segments of AR genes and within which V(D)J recombination has been suggested to take place (7). The formation of these recombination centers depends on the AR enhancers and promoters (6), and correlates with the presence of H3K4me3 (7). Thus, given the specific requirements for chromatin accessibility and H3K4me3 enrichment at J segments to ensure efficient V(D)J recombination (3, 5), we propose that the *Tcrb* locus (and likely other AR loci) has evolved in such a way that a specialized regulation of the transcription process confers a unique long-range epigenetic marking, ultimately allowing the establishment of a highly accessible chromatin structure at the recombining Dβ/Jβ gene segments.

### *Data access*

ChIP-seq and RNA-seq data obtained in this study have been submitted to the National Center for Biotechnology Information's GEO (http://www.ncbi.nlm.nih.gov/geo) under the following accession numbers: GSE63416 (www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE63416), GSE64709 (www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE64709), and from GSM1360722 to GSM1360727 and GSM1359828 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56395). Details are available in Supplemental Table I.

## Acknowledgments

We thank Dr. Eugene Oltz (Washington University, St. Louis, MO) for donating the P5424 cell line.

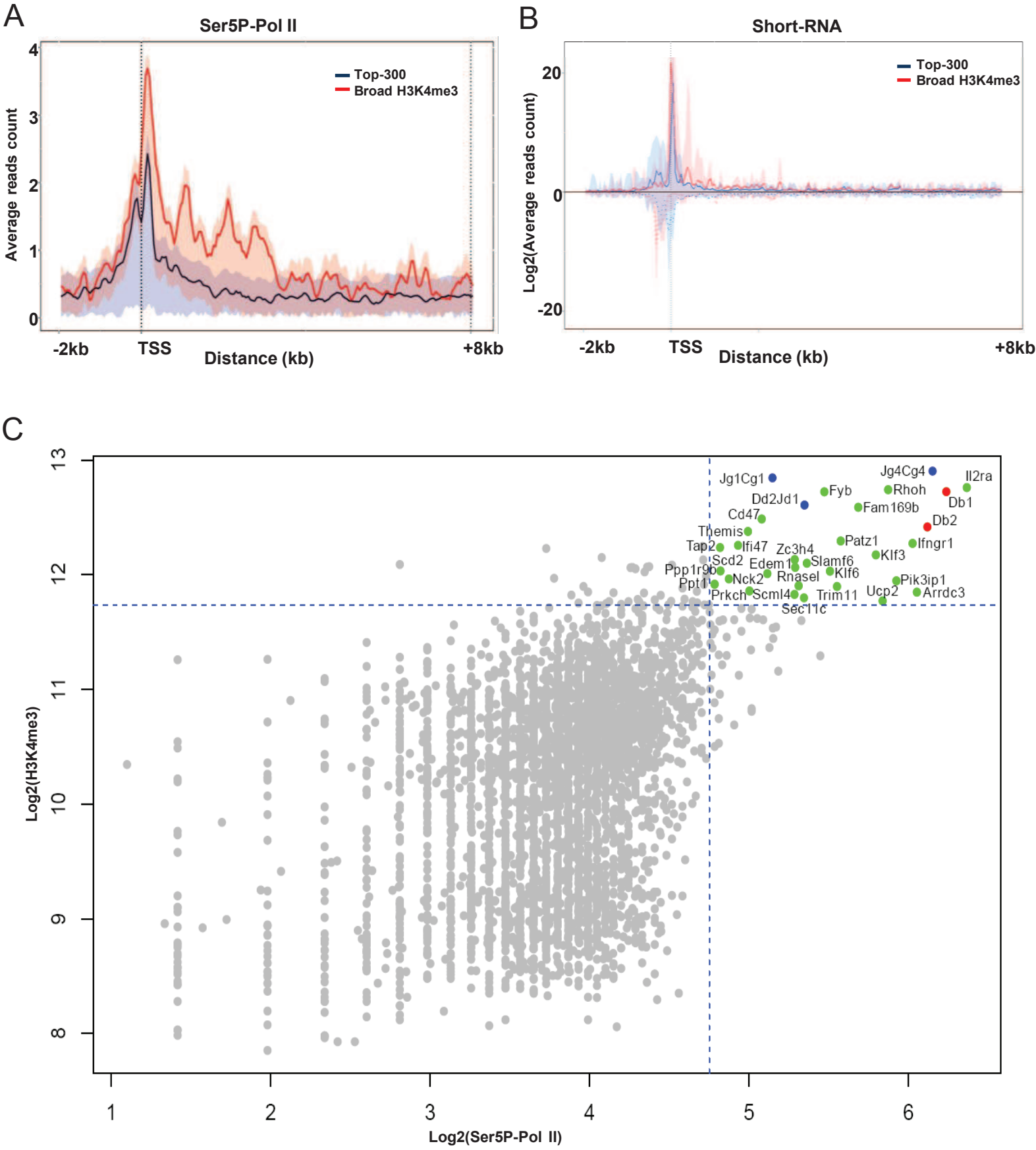
## Disclosures

The authors have no financial conflicts of interest.

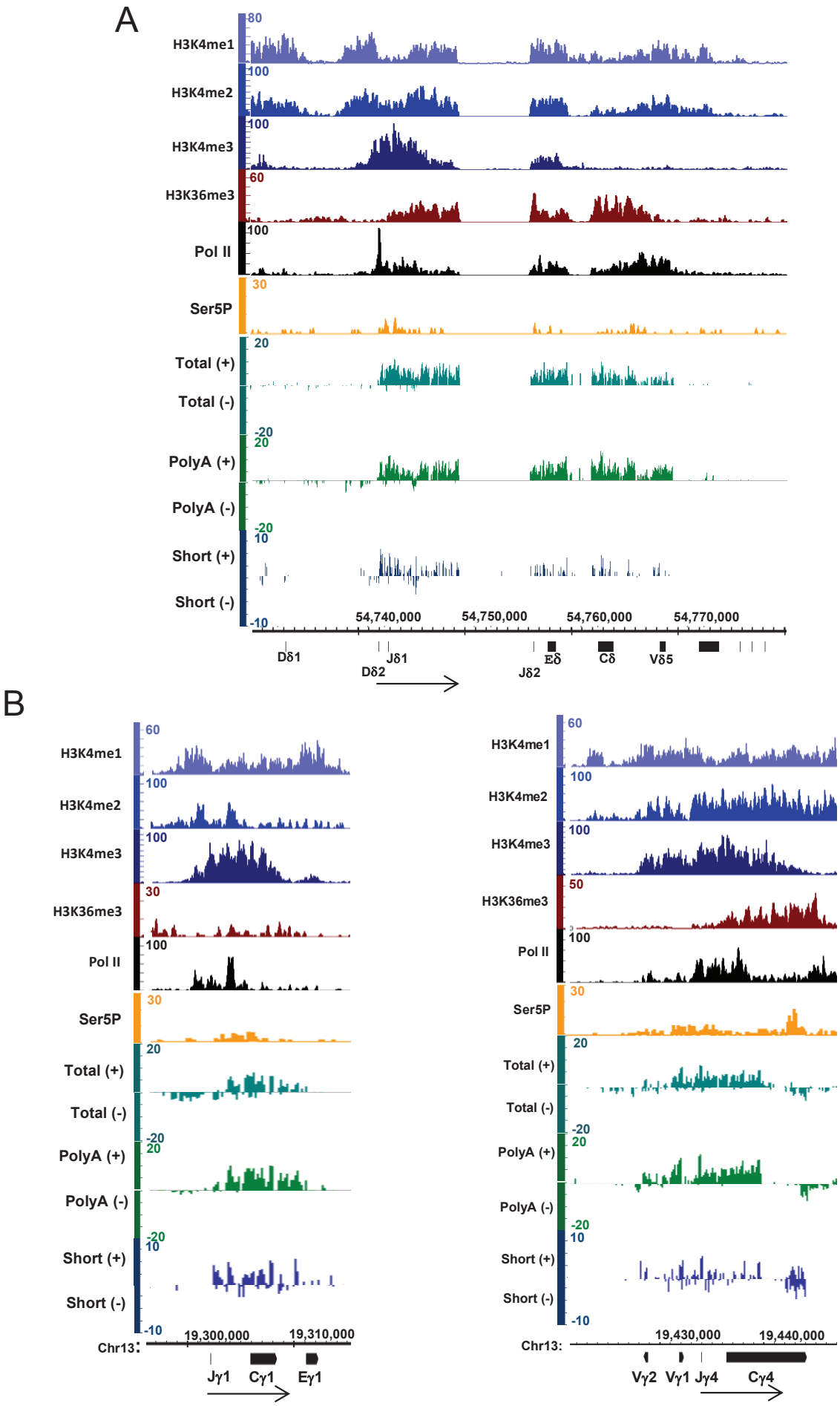
## References

1. Schatz, D. G. 2004. V(D)J recombination. *Immunol. Rev.* 200: 5–11.
2. Jung, D., and F. W. Alt. 2004. Unraveling V(D)J recombination; insights into gene regulation. *Cell* 116: 299–311.
3. Osipovich, O., and E. M. Oltz. 2010. Regulation of antigen receptor gene assembly by genetic-epigenetic crosstalk. *Semin. Immunol.* 22: 313–322.
4. Yancopoulos, G. D., and F. W. Alt. 1986. Regulation of the assembly and expression of variable-region genes. *Annu. Rev. Immunol.* 4: 339–368.
5. Spicuglia, S., J. Zacarias-Cabeza, A. Pekowska, and P. Ferrier. 2010. Epigenetic regulation of antigen receptor gene rearrangement. *F1000 Biol. Rep.* 2: 2.
6. Ji, Y., A. J. Little, J. K. Banerjee, B. Hao, E. M. Oltz, M. S. Krangel, and D. G. Schatz. 2010. Promoters, enhancers, and transcription target RAG1 binding during V(D)J recombination. *J. Exp. Med.* 207: 2809–2816.
7. Ji, Y., W. Resch, E. Corbett, A. Yamane, R. Casellas, and D. G. Schatz. 2010. The in vivo pattern of binding of RAG1 and RAG2 to antigen receptor loci. *Cell* 141: 419–431.
8. Matthews, A. G., and M. A. Oettinger. 2009. RAG: a recombinase diversified. *Nat. Immunol.* 10: 817–821.
9. Abarrategui, I., and M. S. Krangel. 2009. Germline transcription: a key regulator of accessibility and recombination. *Adv. Exp. Med. Biol.* 650: 93–102.
10. Zhou, V. W., A. Goren, and B. E. Bernstein. 2011. Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.* 12: 7–18.
11. Barski, A., S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* 129: 823–837.
12. Pekowska, A., T. Benoukraf, P. Ferrier, and S. Spicuglia. 2010. A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Res.* 20: 1493–1502.

13. Benayoun, B. A., E. A. Pollina, D. Ucar, S. Mahmoudi, K. Karra, E. D. Wong, K. Devarajan, A. C. Daugherty, A. B. Kundaje, E. Mancini, et al. 2014. H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell* 158: 673–688.
14. Koch, F., R. Fenouil, M. Gut, P. Cauchy, T. K. Albert, J. Zacarias-Cabeza, S. Spicuglia, A. L. de la Chapelle, M. Heidemann, C. Hintermair, et al. 2011. Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat. Struct. Mol. Biol.* 18: 956–963.
15. Jackson, A. M., and M. S. Krangel. 2006. Turning T-cell receptor β recombination on and off: more questions than answers. *Immunol. Rev.* 209: 129–141.
16. Spicuglia, S., A. Pekowska, J. Zacarias-Cabeza, and P. Ferrier. 2010. Epigenetic control of Tcrb gene rearrangement. *Semin. Immunol.* 22: 330–336.
17. Michie, A. M., and J. C. Zúñiga-Pflücker. 2002. Regulation of thymocyte differentiation: pre-TCR signals and beta-selection. *Semin. Immunol.* 14: 311–323.
18. Krimpenfort, P., R. de Jong, Y. Uematsu, Z. Dembic, S. Ryser, H. von Boehmer, M. Steinmetz, and A. Berns. 1988. Transcription of T cell receptor beta-chain genes is controlled by a downstream regulatory element. *EMBO J.* 7: 745–750.
19. McDougall, S., C. L. Peterson, and K. Calame. 1988. A transcriptional enhancer 3′ of C beta 2 in the T cell receptor beta locus. *Science* 241: 205–208.
20. Bories, J. C., J. Demengeot, L. Davidson, and F. W. Alt. 1996. Gene-targeted deletion and replacement mutations of the T-cell receptor beta-chain enhancer: the role of enhancer elements in controlling V(D)J recombination accessibility. *Proc. Natl. Acad. Sci. USA* 93: 7871–7876.
21. Bouvier, G., F. Watrin, M. Naspetti, C. Verthuy, P. Naquet, and P. Ferrier. 1996. Deletion of the mouse T-cell receptor beta gene enhancer blocks alphabeta T-cell development. *Proc. Natl. Acad. Sci. USA* 93: 7877–7881.
22. Mathieu, N., W. M. Hempel, S. Spicuglia, C. Verthuy, and P. Ferrier. 2000. Chromatin remodeling by the T cell receptor (TCR)-β gene enhancer during early T cell development: Implications for the control of TCR-β locus recombination. *J. Exp. Med.* 192: 625–636.
23. Spicuglia, S., S. Kumar, J. H. Yeh, E. Vachez, L. Chasson, S. Gorbatch, J. Cautres, and P. Ferrier. 2002. Promoter activation by enhancer-dependent and -independent loading of activator and coactivator complexes. *Mol. Cell* 10: 1479–1487.
24. Ryu, C. J., B. B. Haines, D. D. Draganov, Y. H. Kang, C. E. Whitehurst, T. Schmidt, H. J. Hong, and J. Chen. 2003. The T cell receptor β enhancer promotes access and pairing of Dbeta and Jbeta gene segments during V(D)J recombination. *Proc. Natl. Acad. Sci. USA* 100: 13465–13470.
25. Oestreich, K. J., R. M. Cobb, S. Pierce, J. Chen, P. Ferrier, and E. M. Oltz. 2006. Regulation of TCRbeta gene assembly by a promoter/enhancer holocomplex. *Immunity* 24: 381–391.
26. Bonnet, M., F. Huang, T. Benoukraf, O. Cabaud, C. Verthuy, A. Boucher, S. Jaeger, P. Ferrier, and S. Spicuglia. 2009. Duality of enhancer functioning mode revealed in a reduced TCR β gene enhancer knockin mouse model. *J. Immunol.* 183: 7939–7948.
27. Whitehurst, C. E., S. Chattopadhyay, and J. Chen. 1999. Control of V(D)J recombinational accessibility of the D beta 1 gene segment at the TCR beta locus by a germline promoter. *Immunity* 10: 313–322.
28. Morshead, K. B., D. N. Ciccone, S. D. Taverna, C. D. Allis, and M. A. Oettinger. 2003. Antigen receptor loci poised for V(D)J rearrangement are broadly associated with BRG1 and flanked by peaks of histone H3 dimethylated at lysine 4. *Proc. Natl. Acad. Sci. USA* 100: 11577–11582.
29. Shinkai, Y., G. Rathbun, K. P. Lam, E. M. Oltz, V. Stewart, M. Mendelsohn, J. Charron, M. Datta, F. Young, A. M. Stall, and F. W. Alt. 1992. RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V(D)J rearrangement. *Cell* 68: 855–867.
30. Mombaerts, P., C. Terhorst, T. Jacks, S. Tonegawa, and J. Sancho. 1995. Characterization of immature thymocyte lines derived from T-cell receptor or recombination activating gene 1 and p53 double mutant mice. *Proc. Natl. Acad. Sci. USA* 92: 7420–7424.
31. Medlin, J., A. Scurry, A. Taylor, F. Zhang, B. M. Peterlin, and S. Murphy. 2005. P-TEFb is not an essential elongation factor for the intronless human U2 snRNA and histone H2b genes. *EMBO J.* 24: 4154–4165.
32. Chapman, R. D., M. Heidemann, T. K. Albert, R. Mailhammer, A. Flatley, M. Meisterernst, E. Kremmer, and D. Eick. 2007. Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. *Science* 318: 1780–1782.
33. Giresi, P. G., J. Kim, R. M. McDaniell, V. R. Iyer, and J. D. Lieb. 2007. FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Res.* 17: 877–885.
34. Pekowska, A., T. Benoukraf, J. Zacarias-Cabeza, M. Belhocine, F. Koch, H. Holota, J. Imbert, J. C. Andrau, P. Ferrier, and S. Spicuglia. 2011. H3K4 trimethylation provides an epigenetic signature of active enhancers. *EMBO J.* 30: 4198–4210.
35. Lepoivre, C., M. Belhocine, A. Bergon, A. Griffon, M. Yammine, L. Vanhille, J. Zacarias-Cabeza, M. A. Garibal, F. Koch, M. A. Maqbool, et al. 2013. Divergent transcription is associated with promoters of transcriptional regulators. *BMC Genomics* 14: 914.
36. Rahl, P. B., C. Y. Lin, A. C. Seila, R. A. Flynn, S. McCuine, C. B. Burge, P. A. Sharp, and R. A. Young. 2010. c-Myc regulates transcriptional pause release. *Cell* 141: 432–445.
37. Benoukraf, T., P. Cauchy, R. Fenouil, A. Jeanniard, F. Koch, S. Jaeger, D. Thieffry, J. Imbert, J. C. Andrau, S. Spicuglia, and P. Ferrier. 2009. CoCAS: a ChIP-on-chip analysis suite. *Bioinformatics* 25: 954–955.
38. Heng, T. S., M. W. Painter, Immunological Genome Project Consortium. 2008. The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.* 9: 1091–1094.
39. Huber, W., A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18(Suppl 1): S96–S104.
40. Dennis, G., Jr., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* 4: 3.
41. Carabana, J., A. Watanabe, B. Hao, and M. S. Krangel. 2011. A barrier-type insulator forms a boundary between active and inactive chromatin at the murine TCRβ locus. *J. Immunol.* 186: 3556–3562.
42. Abarrategui, I., and M. S. Krangel. 2006. Regulation of T cell receptor-α gene recombination by transcription. *Nat. Immunol.* 7: 1109–1115.
43. Selth, L. A., S. Sigurdsson, and J. Q. Svejstrup. 2010. Transcript elongation by RNA polymerase II. *Annu. Rev. Biochem.* 79: 271–293.
44. Buratowski, S. 2009. Progression through the RNA polymerase II CTD cycle. *Mol. Cell* 36: 541–546.
45. Osipovich, O., R. M. Cobb, K. J. Oestreich, S. Pierce, P. Ferrier, and E. M. Oltz. 2007. Essential function for SWI-SNF chromatin-remodeling complexes in the promoter-directed assembly of Tcrb genes. *Nat. Immunol.* 8: 809–816.
46. Guenther, M. G., S. S. Levine, L. A. Boyer, R. Jaenisch, and R. A. Young. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* 130: 77–88.
47. Heidemann, M., C. Hintermair, K. Voß, and D. Eick. 2013. Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription. *Biochim. Biophys. Acta* 1829: 55–62.
48. Seila, A. C., J. M. Calabrese, S. S. Levine, G. W. Yeo, P. B. Rahl, R. A. Flynn, R. A. Young, and P. A. Sharp. 2008. Divergent transcription from active promoters. *Science* 322: 1849–1851.
49. Suzuki, M. M., and A. Bird. 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* 9: 465–476.
50. Bieberstein, N. I., F. Carrillo Oesterreich, K. Straube, and K. M. Neugebauer. 2012. First exon length controls active chromatin signatures and transcription. *Cell Reports* 2: 62–68.




**Supplemental Figure 1**: (A-B) Average profiling of Ser5P-PolII ChIP-seq (A) and Short-RNA-seq (B) in ΔRag thymocytes at genomic regions from -2kb to +8kb around the TSS of the set of highly expressed genes (Top-300) and the set of Broad-H3K4me3 associated genes. The shadow represents the border of the 25$^{th}$ and 75$^{th}$ percentiles. (C) Scatter showing the H3K4me3 density in function of Ser5P-Pol II density in the gene body of Refseq genes. The genes with high levels of both H3K4me3 and Ser5P-PolII are highlighted in green. The *Tcrb, d and g* gene clusters are also shown.




**Supplemental Figure 2:** ChIP-seq profiles of histone marks (H3K4me1, 2, 3 and H3K36me3), total and initiating (Ser5P) Pol II, as well as, Total, PolyA and Short RNAs from directional RNA-seq experiments (middle and low tracks) in ΔRag thymocytes for *Tcrd* (A) and *Tcrg* (B) gene clusters. Horizontal arrows indicate the position of the transcription units analyzed in Supplemental Fig. 1C.

**Supplementary Table 1.** Information about ChIP-seq and RNA-seq data generated in this study and submitted to the NCBI Gene Expression Omnibus. The number of mapped reads and the estimated size of DNA fragments are indicated

| Sample | # mapped reads | Estimated fragment size (bp) | GEO ID |
|---|---|---|---|
| **ChIP-seq in the P5424 cell line** | | | |
| H3K4me3 (DMSO) | 29,109,407 | 146 | GSE63416 |
| Polymerase II (DMSO) | 24,534,212 | 176 | |
| H3K4me3 (KM05283) | 26,636,829 | 146 | |
| Pol II (KM05283) | 24,425,084 | 176 | |
| **ChIP-seq in Rag2-/- thymocytes** | | | |
| H3K4me1 Biorep1 | 22,898,035 | 146 | GSM1360722 |
| H3K4me1 Biorep2 | 23,892,863 | 146 | GSM1360723 |
| H3K4me2 Biorep1 | 32,913,113 | 126 | GSM1360724 |
| H3K4me2 Biorep2 | 30,651,343 | 126 | GSM1360725 |
| H3K4me3 Biorep1 | 30,644,661 | 136 | GSM1360726 |
| H3K4me3 Biorep2 | 33,685,813 | 136 | GSM1360727 |
| Ser5P-Pol II | 23,040,876 | 156 | GSE64709 |
| **RNA-seq in Rag2-/- thymocytes** | | | |
| Short RNA | 72,126,257 | n.a | GSM1359828 |

# Article 2:

**L'étude de la dynamique des domaines larges de H3K4me3 dans la différenciation T et les leucémies.**

Comme décrit auparavant, les domaines larges de la méthylation de H3K4 (BH4D) ont été associés à des gènes spécifiques de l'identité cellulaire et des suppresseurs de tumeur. Ainsi, dans ce projet, nous avons utilisés des données générées par mes deux laboratoires de tutelle dans le cadre du projet européen ***Blueprint*** (http://www.blueprint-epigenome.eu/) afin d'identifier les gènes associés à la marque H3K4me3 dans des populations thymiques humaines, une collection de lignées cellulaires leucémiques T et d'échantillons primaires de leucémie lymphoblastique aiguë T. Nous avons constaté que les BH4D présentent une dynamique tout le long de la différenciation des lymphocytes T, sont associés aux oncogènes spécifiques de chaque leucémie et distinguent les cellules normales et leucémiques.

En effet, nous avons constaté que les cellules leucémiques perdent les domaines larges associés aux gènes de l'identité cellulaire T et gagnent de larges domaines dans les gènes liés au cancer. De plus, l'expression des gènes codant et non codant associées aux BH4D est souvent dérégulée dans les leucémies. Enfin, nous montrons que l'expression ectopique du facteur oncogène TLX3 a un impact sélectif sur l'expression des gènes associés aux BH4D.

Nos résultats montrent que la régulation des gènes associés aux BH4D est importante pour la leucémogenèse et suggèrent que l'association avec BH4D pourrait être utilisée comme un outil de priorisation épigénétique des gènes pertinents chez les cancers, y compris les ARN longs non codants (voir Article 4).

**Contributions dans l'article :**

- Réalisation de toutes les analyses bioinformatiques ;
- Participation à l'écriture de l'article ;
- Réalisation de toutes les figures de l'article.

# Genes associated with H3K4me3-broad domains are preferentially deregulated in leukemia and flag key oncogenes

**Author list:** Mohamed Belhocine, Eve-Lyne Mathieu, Agata Cieslak, Laurent Vanhille, Elizabeth Macintyre, Denis Puthier, Vahid Asnafi* and Salvatore Spicuglia*

**Abstract**

Broad domains of H3K4 methylation have been associated with tissue specific, cell identity and tumor suppressor genes. We identify broad H3K4me3 domains (BH4D)-associated genes in human thymic T cell populations and a collection of cell lines and T-Acute Lymphoblastic Leukemia (T-ALL) primary samples. We found that BH4D are highly dynamic throughout T cell differentiation, flag key leukemia associated oncogenes and distinguish between normal and neoplastic cells. Leukemic cells loss broad domains at T cell identity genes and gain broad domains at cancer-related genes. Moreover, the expression of genes (both coding and non-coding) associated with BH4D is frequently deregulated in T-ALL. Finally, we show that ectopic expression of TLX3 oncogenic factor selectively impact on the expression of BH4D-associated genes. Our results show that regulation of BH4D-associated genes is critical for leukemogenesis and suggest that association with BH4D might be used as epigenetic prioritization of cancer-relevant genes, including long non-coding RNAs.

**Highlights:**

- Broad H3K4me3 domains (BH4Ds) mark key T-ALL oncogenes
- BH4D-associated genes (both coding and non-coding) are more frequently deregulated in leukemia
- Leukemic cells loss BH4D at T cell related genes and gain BH4D at cancer-related genes
- Some T-ALL samples have abnormal numbers of BH4D
- Ectopic expression of an oncogenic transcription factor selectively impact on the expression of BH4D-associated genes

**Results**

## Identification of broad H3K4me3 domains (BH4D) in human thymic subpopulations

We first compared to methods to isolate genes associated with broad H3K4me3 domains (**Fig. Sup. 1**). The first method was based on density coverage of H3K4me3 on the 5' side of the gene bodies (**Fig. Sup. 1A**). The second method was based on the association of broad H3K4me3 peaks (MACS tool) with the TSS of genes (**Fig. Sup. 1B**). We found that broad H3K4me3 domains identified with MACS provided a higher enrichment for tissue-specific gene functions (**Fig. Sup. 1E-F**) and provided highly

reproducible results between two biological replicates **(Fig. Sup. 1D)**. We therefore used this method to identify broad H3K4me3 associated genes either coding or non-coding (LncRNAs) in several human thymic sub-populations and several T acute lymphoblastic samples (**Fig. Sup. 2**).

Broad H3K4me3 domains were highly dynamic thought-out T cell differentiation (**Fig. 1A**) and associated with changes in gene expression (**Fig. 1B**). This is highlighted by the epigenomic profiles of several key T cell genes (**Fig. 1C**).

**Broad domains flag key oncogenes in TALL**

We first compared numbers and size of broad domains between normal thymic sub-populations and several primary and cell line T-ALL). We observed that the number of broad domains was more variable among the T-ALL samples as compared to normal cells, with some extreme cases of very low (e.g. T-ALL16, TALL6 and 3) or high (e.g. Loucy, DND41, TALL5, TALL9) numbers of broad domains **(Fig. 2A and sup Table1)**, pointing out to potential global deregulation of epigenetic factors affecting H3K4me3 spreading.

Interestingly, defined oncogenes, such TLX1, TLX3 and TAL1 were found to be associated with broad domains in the T-ALL cell lines were they are overexpressed (**Fig. Sup. 2C**). Moreover, we found that T-ALL oncogenes were generally associated with broad domains in primary T-ALL **(Fig. 2B)**. Several examples provided clear evidence of H3K4me3 broad domains flagging overexpressed oncogenes in T-ALL **(Fig. 2C)**.

We next compared the number of gain and loss of broad domains between normal and T-ALL samples (**Fig. Sup. 3**). We found a strong bias for gain of broad domains in T-ALL samples. Strikingly, clustering based on broad domains clearly separated thymic subpopulations from T-ALL samples and cell lines (**Fig. 3A**), suggesting that broad domains defined key features of the leukemic biology are highly frequent in TALL. Finally, we analyzed statistical enrichments Go terms at thymic and T-ALL specific broad domains associated genes **(Fig. 3B)**. We found that leukemic cells generally loss broad domains at T cell identity genes and gain broad domains at cancer-related genes.

**Broad genes are preferentially deregulated in leukemia**

To explore whether BH4D-associated genes were preferentially deregulated in T-ALL, we analyses the changes in gene expression between normal T cells (LC) and each of the T-ALL samples for both coding and non-coding (lincRNAs) genes. On the one hand, we selected BH4D-associated genes in normal thymocytes and selected a set of genes with same expression pattern but associated with sharp H3K4me3 profiles around the TSS. We observed that BH4D-associated genes were significantly down-regulated in the T-ALL samples than the control set (**Fig. 4A**). On the other hand, BH4D-associated genes in

each of the T-ALL samples were significantly more up regulated as compared to the control set (**Fig. 4B**).

To more globally assess the relevance of deregulation of gene expression in function of BH4D dynamics, we analyzed the fold change expression between normal thymocytes (LC or thymus) and two independent series of T-ALLs (unpublished data from Necker and (Atak et al., 2013) in function of the recurrence of gain and loss of BH4D. As shown in (**Fig. 5**), the frequency of gain and loss of BH4D strictly correlates with the level of, respectively, up and down regulation for both coding and non-coding genes.

Overall, these results show that BH4D-associated genes are more frequently deregulated in leukemia samples, suggesting that mechanisms affecting the expression of BH4D-associated genes might be highly relevant for the oncogenic transformation. Interestingly, this phenomenon was true for both coding and non-coding genes, suggesting that regulation of LincRNAs associated with BH4D might represent relevant candidate markers of leukemogenesis.

**Independent validation in CLL**

To further address the relevance of our findings we analyzed RNA-seq and H3K4me3 ChIP-seq data of normal B cells and primary Chronic lymphocytic leukemia (CLL) generated by the blueprint consortium. Clustering based on broad domains associated genes clearly separated normal B cells from CLL (**Fig. Sup. 4A**). The fold change expression between CLL and naïve B cells was significantly higher for genes associated with broad domains as compared with a control set of genes (**Fig. Sup. 4B**). Thus preferential deregulation of broad associated gens appears to be a more general feature of leukemic cells.

**Experimental model of T-ALL**

To further explore the link between oncogenic transformation and deregulation of broad domains, we set up an experimental model of T-ALLL by overexpressing TLX3, a major transcription factor oncogene, in mouse developing thymocytes (**Fig. 6A**). Sorted CD4-CD8- primary thymocytes from wild-type mice were infected with TLX3-expressing retrovirus and co-cultured on OP9 stroma cells. A homogenous population of TLX3 expressing cells was obtained which proliferate indefinitely and resemble leukemic cells (hereafter, DN-TLX3 cells). DN-TLX3 were subjected to gene expression and ChIP-seq (TLX3 and H3K4me3) analyses and compared to our previously published data from DN thymocytes derived from Rag2-/- mice (Rag-DN). Broad H3K4me3 domains were highly dynamics (**Fig. 6B**) and gain and loss of broad domains was significantly associated with, respectively, up and down regulation of associated genes (**Fig. 6C, left panel**), including direct target genes (**Fig. 6C, right panel**). Strikingly, the expression of genes associated with broad domains in DN-TLX3 cell line was significantly more up-regulated

as compared to a set of expression-matched genes with sharp domains (**Fig. 6D, left panel**). On the contrary, the expression of genes associated with broad domains in Rag-DN cells was significantly more down-regulated in DN-TLX3 cells as compared to a set of expression-matched genes with sharp domains (**Fig. 6D, right panel**). Finally, lost broad domains in the leukemic DN-TLX3 cell line were associated with genes related to normal immunological and T cell functions, while gained broad domains were associated with genes related with cancer and leukemogenesis (**Fig. 6E**). Thus, ectopic expression of TLX3 oncogenic factor selectively impact on the expression of broad domain-associated genes, likely reproducing the situation in naturally occurring leukemia.

## Conclusions & Perspectives

In this study we identified BH4D-associated genes in human thymic T cell populations and a collection of T-ALL cell lines and primary samples. BH4D were found to distinguish between normal and neoplastic cells and to flag key leukemia associated oncogenes. Moreover, the dynamic of BH4D and of the expression of BH4D-associated genes between normal and leukemic samples appeared to be highly relevant to leukemogenesis. The fact that BH4D are also associated with lincRNAs suggests that this criterion could be used as an epigenetic prioritization of cancer-relevant LncRNAs.

## Materiel and Methods

### RNA-seq alignment and quantification

The adapters were removed from reads using CutAdapt and trimmed for low quality reads (less than 20). The quality control statistics for the reads were obtained using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The trimmed reads were aligned to the human genome version Hg19 using Tophat (Trapnell et al., 2009). Reads that were aligned to unique chromosomal positions were retained. The aligned RNA-seq reads were used to generate a separate density profiles in BigWig format for the positive and negative strand. The total numbers of reads that overlap with the coding and LncRNA transcripts or genes (exonic region only) were obtained using the cuffdiff function of cufflinks (Trapnell et al., 2010).

### Identification of differentially expressed genes

The differential expression analysis for the coding and no coding transcripts were performed using DESeq2 (Anders and Huber, 2010) using the count data. Transcripts were said to be differentially expressed (DE) if there was more than one Log2fold change in expression between normal and leukemic samples with a p-value less than or equal to 0.01.

**Clustering of differentially expressed genes**

Log2 (n + 1) transformed reads per kilo-base per million mapped reads (RPKM) for the DE transcripts were clustered using hierarchical clustering with Pearson correlation and average linkage and plotted as a heat map using Gene-E.

**Dynamic expression patterns identification**

All transcripts were clustered into 20 expression patterns using k-means clustering with Euclidean distance metric and 100 maximum iterations (stats package in R) during developmental transitions in the progression of hematopoietic commitment. The standardized expression values of the transcripts in the different expression patterns were plotted in boxplot. The Cluster contains the genes for which the expression decrease during beta selection (ISP > EC > LC) was identified.

**Functional annotation**

Over-represented Gene Ontology (GO) terms were identified using g:Profiler (Reimand et al., 2016) with biological processes ontology and the default set of all human genes as the standard background set. Over-represented KEGG pathways for gene sets were obtained using *g:Profiler* similar to GO term enrichment analysis. The top 20 over-represented GO terms with the best p-values were identified for each of the cell lines.

**CHIP-seq alignment and Peak Calling**

ChIP-seq for H4K4me3 in thymic T cell subpopulations and T-ALL primary samples were generated within the frame of the Blueprint consortium. The reads were trimmed for low quality reads with sickle (Joshi NA, Fass JN. 2011) and mapped to reference genome using Bowtie.12.7 (Langmead and Salzberg, 2012). The peaks were called using MACS2.08 (Zhang et al., 2008) with a p-value cutoffs less than 0.1 and using the "--broad" and "-bdg" options. For each cell population, the respective ChIP-seq inputs were used as control data during peak calling. In addition to the peak files, MACS2 generated Bedgraph files which are transformed to bigwig files with bedGraphToBigWig. Peaks were assigned to the transcripts with the overlapped TSS [between TSS and TSS+1.5kb]. ChIP-seq data were visualized using the Integrated Genome Browser (http://bioviz.org/igb/).

**Peak classification**

H3K4me3 broad, intermediate or sharp domains were determined by identifying high (IpH) and low (IpL) inflection points of the peak length (PL) versus gene rank. The inflection points were computed using the R package "inflection" (Demetris T. C. et al.,

2013). The peaks are classified as broad if PL > IpH; Sharp if PL < IpL and intermediate if IpL < PL < IpH.

**Control list**

Control lists were determined based on a reference gene list computes as optimally matched for signal level of expression using R script.

**Average tag density profile of histone marks**

The average tag density profiles for histone mark H3K4me3 in a given cell line, around the TSS were generated by calculating the mean tag density normalized as coverage per million (CPM), within -5kb of the TSSs of reference and control lists genes. The TSSs were grouped into different classes based on their H3K4me3 breadth at each T cell developmental stage and different cell lines. The tag density profiles were computed individually for each class and average density over 50bp windows were used to generate the density profiles. The average densities were computed using R script and the profiles were plotted using ggplot2 in R.

**Legends**

**Figure 1: Dynamic of BH4D during early T cell differentiation.** A) Overlap of BH4D associated with coding or LncRNA genes between the thymic T cell subpopulations. B) Expression dynamic of BH4D-associated genes in each T cell subpopulation. C) Examples of BH4D loci.

**Figure 2: Characterization of BH4D in T-ALL samples.** A) Number of BH4D in the thymic T cell populations, the primary T-ALL and the T-ALL cell lines. B) Heatmap showing the association of BH4D with oncogenes frequently activated in T-ALL. C) Examples of key oncogenes associated with BH4D in three T-ALL cell lines (top). Detailed view of BH4D associated with TAL1 oncogene in the Jurkat cell line.

**Figure 3: Comparison of BH4D between normal T cell populations and T-ALL.** A) Classical multidimensional scaling classification of T cell populations and T-ALL samples in function of BH4D. The dotted red line indicated separation between the normal and leukemic samples. The Loucy cell line was excluded from this analyze because of the exceptional elevated number of BH4D. B) Cluster of normal and T-ALL samples based on the GO terms enrichment of each sample.

**Figure 4: Dynamic expression of BH4D associated genes.** A-B) Two sets of equally expressed genes associated with broad (B) or sharp (S) H3K4me3 peaks in the indicated T-ALL sample (A) or the normal T cell population (B) were defined. The fold change of the expression between the T-ALL sample and the normal T cell population (either Thymus or LC) was determined.

**Figure 5: Gene expression in function of the recurrent association with BH4D.** Genes specifically associated with BH4D in T-ALL samples were separated in function of

their frequency. The fold change expression between the thymus and a series of T-ALL from public resources (Atak et al. 2011) was analyzed.

**Figure 6: Dynamics of BH4D in a mouse model of T-ALL.** A) Experimental workflow. Triple negative thymocytes were infected with a retroviral vector expressing TLX3 oncogene and the GFP. After one month of co-culture with OP9 stromal cells, the GFP positive cells were purified and analyzed by FACS (TLX3 expression was confirmed by western blot, not shown). B) Overlap between the BH4D found in DN thymocytes from Rag2-/- mice and TLX3+ leukemic cells (DN-TLX3). C) Boxplot showing the fold change of expression of BH4D associated genes only in DN-TLX3 (gain) or Rag-DN (loss) (left panel). Right panel: only the direct targets of TLX3 were analyzed. D) Two sets of equally expressed genes associated with broad (B) or sharp (S) H3K4me3 peaks in DN-TLX3 cells were defined. The fold change of the expression between DN-TLX3 and Rag-DN was determined. E) GO-terms enrichment of genes associated with BH4D only in DN-TLX3 (gain) or in Rag-DN (loss). The inverted log10 of the Benjamin-corrected *P* value is shown.

**Supplemental Figure 1: Comparison between the two strategies to identify BH4D-associated genes.** A-B) Strategy based on the identification of Broad peaks using MACS2 (A) or based on the coverage of the signal (B). C) Overlap between the two strategies. D) Correlation between two experimental replicates of LC thymocytes. E-F) Comparison of GO-term (E) and pathway (F) enrichments between the two strategies.

**Supplemental Figure 2: Identification of BH4D in the different samples.** The H3K4me3 peaks were ranked in function of the size for thymic T cell subpopulations (A), primary T-ALL samples and T-ALL cell lines (C). The inflection point allowing the selection of BH4D is indicated by a red line.

**Supplemental Figure 3: Overlap between broad associated genes in T-ALL and normal T cell samples.**

**Supplemental Figure 4: Analyses of BH4D in Chronic Lymphocytic Leukemia (CLL).** A) MDS classification of normal B cell and CLL samples. B) Dynamic expression of BH4D-associated genes. Analyses are as in Figure 4.

**Supplemental Table 1: Summary table of BH4D found in each sample and their association with coding and non-coding genes.**

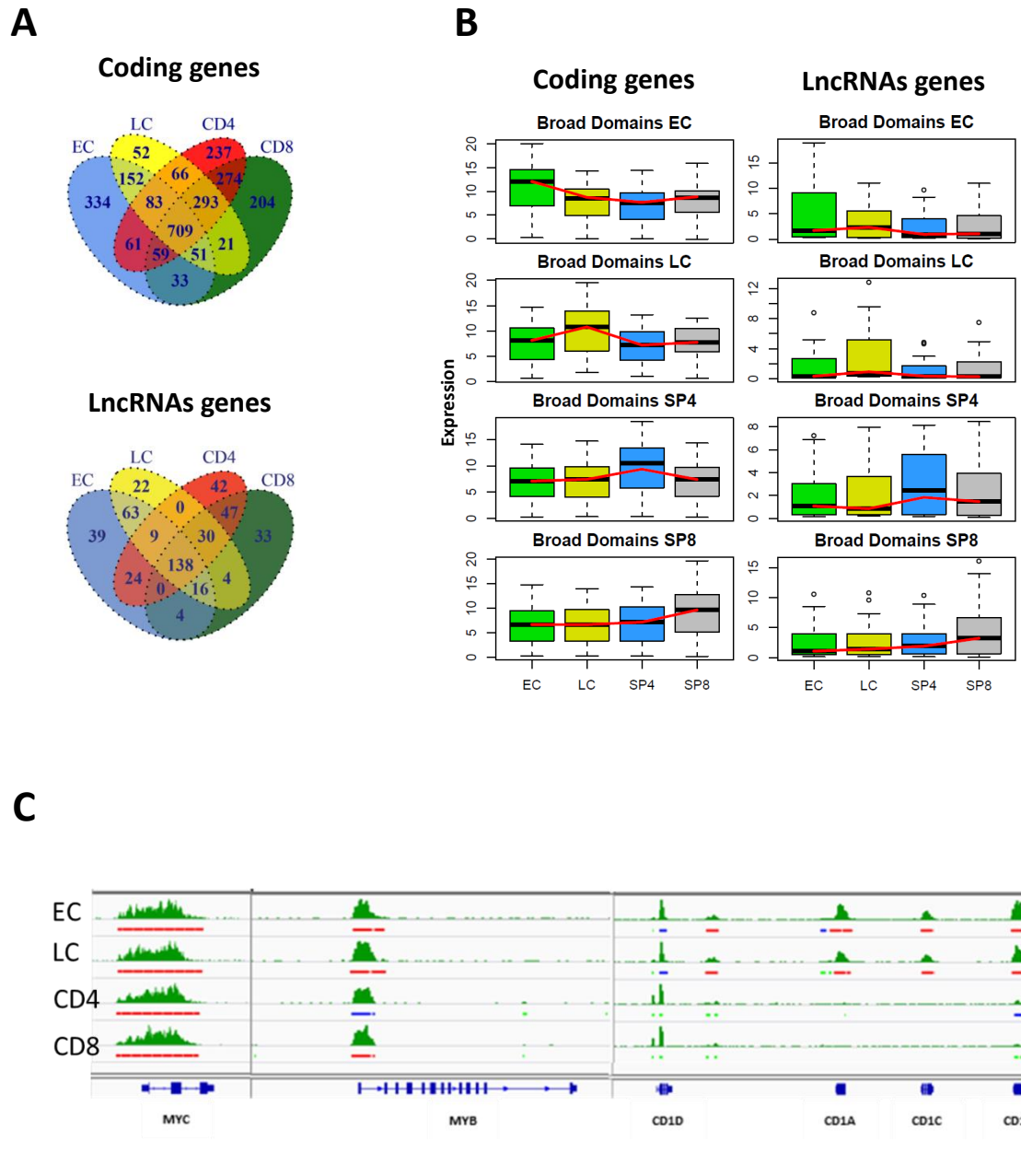

A
Coding genes
EC
LC
CD4
CD8
52
237
66
152
274
334
83
293
204
709
61
21
59
51
33
LncRNAs genes
EC
LC
CD4
CD8
22
42
0
63
47
39
9
30
33
138
24
4
0
16
4
B
Coding genes
LncRNAs genes
Broad Domains EC
Broad Domains LC
Broad Domains SP4
Broad Domains SP8
Expression
EC
LC
SP4
SP8
C
EC
LC
CD4
CD8
MYC
MYB
CD1D
CD1A
CD1C
CD1B
CD1E

Figure1

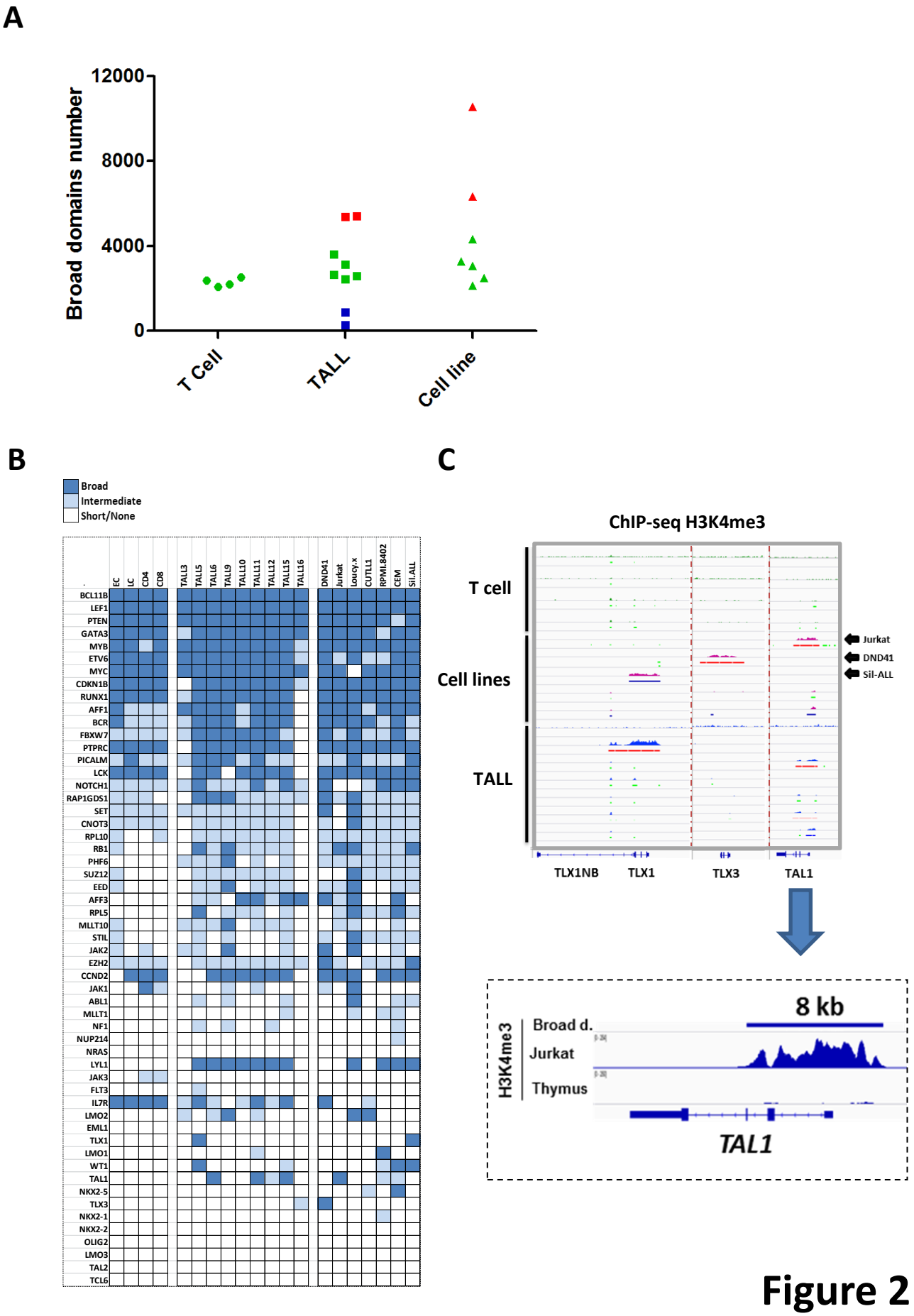
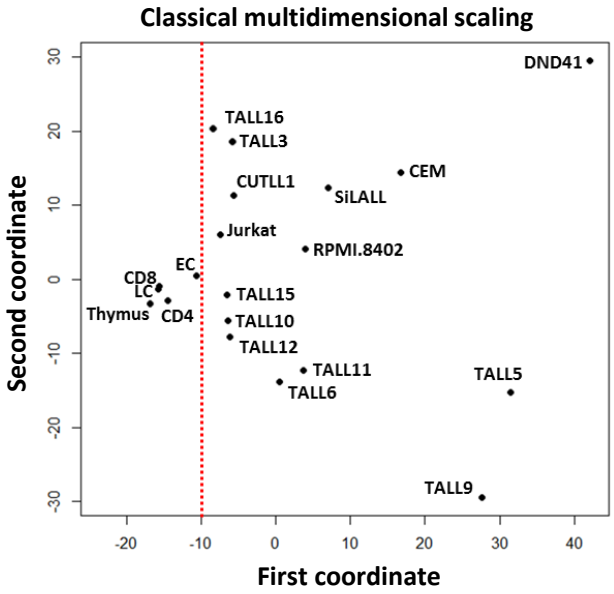

A
Broad domains number
12000
8000
4000
0
T Cell
TALL
Cell line
B
Broad
Intermediate
Short/None
EC
LC
CD4
CD8
TALL3
TALL5
TALL6
TALL9
TALL10
TALL11
TALL12
TALL15
TALL16
DND41
Jurkat
Loucy.x
CUTLL1
RPMI.8402
CEM
Sil.ALL
BCL11B
LEF1
PTEN
GATA3
MYB
ETV6
MYC
CDKN1B
RUNX1
AFF1
BCR
FBXW7
PTPRC
PICALM
LCK
NOTCH1
RAP1GDS1
SET
CNOT3
RPL10
RB1
PHF6
SUZ12
EED
AFF3
RPL5
MLLT10
STIL
JAK2
EZH2
CCND2
JAK1
ABL1
MLLT1
NF1
NUP214
NRAS
LYL1
JAK3
FLT3
IL7R
LMO2
EML1
TLX1
LMO1
WT1
TAL1
NKX2-5
TLX3
NKX2-1
NKX2-2
OLIG2
LMO3
TAL2
TCL6
C
ChIP-seq H3K4me3
T cell
Cell lines
TALL
Jurkat
DND41
Sil-ALL
TLX1NB
TLX1
TLX3
TAL1
8 kb
H3K4me3
Broad d.
Jurkat
Thymus
TAL1

Figure 2

**A**


Classical multidimensional scaling
Second coordinate
First coordinate
DND41
TALL16
TALL3
CEM
CUTLL1
SiLALL
Jurkat
RPMI.8402
EC
CD8
LC
TALL15
Thymus
CD4
TALL10
TALL12
TALL11
TALL5
TALL6
TALL9
30
20
10
0
-10
-20
-30
-20
-10
0
10
20
30
40

**B**

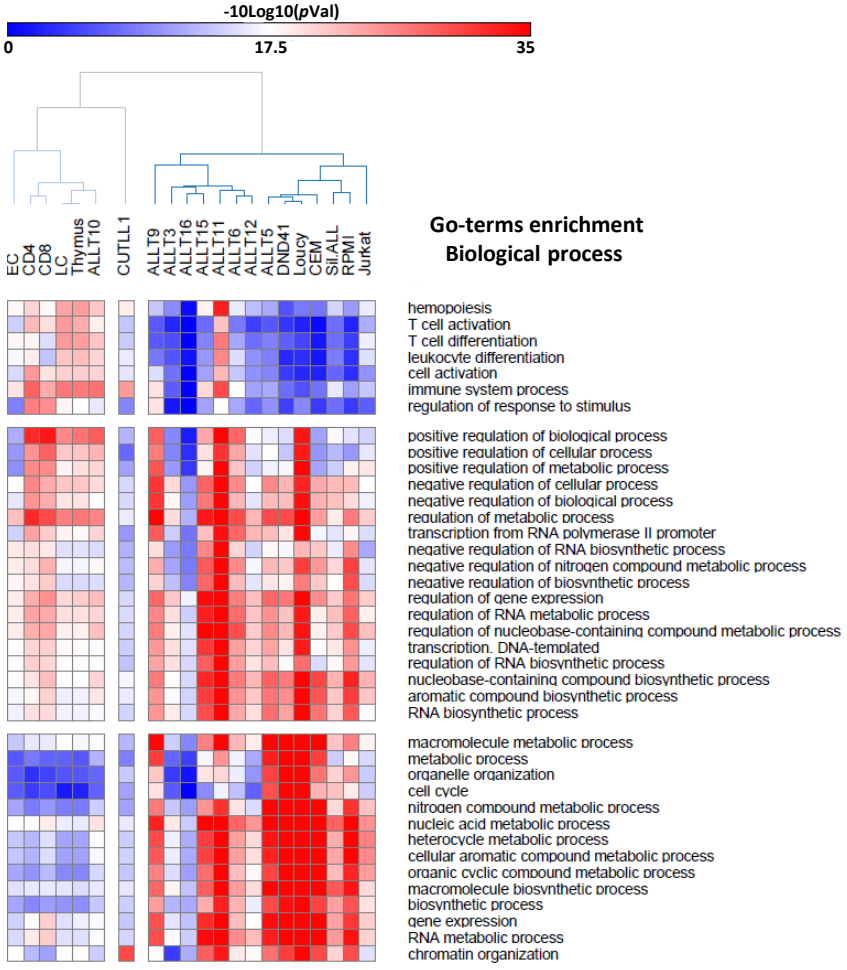
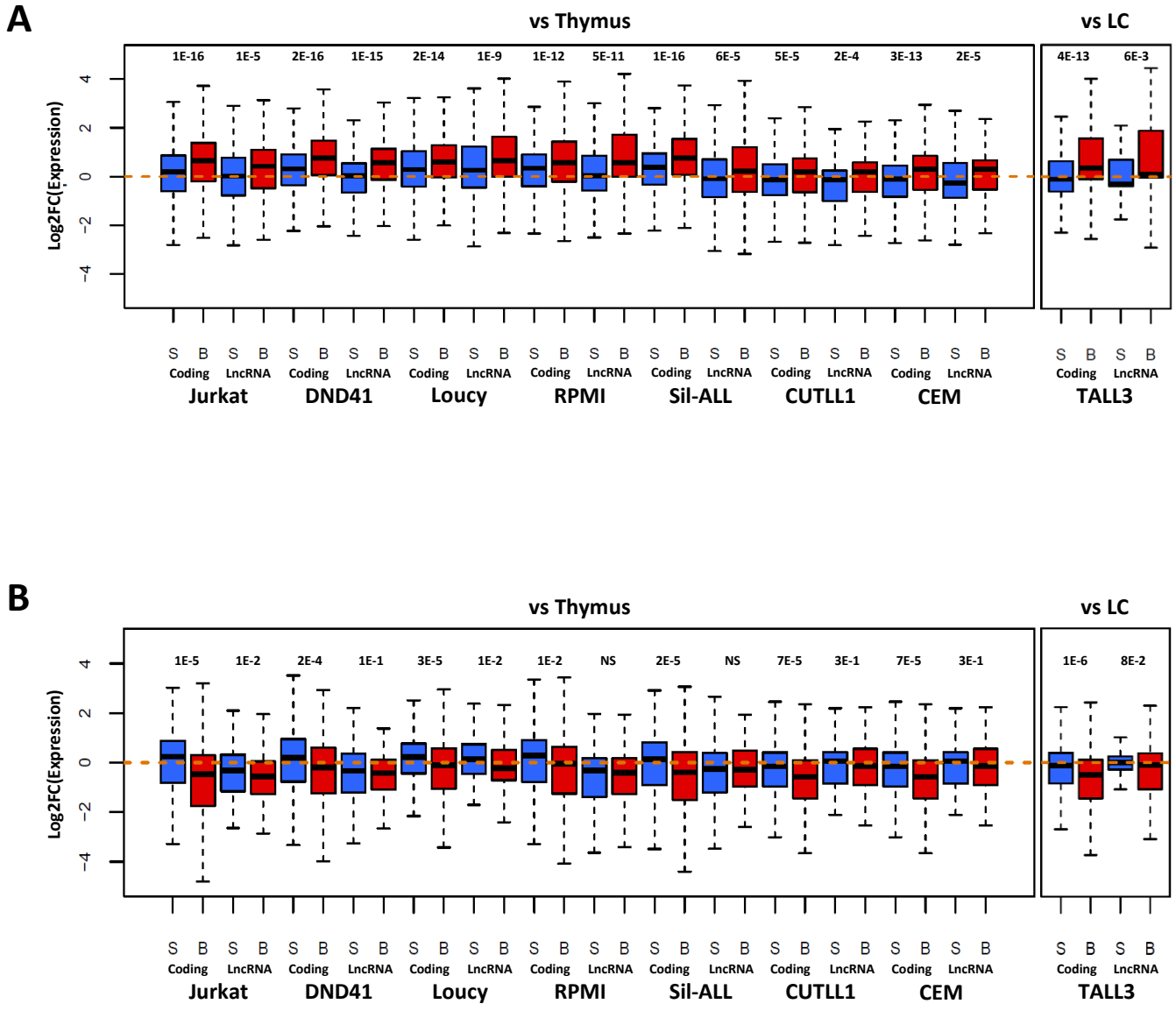

-10Log10(pVal)
0
17.5
35
EC
CD4
CD8
LC
Thymus
ALLT10
CUTLL1
ALLT9
ALLT3
ALLT16
ALLT15
ALLT11
ALLT6
ALLT12
ALLT5
DND41
Loucy
CEM
SiLALL
RPMI
Jurkat
Go-terms enrichment
Biological process
hemopoiesis
T cell activation
T cell differentiation
leukocyte differentiation
cell activation
immune system process
regulation of response to stimulus
positive regulation of biological process
positive regulation of cellular process
positive regulation of metabolic process
negative regulation of cellular process
negative regulation of biological process
regulation of metabolic process
transcription from RNA polymerase II promoter
negative regulation of RNA biosynthetic process
negative regulation of nitrogen compound metabolic process
negative regulation of biosynthetic process
regulation of gene expression
regulation of RNA metabolic process
regulation of nucleobase-containing compound metabolic process
transcription, DNA-templated
regulation of RNA biosynthetic process
nucleobase-containing compound biosynthetic process
aromatic compound biosynthetic process
RNA biosynthetic process
macromolecule metabolic process
metabolic process
organelle organization
cell cycle
nitrogen compound metabolic process
nucleic acid metabolic process
heterocycle metabolic process
cellular aromatic compound metabolic process
organic cyclic compound metabolic process
macromolecule biosynthetic process
biosynthetic process
gene expression
RNA metabolic process
chromatin organization

**Figure 3**


A
vs Thymus
vs LC
Log2FC(Expression)
1E-16 1E-5 2E-16 1E-15 2E-14 1E-9 1E-12 5E-11 1E-16 6E-5 5E-5 2E-4 3E-13 2E-5
4E-13 6E-3
S B S B S B S B S B S B S B S B S B S B S B S B S B S B S B
Coding LncRNA Coding LncRNA Coding LncRNA Coding LncRNA Coding LncRNA Coding LncRNA Coding LncRNA Coding LncRNA Coding LncRNA
Jurkat DND41 Loucy RPMI Sil-ALL CUTLL1 CEM TALL3


B
vs Thymus
vs LC
Log2FC(Expression)
1E-5 1E-2 2E-4 1E-1 3E-5 1E-2 1E-2 NS 2E-5 NS 7E-5 3E-1 7E-5 3E-1
1E-6 8E-2
S B S B S B S B S B S B S B S B S B S B S B S B S B S B S B
Coding LncRNA Coding LncRNA Coding LncRNA Coding LncRNA Coding LncRNA Coding LncRNA Coding LncRNA Coding LncRNA Coding LncRNA
Jurkat DND41 Loucy RPMI Sil-ALL CUTLL1 CEM TALL3

**Figure 4**

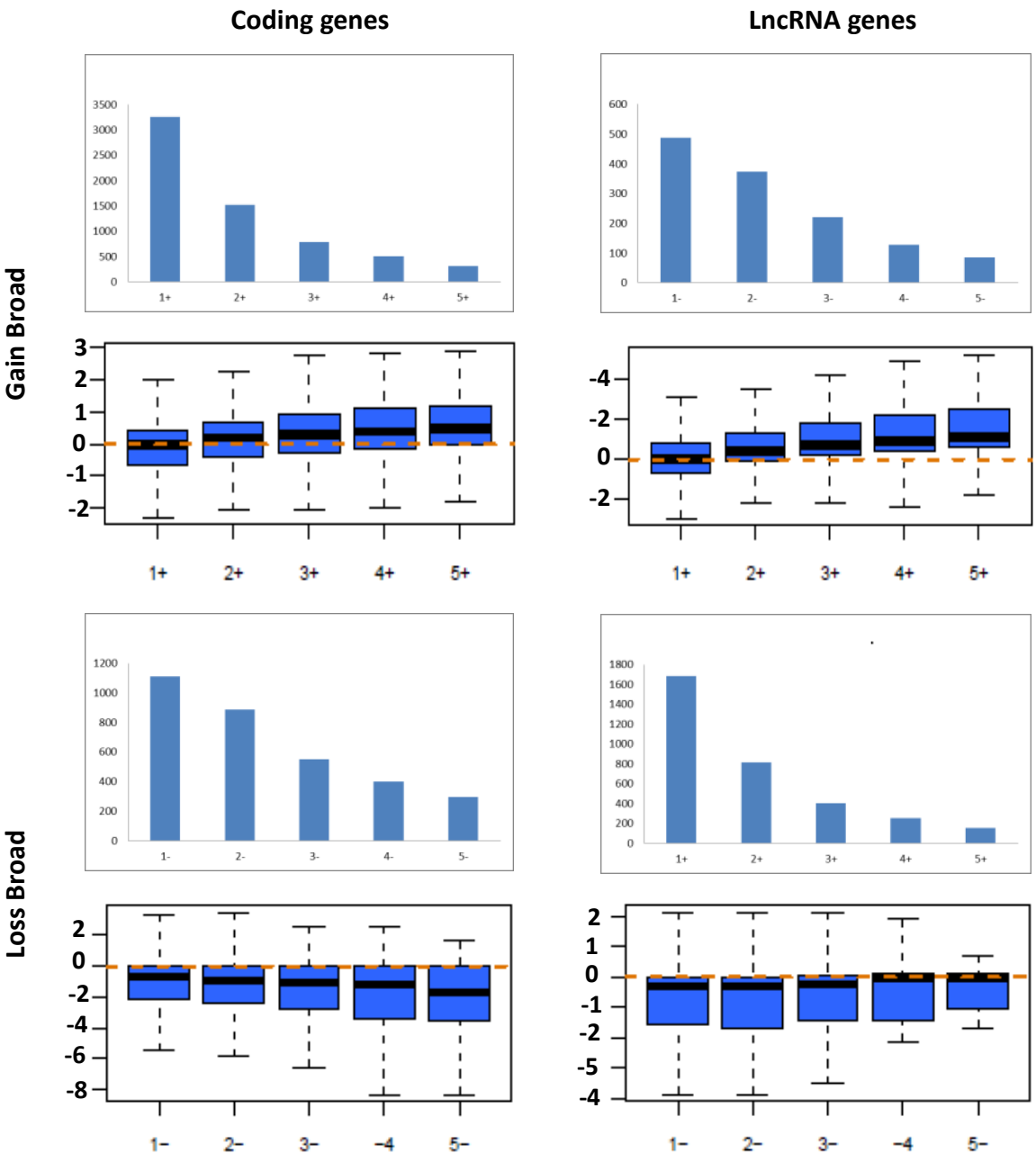

Coding genes
LncRNA genes
Gain Broad
Loss Broad
1+
2+
3+
4+
5+
1-
2-
3-
4-
5-

Figure 5

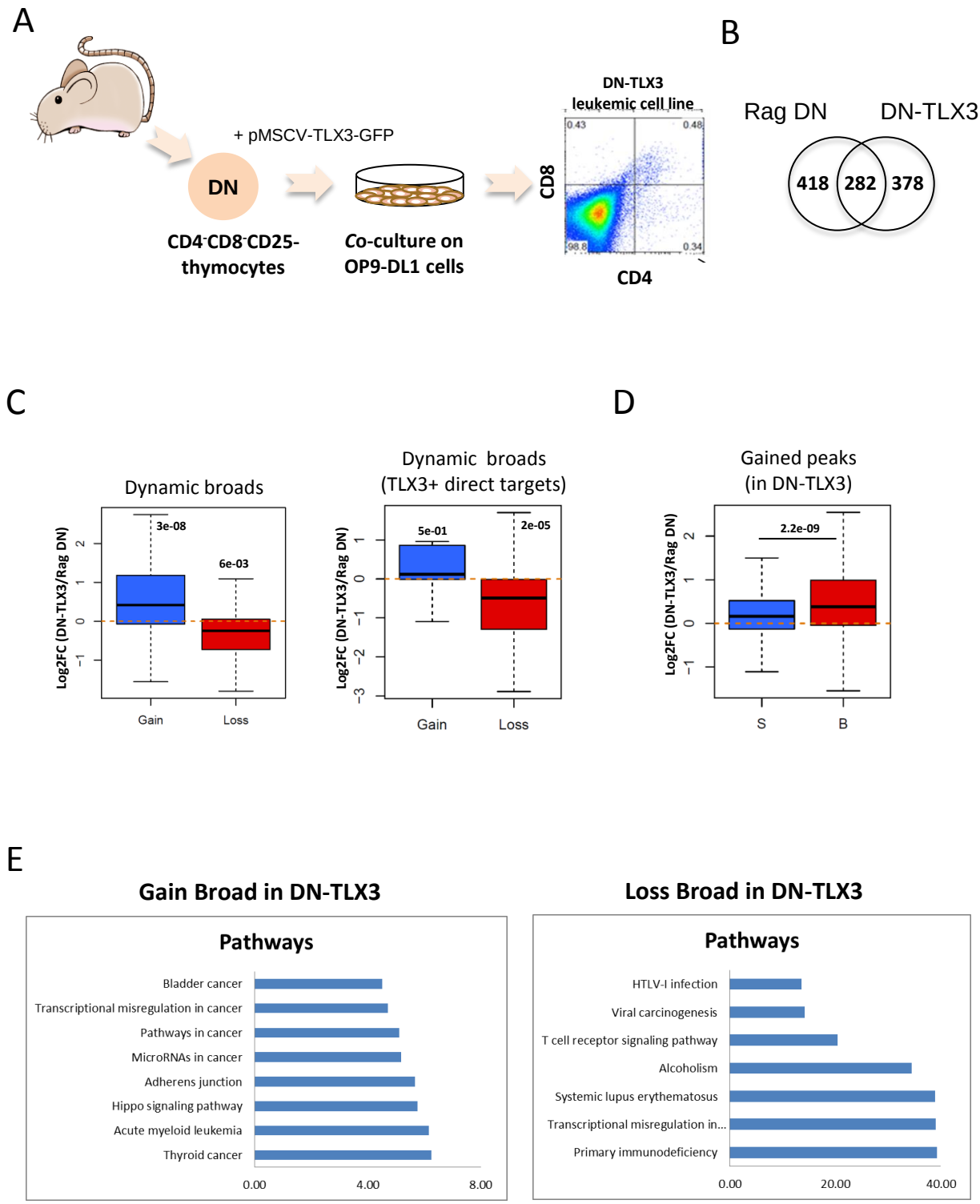
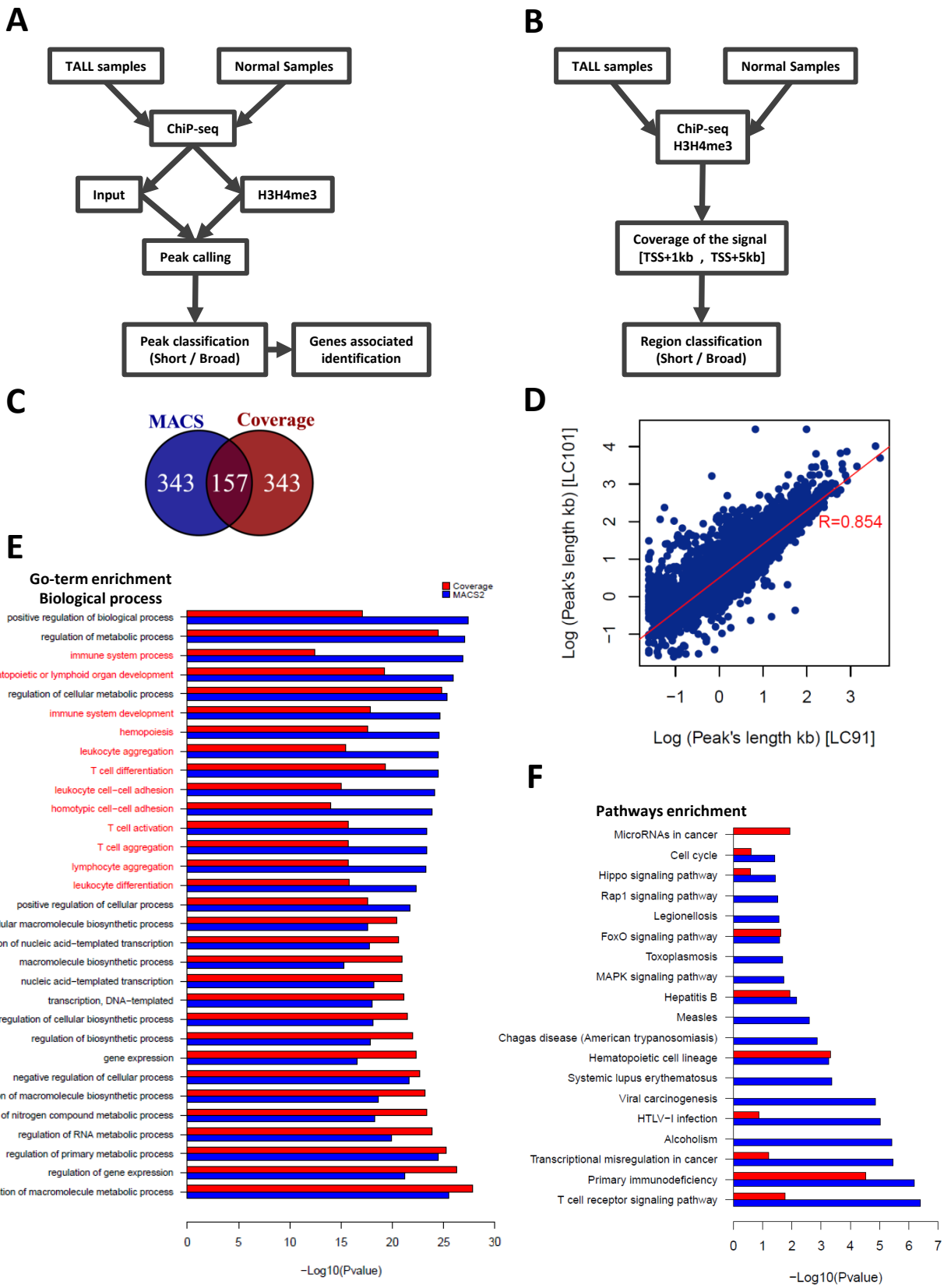
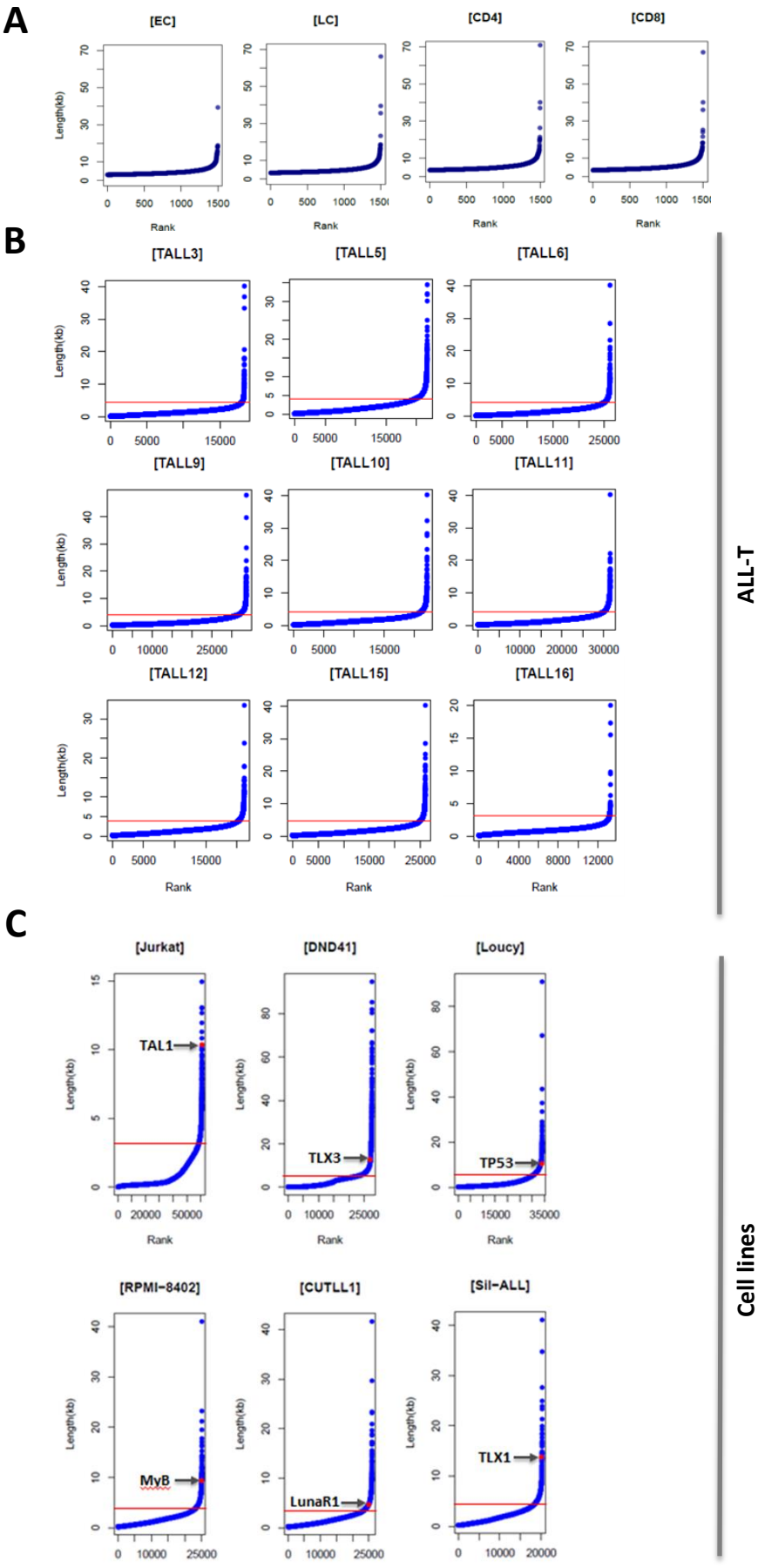

A
+ pMSCV-TLX3-GFP
DN
CD4-CD8-CD25-
thymocytes
Co-culture on
OP9-DL1 cells
DN-TLX3
leukemic cell line
CD8
CD4
0.43
0.48
98.8
0.34
B
Rag DN
DN-TLX3
418
282
378
C
Dynamic broads
Log2FC (DN-TLX3/Rag DN)
3e-08
6e-03
Gain
Loss
Dynamic broads
(TLX3+ direct targets)
5e-01
2e-05
D
Gained peaks
(in DN-TLX3)
2.2e-09
S
B
E
Gain Broad in DN-TLX3
Pathways
Bladder cancer
Transcriptional misregulation in cancer
Pathways in cancer
MicroRNAs in cancer
Adherens junction
Hippo signaling pathway
Acute myeloid leukemia
Thyroid cancer
0.00
4.00
8.00
Loss Broad in DN-TLX3
Pathways
HTLV-I infection
Viral carcinogenesis
T cell receptor signaling pathway
Alcoholism
Systemic lupus erythematosus
Transcriptional misregulation in...
Primary immunodeficiency
0.00
20.00
40.00

**Figure 6**



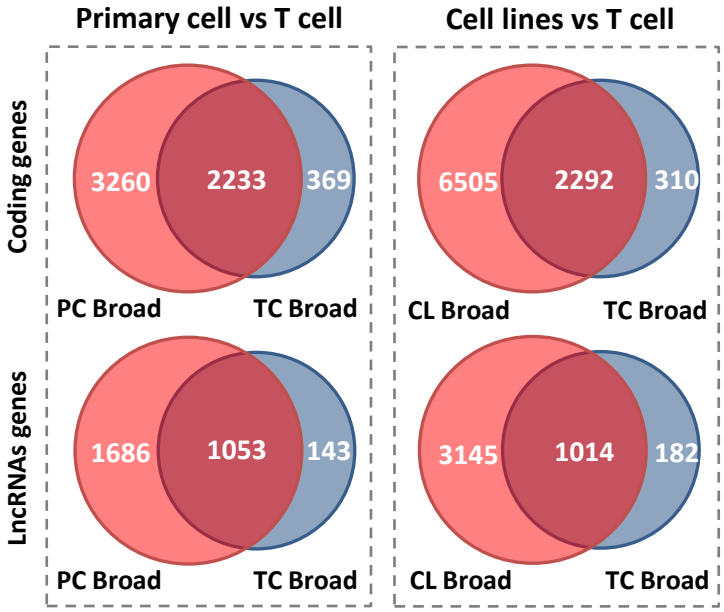

**Supplemental Figure 1**




**Supplemental Figure 2**

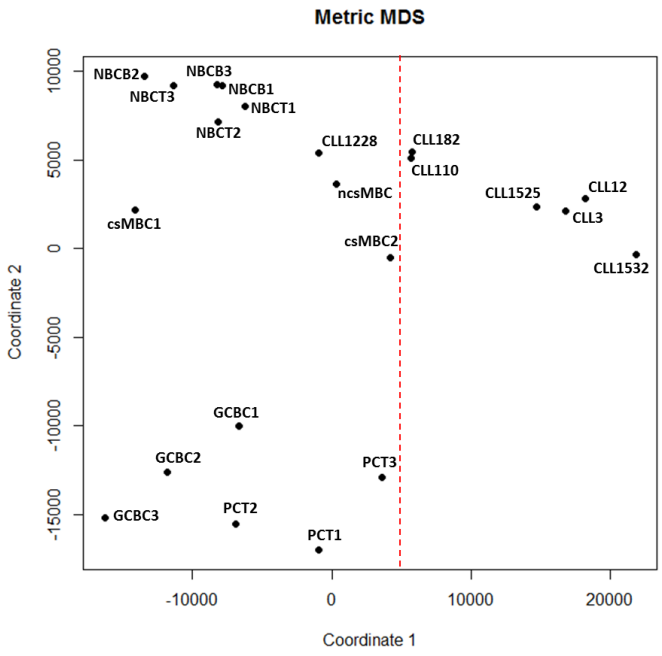
# Supplemental Table 1

| Type | Sample | NB Peaks | Threshold1 | Threshold2 | Coding None | Coding short | Coding intermediate | Coding broad | LncRNA None | LncRNA short | LncRNA intermediate | LncRNA broad | LincRNA None | LincRNA short | LincRNA intermediate | LincRNA broad |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SubPop | EC | 17684 | 2192 | 3692 | 12283 | 13936 | 4916 | 1460 | 22902 | 5364 | 1959 | 729 | 14743 | 1852 | 448 | 286 |
| SubPop | LC | 19900 | 2540 | 4038 | 11264 | 16302 | 3624 | 1405 | 22381 | 6451 | 1450 | 672 | 14524 | 2208 | 321 | 276 |
| SubPop | CD4 | 23371 | 2570 | 4071 | 10226 | 16302 | 4305 | 1762 | 21864 | 6553 | 1787 | 750 | 14262 | 2369 | 415 | 283 |
| SubPop | CD8 | 23067 | 2683 | 4183 | 10206 | 17041 | 3724 | 1624 | 21870 | 6806 | 1531 | 747 | 14243 | 2483 | 336 | 267 |
| TALL | TALL3 | 18352 | 3001 | 4500 | 13615 | 16362 | 2111 | 507 | 23672 | 5932 | 991 | 359 | 15047 | 1855 | 268 | 159 |
| TALL | TALL5 | 21766 | 2600 | 4099 | 13200 | 10134 | 5689 | 3572 | 22844 | 4087 | 2183 | 1840 | 14567 | 1578 | 546 | 638 |
| TALL | TALL6 | 26105 | 2643 | 4144 | 9880 | 15550 | 5032 | 2133 | 21737 | 6294 | 1921 | 1002 | 14190 | 2370 | 430 | 339 |
| TALL | TALL9 | 33645 | 2570 | 4068 | 11700 | 11153 | 6119 | 3623 | 22111 | 4923 | 2160 | 1760 | 14269 | 2029 | 541 | 490 |
| TALL | TALL10 | 22084 | 2670 | 4149 | 10406 | 16131 | 4312 | 1746 | 22004 | 6350 | 1704 | 896 | 14284 | 2317 | 401 | 327 |
| TALL | TALL11 | 31535 | 2703 | 4204 | 9006 | 16302 | 4749 | 2538 | 20947 | 7409 | 1533 | 1065 | 13676 | 2917 | 395 | 341 |
| TALL | TALL12 | 21198 | 2485 | 3983 | 11086 | 15590 | 4280 | 1639 | 2248 | 6046 | 1624 | 801 | 14548 | 2139 | 361 | 281 |
| TALL | TALL15 | 25961 | 2720 | 4785 | 10350 | 14294 | 6234 | 1717 | 21818 | 5832 | 2444 | 860 | 14174 | 2223 | 605 | 327 |
| TALL | TALL16 | 13282 | 2003 | 3010 | 14320 | 16914 | 1174 | 187 | 24817 | 5534 | 505 | 98 | 15618 | 1510 | 143 | 58 |
| CellLine | Jurkat | 60859 | 2630 | 4133 | 12063 | 12805 | 6300 | 1427 | 21857 | 5986 | 2405 | 706 | 14057 | 2632 | 387 | 253 |
| CellLine | DND41 | 27566 | 4810 | 6313 | 14047 | 9332 | 5074 | 4142 | 22510 | 4241 | 2003 | 2200 | 14294 | 1783 | 454 | 798 |
| CellLine | Loucy | 34010 | 3650 | 5148 | 12991 | 8092 | 4022 | 7490 | 22417 | 4042 | 1410 | 3085 | 14471 | 1817 | 346 | 695 |
| CellLine | CUTLL1 | 26018 | 3201 | 4700 | 12936 | 14370 | 3684 | 1605 | 22420 | 6061 | 1590 | 883 | 14262 | 2290 | 408 | 369 |
| CellLine | Sil-ALL | 20218 | 3167 | 5170 | 14534 | 10455 | 5470 | 2136 | 23253 | 4424 | 2144 | 1133 | 14750 | 1641 | 535 | 403 |
| CellLine | CCRF-CEM | 33481 | 2499 | 4800 | 11529 | 8309 | 9876 | 2881 | 21541 | 4245 | 3723 | 1445 | 13846 | 2032 | 983 | 468 |
| CellLine | RPMI-8402 | 25095 | 3110 | 4619 | 11780 | 14585 | 4216 | 2014 | 22249 | 5963 | 1684 | 1058 | 14351 | 2181 | 453 | 344 |

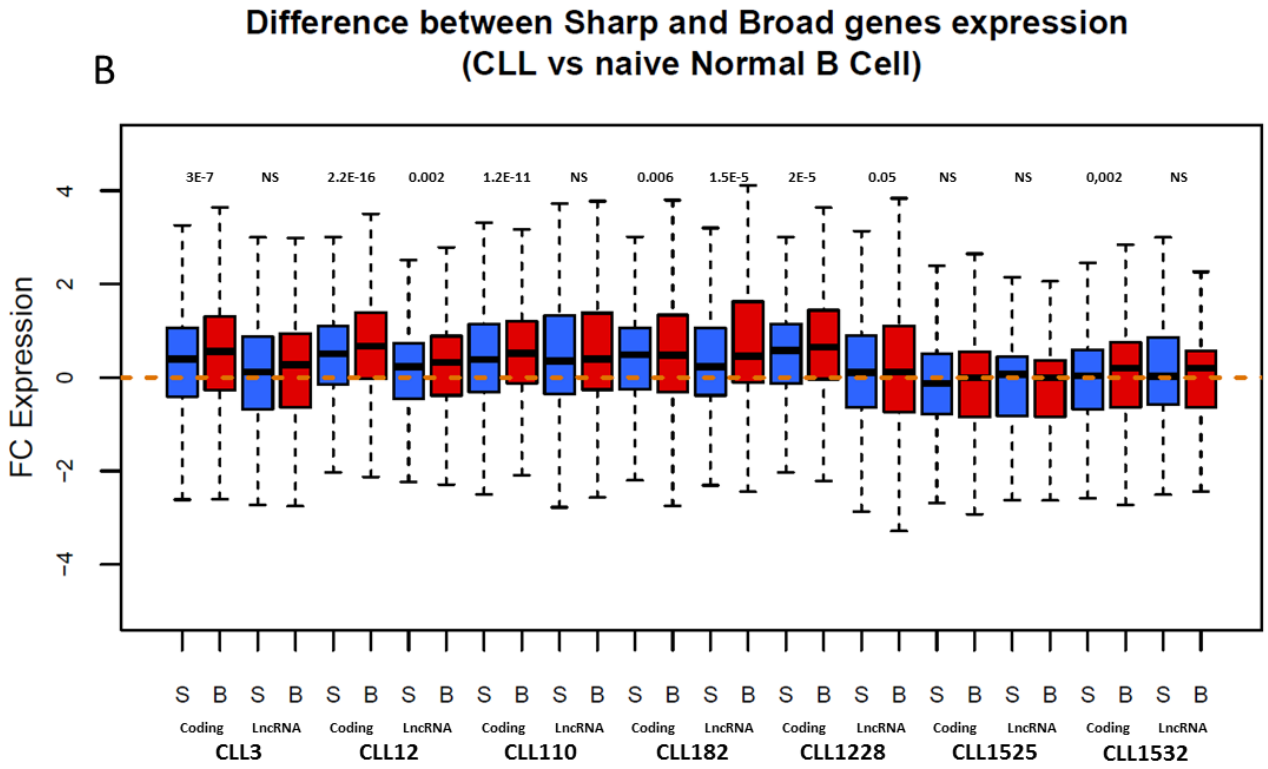



**Supplemental Figure 3**

A




B




**Supplemental Figure 4**

# CHAPITRE III :

## Etude des ARNs longs non codants au cours de la differentiation des lymphocytes T et dans les léucemies

## 1. Introduction:

Seulement quelques études à l´échelle du génome se sont focalisées sur l'identification des LncRNA et elles se sont concentrées sur un nombre limité de tissus ou de lignées cellulaires. Par conséquent, il est important d'identifier systématiquement les espèces de transcrits qui sont impliquées dans les processus développementaux complexes et d'explorer leurs fonctions et mécanismes d'action.

Donc, pour comprendre ces LncRNAs, deux défis importants doivent être surmontés : (1) l'identification des LncRNAs potentiellement fonctionnels et (2) leur attribution de fonctions vérifiables expérimentalement. Nous assumons que l'étude des LncRNAs qui sont exprimés différemment durant les stages de développement cellulaire pourra permettre l'identification des LncRNAs dont les implications dans la différenciation cellulaire des lymphocytes T et dans les leucémies sont les plus probables.

Jusqu'à présent très peu d'études se sont intéressés au rôle des LncRNAs dans un contexte hématopoïétique normal (Alvarez-Dominguez et al., 2014), ou leucémique(Garzon et al., 2014, Guo et al., 2014). L'importance des LncRNA dans les LAL-T a été récemment mise en évidence par une étude qui s'est intéressée aux LncRNAs régulés par la voie de signalisation *Notch* ce qui les a conduit à l'identification d'un nouveau LncRNA appelé *LUNAR1* et qui est impliqué dans la croissance leucémique (Trimarchi et al., 2014). Ces données appuient l'hypothèse d'un rôle important des LncRNAs comme régulateurs épigénétiques clés dans la leucémogenèse.

Ainsi, durant ma thèse je me suis intéressé à deux parties : (i) l'identification systématique des transcrits non codants divergents dans les tissus normaux et l'analyse de leurs caractéristiques épigénétiques ; (ii) L'identification et caractérisation des LncRNAs impliqués dans le développement des lymphocytes T et dans la leucémie lymphoblastique aiguë des cellules T.

# Article 3 :

**La transcription divergente est associée aux promoteurs des facteurs de transcription**

La transcription divergente est un phénomène répandu chez les mammifères (Fig. 36). En particulier, les transcrits courts sens/antisens sont une caractéristique des promoteurs actifs (Core et al., 2008) ; de plus, les longs transcrits instables en amont de la plupart des gènes exprimés qu'on peut détecter dans des conditions où la machinerie de dégradation de l'ARN est inhibée (Preker et al., 2008). Cependant, de nombreux LncRNAs décrits sont des transcrits antisens à partir des promoteurs des gènes codants (Guttman et al., 2010, Cabili et al., 2011, Derrien et al., 2012, Sigova et al., 2013). Par contre, l'importance de ces paires LncRNA/ARNm est encore mal comprise.

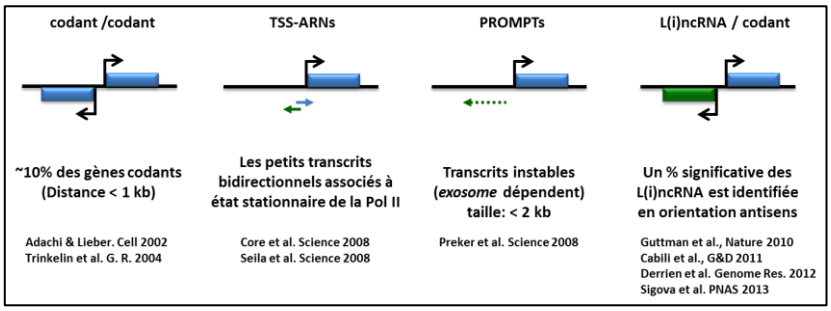



**Fig. 36: Les différents types de la transcription bidirectionnelle.**

En conséquence, nous avons utilisé des données RNA-Seq sens-spécifique afin d'identifier les transcrits antisens dans les tissus primaires de souris. Nous avons utilisé également des données ChIP-Seq afin d'analyser leurs caractéristiques épigénétiques. Nous avons trouvé que la transcription divergente est associée de manière significative à des gènes liés à la régulation de la transcription et le développement. Ces promoteurs présentent des caractéristiques génétiques et épigénétiques uniques. Ces résultats suggèrent fortement un lien fonctionnel entre la transcription divergente et la régulation des gènes codant pour des facteurs de transcription et de développement et ainsi pourraient illustrer un mécanisme supplémentaire pour affiner l'expression des gènes.

**Contributions dans l'article :**

- Réalisation des analyses bioinformatiques pour la caractérisation épigénétique des promoteurs associés à la transcription divergente ;
- Participation à l'écriture de l'article ;
- Réalisation des figures de l'article :
  - Fig. 1C ;
  - Fig. 5E et I ;
  - Fig. 6A, B, C et D ;
  - Fig. 7A et C.

BMC Genomics

RESEARCH ARTICLE Open Access

# Divergent transcription is associated with promoters of transcriptional regulators

Cyrille Lepoivre[1,2,3,10†], Mohamed Belhocine[1,2,3†], Aurélie Bergon[1,2,3,4], Aurélien Griffon[1,2,3], Miriam Yammine[1,2,3,4], Laurent Vanhille[1,2,3,5,6,7], Joaquin Zacarias-Cabeza[5,6,7], Marc-Antoine Garibal[1,2,3], Frederic Koch[5,6,7], Muhammad Ahmad Maqbool[5,6,7], Romain Fenouil[5,6,7], Beatrice Loriod[1,2,3,4], Hélène Holota[1,2,3,4], Marta Gut[8], Ivo Gut[8], Jean Imbert[1,2,3,4], Jean-Christophe Andrau[5,6,7*], Denis Puthier[1,2,3,9*] and Salvatore Spicuglia[1,2,3*]

## Abstract

**Background:** Divergent transcription is a wide-spread phenomenon in mammals. For instance, short bidirectional transcripts are a hallmark of active promoters, while longer transcripts can be detected antisense from active genes in conditions where the RNA degradation machinery is inhibited. Moreover, many described long non-coding RNAs (lncRNAs) are transcribed antisense from coding gene promoters. However, the general significance of divergent lncRNA/mRNA gene pair transcription is still poorly understood. Here, we used strand-specific RNA-seq with high sequencing depth to thoroughly identify antisense transcripts from coding gene promoters in primary mouse tissues.

**Results:** We found that a substantial fraction of coding-gene promoters sustain divergent transcription of long non-coding RNA (lncRNA)/mRNA gene pairs. Strikingly, upstream antisense transcription is significantly associated with genes related to transcriptional regulation and development. Their promoters share several characteristics with those of transcriptional developmental genes, including very large CpG islands, high degree of conservation and epigenetic regulation in ES cells. In-depth analysis revealed a unique GC skew profile at these promoter regions, while the associated coding genes were found to have large first exons, two genomic features that might enforce bidirectional transcription. Finally, genes associated with antisense transcription harbor specific H3K79me2 epigenetic marking and RNA polymerase II enrichment profiles linked to an intensified rate of early transcriptional elongation.

**Conclusions:** We concluded that promoters of a class of transcription regulators are characterized by a specialized transcriptional control mechanism, which is directly coupled to relaxed bidirectional transcription.

**Keywords:** Divergent transcription, lncRNA, Bidirectional promoter, GC skew, Developmental transcription factor

## Background

Transcription of eukaryotic genomes generates a myriad of non-coding RNAs that show complex overlapping patterns of expression and regulation [1]. The complexity of the eukaryotic transcriptome, transcribed by RNA Polymerase (Pol) II, goes far beyond the coding genome and expands to many short RNA populations (such as miRNAs, siRNAs, piwiRNAs, eRNAs, TSS-RNAs) as well as long non-coding RNAs (lncRNAs) [2]. LncRNAs form a heterogeneous group of RNAs transcribed from intergenic or intragenic regions, which vary in length from 200 nucleotides to over 100 kb [3]. Intragenic non-coding transcripts might be further subdivided depending on the way they overlap protein-coding genes and/or the orientation with respect to protein-coding genes (sense or antisense) [4]. Although the biological relevance of many non-coding transcripts has been unambiguously established, this unanticipated level of complexity has led to the notion of pervasive transcription, which refers to the fact that transcription is not restricted to well-defined functional features, such as genes [5-7].

A large proportion of lncRNAs are transcribed in antisense orientation of protein-coding genes, with which they

* Correspondence: andrau@ciml.univ-mrs.fr; puthier@tagc.univ-mrs.fr; salvatore.spicuglia@inserm.fr
†Equal contributors
[5]Centre d'Immunologie de Marseille-Luminy, Campus de Luminy, Case 906, 13288 Marseille cedex 9, France
[1]Technological Advances for Genomics and Clinics (TAGC), Case 928, 163 Avenue de Luminy, 13288, Marseille cedex 09, France
[2]Aix-Marseille Université, UMR1090 TAGC, Marseille F-13288 France
Full list of author information is available at the end of the article

often share sequence complementarities [8,9]. Antisense RNAs could potentially exert a regulatory function on their corresponding sense mRNA at different levels. Recent findings have shown that some antisense transcripts act as epigenetic regulators of gene expression and chromatin remodeling [8], while others play a role at the level of translation efficiency [10]. Besides these transcripts, the existence of non-coding antisense transcripts emanating from the promoters of protein-coding genes (i.e. head-to-head conformation) has also emerged as a widespread phenomenon from yeast to mammals [11]. On the one hand, the presence of short bidirectional transcripts appears to be a hallmark of active promoters in mammals [12-14]. On the other hand, relatively longer non coding antisense transcripts can be detected upstream of most expressed genes in conditions where the RNA degradation machinery is inhibited [15-18]. Moreover, lncRNAs (including long intergenic non-coding RNAs or lincRNA) are preferentially localized at the vicinity of gene promoters in antisense orientation [4,19-21]. For instance, ~60% of lncRNAs expressed in ES cells were found to originate close to the TSS of protein-coding genes [21]. Whether long antisense transcripts emanating from bidirectional promoters have general functional implications in gene regulation is currently unknown [11].

In order to systematically identify and characterize bidirectional promoters associated with long non-coding antisense transcription, we took advantage of strand-specific RNA-seq experiments, which provide an unprecedented opportunity to analyze and categorize transcripts [22]. Thorough analyses of RNA-seq data from early developing thymocytes and other mouse tissues indicated that long-range bidirectional transcription is an intrinsic property of a class of promoters whose associated genes mainly encode for transcriptional regulators involved in development and cell differentiation. Accordingly, these promoters are characterized by large CpG islands, high degree of conservation and are generally repressed by Polycomb complexes in ES cells. Moreover, they display a unique GC skew profile, while the associated coding genes have large first exons, both properties likely reminiscent of their bidirectional activity. Surprisingly, coding genes associated with upstream antisense lncRNAs display an increased rate of immature transcription, highlighting an additional level of transcriptional control. Thus, expression of long non-coding antisense transcripts appears as a common feature of a subset of mammalian protein-coding gene promoters with functional implications for gene regulation.

## Results

### Systematic identification of genes associated with long upstream antisense transcripts

We sought to assess whether production of long antisense transcripts is a general feature of mammalian gene promoters. To this goal, we initially performed strand-specific paired-end Total RNA-seq with high sequencing depth from ΔRag thymocytes (Additional file 1: Table S1), corresponding mainly to $CD4^{-}CD8^{-}$ T-cell precursors (hereafter, double negative or DN thymocytes). We selected the set of protein-coding RefSeq transcripts whose promoter regions (from −5 kb to the transcription start site, TSS) do not overlap with transcripts of any other coding gene (a total of 17,186 transcripts; see Methods). We then calculated the total RNA-seq signal in the sense and antisense orientation for the region −5 kb to +5 kb with respect to each TSS and ordered the selection in function of the level of upstream antisense (AS) transcription (from −5 kb to the TSS; Figure 1A). Using a stringent threshold ($p < 0.005$; see Methods) we found 6.8% (1,177) of coding RefSeq transcripts to be associated with long upstream antisense transcripts (hereafter, LUATs), of which 236 overlap with previously annotated non-coding transcripts. Several examples are shown in Figure 1B. Interestingly, these antisense transcripts are generally polyadenylated as shown by the average profile of strand-specific and PolyA-enriched RNA-seq signal generated from ΔRag thymocytes (Figure 1B and C; Additional file 1: Table S1). To confirm our observation in a different tissue, we analyzed strand-specific Total RNA-seq data from mouse embryonic kidney [23] and obtained consistent results (Additional file 1: Table S1 and Additional file 2: Figure S1).

In order to infer the structure of LUATs, we used the transcript assembly tool Cufflinks [24,25]. We selected antisense Cufflinks transcripts starting within the region +/− 1.5 kb around the TSS and longer than 200 nt, and inferred antisense transcript models for 992 out of 1,177 RefSeq genes associated with divergent transcription in ΔRag thymocytes, as defined above (Additional file 3: Table S2). As expected, LUATs have very low or no coding potential as determined by PhyloCSF [26] analyses (Additional file 4: Figure S2). Assessment of subcellular localization of LUATs using recently published RNA-seq data obtained from fractionated chromatin-associated, nucleoplasmic and cytoplasmic transcripts (Bhatt et al. [27]), showed that they remains mainly associated with the chromatin fraction (Additional file 5: Figure S3), consistent with their lack of coding potential. Interestingly, 58% of antisense transcripts start within the region 500 bp upstream the TSS of the associated-coding genes (Figure 1D), suggesting that both sense and antisense transcripts originate from the same promoter elements.

LUATs were found to be expressed at relatively low level with a median expression value of 0.8 fragments per kilobase per million fragments mapped or FPKM (Figure 1E; see also Methods for details on quantification of LUAT expression). Strikingly however, the LUAT-associated coding genes are expressed at high levels (median expression value 3.9 FPKM; Figure 1E). Indeed, expression levels of

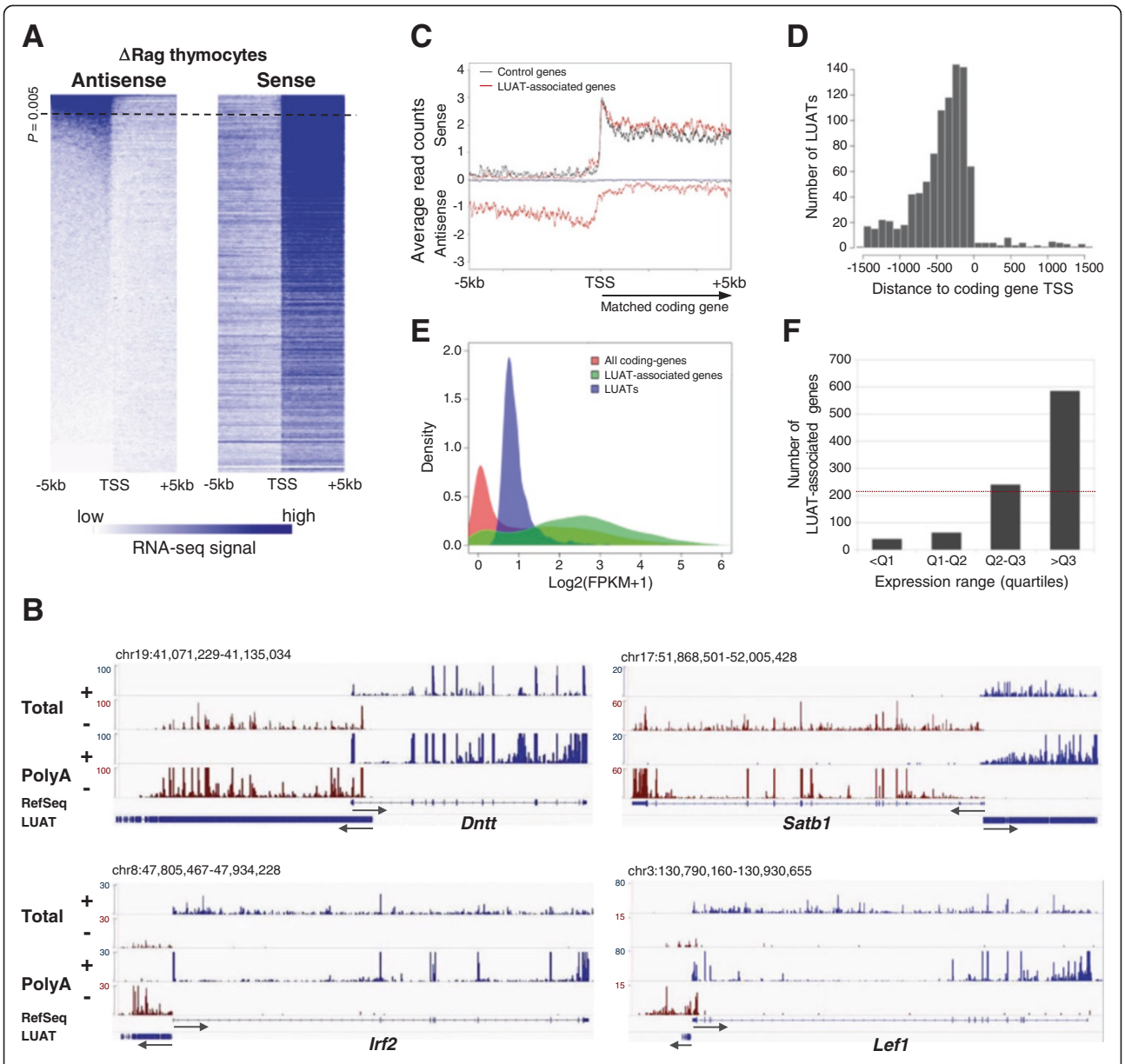



**Figure 1 Identification of genes associated with long upstream antisense transcripts in DN thymocytes. A)** Heatmap showing the Total RNA-seq signal from ΔRag (DN) thymocytes (SoliD platform) found in the 5 kb region surrounding the TSS of all non-overlapping Refseq genes. Signal was computed based on the number of reads per 100 bp binned regions originated from either the antisense or sense strand with respect to gene orientation (left and right panels, respectively). The heatmap is ordered according to the antisense signal for the [−5 kb; 0] region. The threshold for significantly expressed antisense transcripts is shown by a dotted line (see Methods section). **B)** Examples of genes associated with LUAT in ΔRag thymocytes. The Total and PolyA RNA-seq signals for the plus and minus strands are shown. Arrows indicate transcript orientation. The scales and genomic coordinates are shown on the left and top of each panel, respectively. Note that the scales were independently fixed for the plus and minus strands in order to properly visualize sense and antisense transcripts. **C)** Average profiles of PolyA- RNA-seq signal in ΔRag thymocytes for LUAT-associated genes (red line) and a control set of similarly expressed genes (black line). Signals corresponding to the orientation of the coding genes are represented as positive values while antisense signals as negative values. **D)** Histogram of the positions of 5′ end of LUATs relative to the TSS of their associated coding-genes. **E)** Distribution of expression of all coding genes (red), LUAT-associated genes (green) and LUATs (blue) in ΔRag thymocytes. **F)** Number of LUAT-associated genes in each expression quartile of all coding genes (Q1 = 3.05e-6 FPKM; Q2 = 0.013 FPKM; Q3 = 1.99 FPKM). The red line indicates the expected (random) distribution.

the majority of genes displaying divergent promoters were found to be above the 3rd quartile of expression value distribution in ΔRag thymocytes (Figure 1F). Overall, these results suggest that antisense transcription is initiated from active coding-gene promoters, leading to concomitant expression of the two divergent transcripts.

### LUAT-associated genes are related to transcription and developmental functions

The above results indicated that a substantial fraction of mammalian promoters sustain divergent transcription of lncRNA/mRNA gene pairs. We next assessed whether LUAT-associated genes were enriched for specific categories of genes. We found that this set of genes is highly enriched with transcription- and chromatin regulation-related GO terms (Figure 2A). This observation was specific to LUAT-associated genes as compared to a control set composed of genes with similar expression level distribution, but without antisense transcription, which was not significantly enriched for any GO terms (considering Benjamini-corrected $p < 0.001$ as a threshold). Strikingly, the list of LUAT-associated genes includes most transcription regulators known to be important for early T-cell differentiation [28], including *Tcf7* (TCF1), *Lef1*, *Tcf12* (E47), *Satb1, Dntt, Gfi1*, *Myb*, *Tox*, *Notch1, Bcl11a*, *Rorc* (Rorγt) and *Ikzf1*. Consistent with a tissue-specific function, LUATs are significantly associated with a higher proportion of thymocyte-specific genes (Figure 3A) and with genes involved in T cell differentiation (ToppGene analysis for "Mouse Phenotypes" [29]; Bonferroni-corrected *p*-value: 0.004), as compared to the control set. Enrichment for transcription- and development-related functions was also found with LUAT-associated genes isolated from kidney RNA-seq data (Figure 2B; Additional file 2: Figure S1), comprising important regulators of kidney development, such as *Irx2*, *Irx3*, *Hnf1b*, *Lhx1* and *Smad4*.

To obtain a more general view of the functional relevance of LUAT-associated genes, we analyzed recently published directional and PolyA-enriched RNA-seq data from 17 mouse tissues [30]. Although the sequencing depth was generally low, we were able to isolate LUATs for all analyzed tissues and to annotate a non-redundant set of 1,118 LUAT-associated genes (Additional file 6: Table S3). Consistent with the initial studies from thymus and kidney (Figure 2A and B), this set of genes was highly enriched for genes involved in transcription regulation and developmental functions (Figure 2C). In contrast, head-to-head coding genes (coding-coding) or randomly selected genes with unidirectional promoters are not enriched for transcription-related functions (Figure 2C; Note that control sets were chosen by selecting genes for which the breadth of expression matched those of the LUAT-associated gene set; Additional file 7: Table S4). Next, we addressed the question of whether LUAT-associated genes were specifically enriched for developmental genes involved in transcription regulation. We found that LUAT-associated genes, but not control genes, were enriched in the set of genes annotated for both "Developmental process" (GO:0032502, biological process) and "Transcription factor activity" (GO:0003700, molecular function) GO terms ($p < 1 \times 10^{-8}$, Fisher's exact test; see Material and Methods). In conclusion, LUAT-associated genes are commonly involved in transcription regulation related to developmental functions.

### LUAT and their associated coding-genes are co-regulated throughout cell differentiation and development

The co-expression of the two divergent transcripts suggests that the expression of LUATs and their neighboring genes might be co-regulated throughout development and cell differentiation. To address this hypothesis, we first compared two subsequent stages of early T-cell development. During normal T-cell differentiation, preTCR-induced signaling leads DN thymocytes to cross the so-called β-selection checkpoint, which results in massive cell proliferation and the induction of a developmental process marked by the expression of both CD4 and CD8 co-receptors, thus generating DP thymocytes [31]. To determine whether LUATs and associated genes were co-regulated during the β-selection process, we used previously published Total RNA-seq from DP thymocytes [32] and produced a new set of RNA-seq data from ΔRag (DN) thymocytes, using the same RNA-seq procedure and sequencing platform (Additional file 1: Table S1). We then selected a non-redundant list of 758 LUAT expressed in either DN or DP cell stages and compared their differential expression ratio along with the expression ratio of the associated coding-genes (Figure 3B; Additional file 8: Table S5). Interestingly, we observed a significant association between developmental regulation of LUAT and their associated genes when considering transcripts with an expression ratio of at least twofold ($p < 0.0001$; Chi-squared test). Examples of co-regulated LUAT-gene pairs are shown in Figure 3C.

To have a more thorough dynamical view of the regulated expression of LUATs and their associated genes, we analyzed recently published RNA-seq data from several stages of early T-cell differentiation [33]. Although the absence of strand-specific information did not allow genome-wide isolation of LUAT in these data sets, visual inspection of the RNA-seq revealed clear examples where the LUAT and the associated gene followed the same kinetics throughout T-cell differentiation (Additional file 9: Figure S4). In the same line, we also observed a tight co-regulation of LUAT and associated gene pairs between thymocytes and embryonic kidney (Additional file 2: Figure S1B).

One expectation from this observation is that the expression of LUATs and their associated genes would be correlated across different tissues. To address this possibility, we analyzed the expression patterns of the 1,118 LUATs and their associated genes found in the multi-tissue analysis. The vast majority of LUATs exhibit tissue-specific expression patterns as underlined by unsupervised clustering of expression profiles (k-means algorithm; Figure 4A) and the restricted number of tissues where

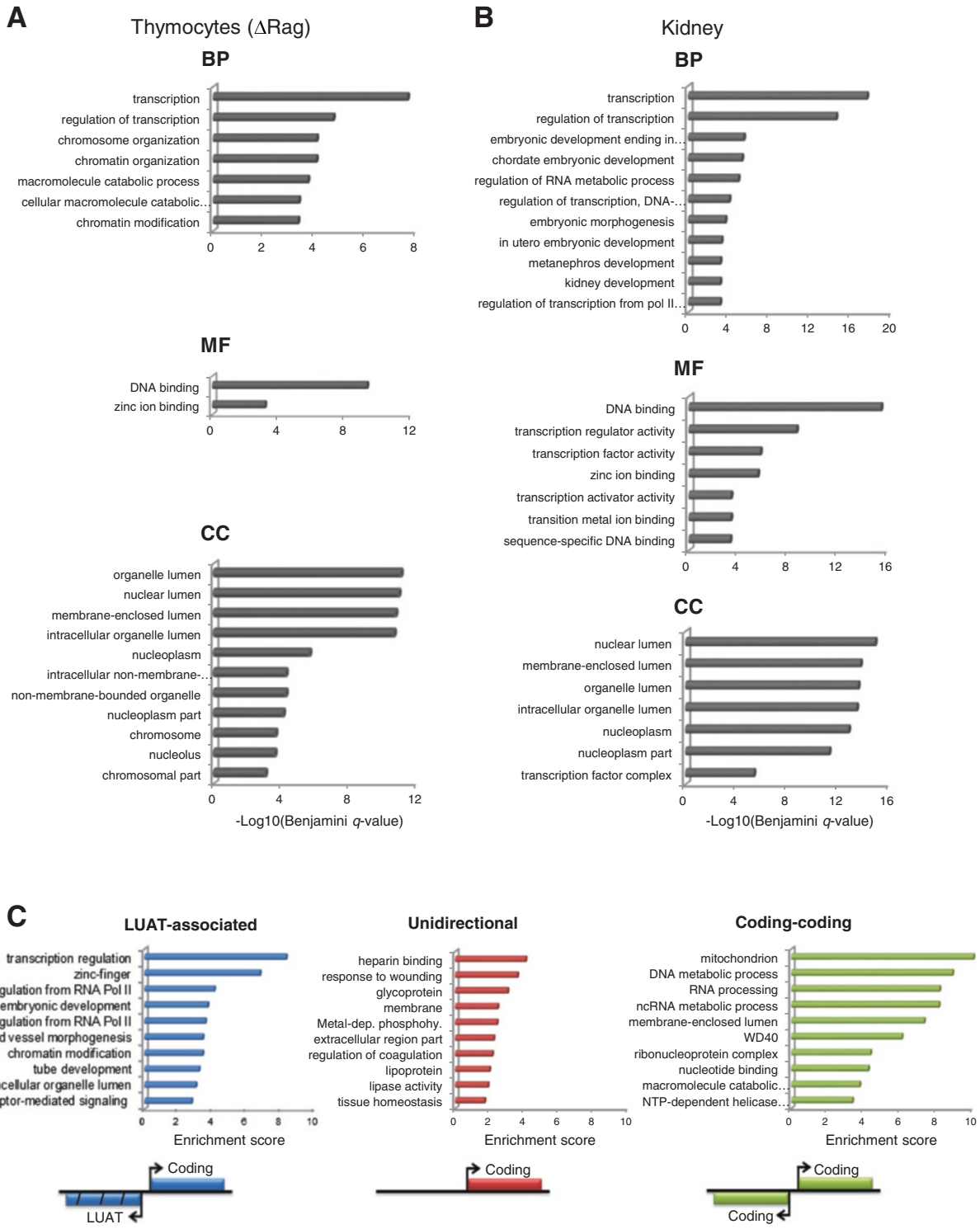



**Figure 2 Functional analysis of LUAT-associated genes. A-B)** Functional enrichment analyses for LUAT-associated genes found in ΔRag thymocytes **(A)** and embryonic kidney **(B)**. Significant GO terms for Molecular Function (MF), Biological Process (BP) and Cellular Component (CC) with a Benjamini-corrected $p < 10^{-3}$ are shown. Note that using this threshold, a set of similarly expressed control genes retrieved no significant enrichment for GO terms. **C)** Enrichment scores of functional groups found using the Functional Classification Tool from DAVID [77]. Results are shown for LUAT-associated genes found in the multi-tissue analysis, bidirectional protein-coding gene pairs (coding-coding) and genes with unidirectional promoters. The top ten groups are shown for each set of genes. The functional groups are named based on the term with the lowest *p* value found in each group.

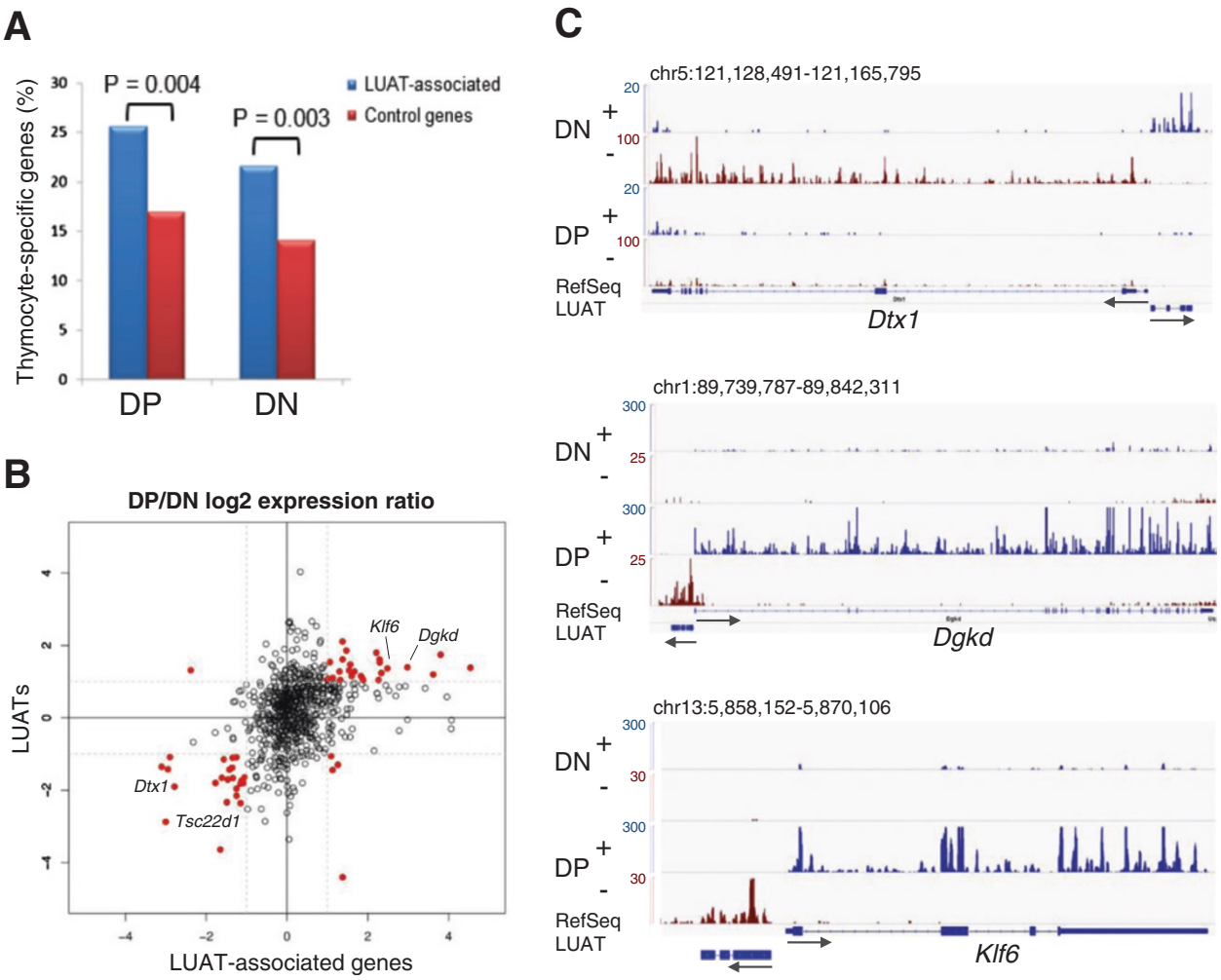



**Figure 3 Regulation of LUAT and associated genes during early T-cell differentiation. A)** Percentage of thymocyte-specific genes in the LUAT-associated gene sets from either DP or DN (ΔRag) thymocytes or in control sets of similarly expressed genes (see Methods). The *p* values results from a Chi-2 test are shown. **B)** Scatterplot showing the log2 ratio of Total RNA-seq signals (FPKM) between DN (ΔRag) and DP thymocytes (both from Illumina platform) for LUAT and associated genes. Shown in red are LUATs and associated gene pairs which are both considered as differentially expressed (2-fold change). **C)** Examples of co-regulated LUATs and associated genes between DN (ΔRag) and DP thymocytes. The Total RNA-seq signal for the plus and minus strands is shown. Legends are as in Figure 1B.

each of them was found (Additional file 10: Figure S5). Moreover, these LUATs and their coding neighbors are more correlated to each other than random gene pairs, and even slightly more than head-to-head protein-coding gene pairs (Figure 4B). In agreement, we found many examples where LUATs expression is strictly associated with the expression of their neighboring genes (Figure 4C). Taken together, these results suggest that LUAT expression likely reflects the activity of associated coding-gene promoters throughout cell differentiation and development.

### LUAT-associated promoters share characteristics with those of developmentally regulated genes

The close proximity between LUAT and TSS of associated coding genes, along with the tight correlation of their expression, strongly suggest sharing of common regulatory elements. Hence, we asked whether the bidirectional property of LUAT-associated promoters might be linked to intrinsic sequence specificities. We compared several sequence characteristics between the promoters of the three gene sets described above. We found that GC content differs between these sets. In the region upstream of the TSS, both LUAT-associated and coding-coding gene promoters have a significantly higher GC content than unidirectional promoters (*t*-test in region [–500 bp; TSS]; $p < 10^{-50}$ and $p < 10^{-81}$, respectively), whereas in the downstream region, LUAT-associated promoters have a higher GC content than the unidirectional and coding-coding gene sets (*t*-test in region [TSS; +500 bp], $p < 10^{-38}$ and $p < 10^{-45}$, respectively) (Figure 5A, left panel). Moreover, 80%

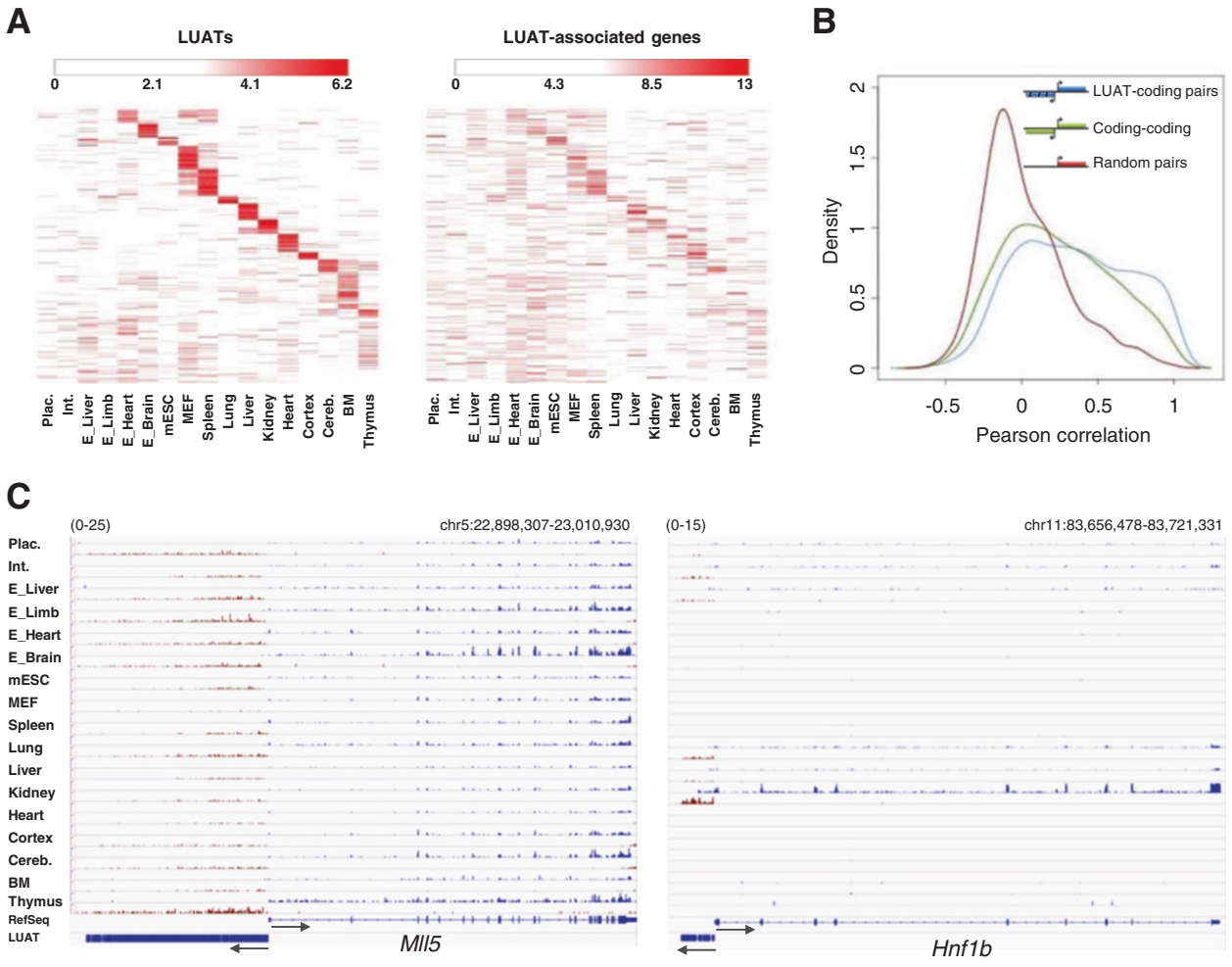



**Figure 4 Co-expression of LUAT and associated genes. A)** Left panel: heatmap of expression profiles of LUATs in the 17 indicated tissues. Expression profiles were partitioned using K-means algorithm (k = 14). Right panel: heatmap of expression profiles of LUAT-associated genes in the same tissues. Lines are ordered according to the corresponding LUATs. **B)** Distribution of Pearson correlation coefficient between expression values of indicated gene pairs across the 17 tissues. Red: randomly paired genes (with unidirectional promoters); green: head-to-head protein-coding genes; blue: LUAT and their associated genes. **C)** Examples of LUAT and associated genes across the 17 tissues. The PolyA RNA-seq signals for the plus and minus strands are shown in blue and red, respectively. The scale is indicated in the top-left of each panel.

of LUAT-associated and 89% of coding-coding gene promoters are covered by CpG islands within the region [−500 bp; +500 bp], as compared to only 56% of unidirectional gene promoters (Figure 5A, right panel). Strikingly however, analysis of CpG island size demonstrated that LUAT-associated gene promoters contain particularly longer CpG islands (Figure 5B; 46.2% of LUAT-associated promoters have a CpG island size greater than 1 kb, as compared to 23.6% and 26.5% of bidirectional coding-gene and unidirectional promoters, respectively). Sequence motif analyses revealed that both LUAT-associated and coding-coding gene promoters are depleted of TATA boxes, as compared to unidirectional promoters (Figure 5C). Finally, we found that LUAT-associated promoters contain more conserved elements than the other two sets in regions close to and downstream of the TSS (*t*-test in region [TSS; +500 bp]; $p < 10^{-166}$ and $p < 10^{-224}$, respectively; Figure 5D). Interestingly, it has been described that developmentally regulated genes are associated with Genomic Regulatory Blocks (GRB) which are highly conserved genomic regions characterized by a number of unique features, including very large CpG islands and TATA-box depletion [34,35]. Therefore, the bidirectional property of LUAT-associated promoters might be linked to intrinsic regulatory properties related to genes encoding for transcriptional and developmental regulators.

Developmental genes have also been shown to be actively repressed by Polycomb complexes in Embryonic Stem (ES) cells, and therefore are often found associated with trimethylation of H3 lysine 27 (H3K27me3) [36]. By analyzing ChIP-seq data from ES cells [37], we found that LUAT-associated promoters were specifically enriched for

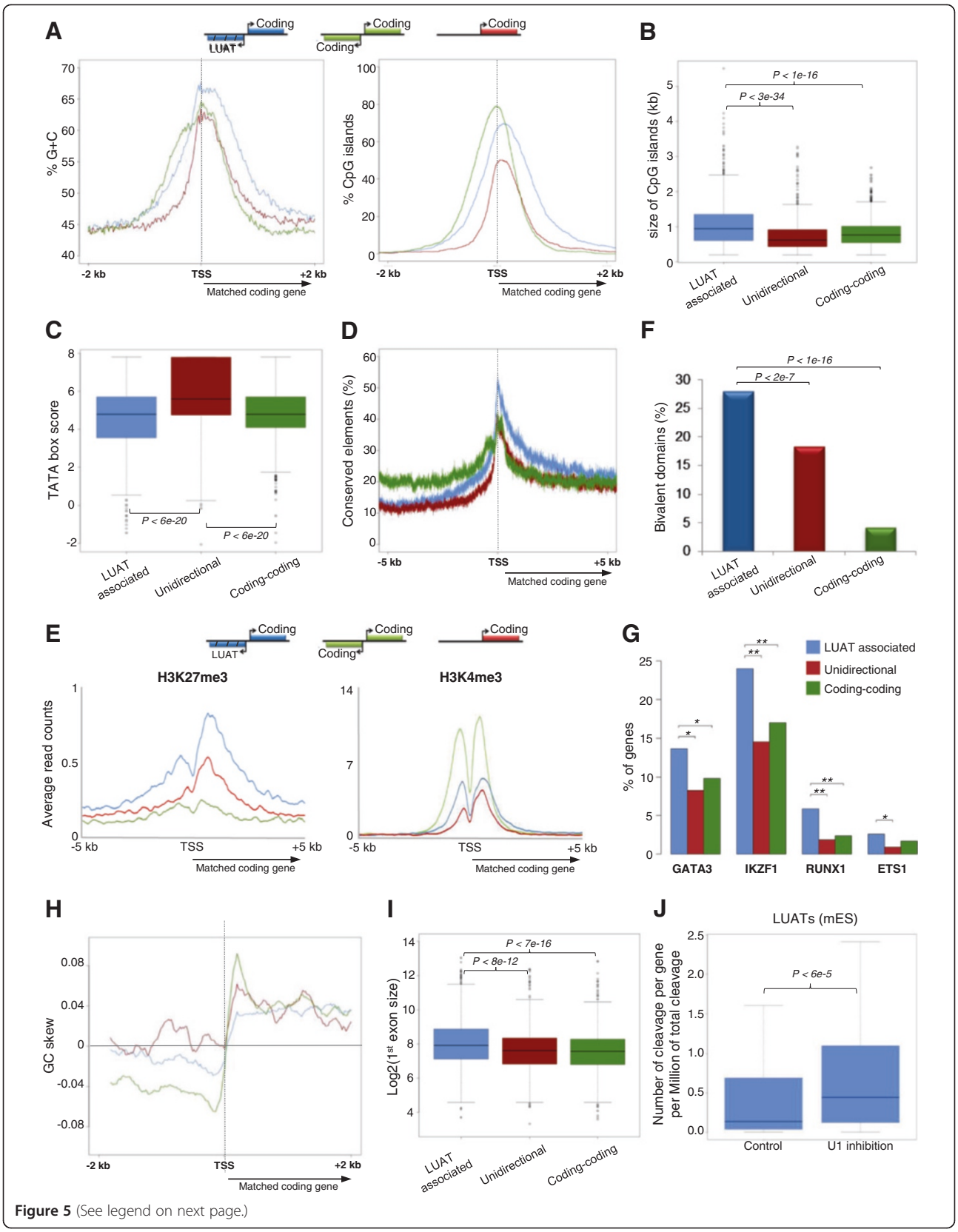



**Figure 5** (See legend on next page.)

(See figure on previous page.)
**Figure 5 Characterization of sequence content and regulatory features of LUAT-associated promoters.** Results in A-F and H-I are shown for the three set of genes described in Figure 2C. **A)** Average GC content (left panel) and percentage of CpG islands (right panel) around the TSS (bidirectional promoters are centered on the TSS from the genes that has been used to match the expression with the LUAT associated genes). **B)** Boxplot showing the distribution of sizes of the CpG islands overlapping the 2 kb region around the TSS (when several CpG islands were found, the sum was calculated). **C)** Boxplot showing the distribution of TATA box motif scores found in a 500 bp region around the TSS. **D)** Percentage of sequences with a conserved element at each position around the TSS. **E)** Average profiles of indicated ChIP-seq data in ES cells around the TSS. **F)** Percentage of genes having a bivalent domain in their promoter, as defined in [37]. Statistical significances were computed using the hypergeometric test. **G)** Percentage of genes associated with lymphoid-specific transcription factors. The histogram shows the overlap between indicated transcription factor peaks and regions around TSS (+/−5 kb) for the genes selected in DP thymocytes. Statistical significances were computed using the hypergeometric test (**$p$ value < 0.01; *$p$ value < 0.05). **H)** Average GC skew profiles, computed as (#G-#C)/(#G+#C). I) Boxplot showing the distribution of first exon length. **J)** The normalized number of cleavage sites in antisense orientation identified in two control and two U1 inhibition experiments in ES cells [43] was computed for a 5 kb region upstream the TSS of genes for which an associated LUAT was expressed in mouse ES cells (FPKM >1). In panels **B**, **C**, **I** and **J**, $p$ values of the Wilcoxon rank sum test are shown.

H3K27me3 within the 2 kb around the TSS, as compared to unidirectional and coding-coding gene sets (Mann–Whitney U test: $p = 0.007$ and $p = 0.001$, respectively), but not for H3K4me3 (Figure 5E). Moreover, they were more frequently associated with bivalent domains (Figure 5F), composed of concomitant H3K27me3 and H3K4me3 peaks, a feature related to silencing of developmental genes in ES cells, while keeping them poised for activation [37,38]. One additional expectation for developmentally regulated promoters is that they should be bound by tissue specific transcription factors. To test this, we analyzed ChIP-seq data performed in DP thymocytes for a series of lymphoid-specific transcription factors, including GATA3 [39], Ikaros [40], ETS1 [32] and RUNX1 (this study). These transcription factors were found in a higher frequency at LUAT-associated promoters active in DP thymocytes, as compared with a set of control genes (Figure 5G). Thus, LUAT-associated promoters appear to be regulated in a more specific way than other similarly expressed genes.

Overall, we found that LUAT-associated promoters share characteristics with those of developmentally regulated genes. It seems likely that the divergent transcription observed at LUAT-associated promoters is linked to intrinsic genomic characteristics of these promoters.

### Bidirectional transcription at LUAT-associated promoters is linked to a specific GC skew profile and longer first exon

Directionality of transcription is thought to be mediated, at least in part, by an asymmetric distribution of G and C content between the two DNA strands around the promoter, a property known as GC skew [41], possibly constraining the orientation of the transcription initiation complex. GC-rich promoters are characterized by a significant excess of G over C residues (positive GC skew) immediately downstream the TSSs [42]. To test the link between GC skew and bidirectionality, we computed GC skew profiles for each of the three gene promoter groups described above. As expected, unidirectional genes show a positive GC skew immediately downstream the TSS, while coding-coding genes show two sharp and inverted GC skew peaks, one negative and one positive, respectively upstream and downstream the TSS (Figure 5G). Strikingly, LUAT-associated genes also display two inverted GC skew profiles, but the GC bias is less pronounced than at both sides of the TSS ($p < 5 \times 10^{-5}$ and $p < 2 \times 10^{-10}$; $t$-test for the regions [−500 bp; TSS] and [TSS; +500 bp], respectively). In addition, the positive GC skew downstream the TSS is also less pronounced than at unidirectional genes ($p < 1 \times 10^{-5}$; $t$-test for the region [TSS; +500 bp]). This result suggests that bidirectional transcription at LUAT-associated promoters might be linked to a unique GC skew profile resulting in lower constraints on the directionality of the transcription initiation complex.

Promoter-proximal 5′ splicing sites and first exon length have been recently suggested to play a role in directionality of transcription [43-45]. We computed the average length of the first exon of genes in each gene set (Figure 5I). Strikingly, the set of LUAT-associated genes has the longest first exon with a median length of 242 bp, compared to 195 bp and 190 bp for the unidirectional and coding-coding gene sets. Consistently, 5′ splicing sites are relatively depleted immediately downstream the TSS of LUAT-associated genes, as compared to the control set of genes (Additional file 11: Figure S6). To assess whether splicing might play a role in controlling the expression of LUATs, we analyzed recent published data of 3′ ends of polyadenylated RNA-seq in mouse ES cells in which U1 small nuclear ribonucleoprotein (snRNP) has been functionally inhibited [43]. Interestingly, the expression (as measured by the level of 3′ ends of polyadenylated RNA) of a selection of LUAT normally expressed in ES cells (FPKM > 1) is significantly increased after inhibition of U1 snRNP (Figure 5J), as compared to control ES cells. Taken together, these observations indicate that a unique GC skew profile along with large first exon might both contribute to the bidirectionality of transcription at LUAT-associated genes.

### LUAT-associated genes harbor specific chromatin features

In order to assess whether LUAT-associated promoters display specific chromatin features, we analyzed several histone modification marks and general transcription factors in DP thymocytes that were either performed in this study or already published [32] (Figure 6). We compared the surrounding regions of three sets of promoters displaying similar expression level distribution based on Total RNA-seq signal at exons (FPKM) in DP thymocytes: LUATs-associated promoters, unidirectional promoters, as well as promoters of bidirectional coding gene pairs (coding-coding set). As expected, we found chromatin features common to both sets of bidirectional promoters (Figure 6A). Bidirectional promoters display higher and/or wider level of histone modifications linked to open and active chromatin (H3K4me1/2/3 and H3K27ac) at the region immediately upstream of the TSS (Figure 6A), consistent with the bidirectional activity at these promoters. This property is also associated with additional peaks upstream of the TSS for either total or initiating (Ser5 phosphorylated: Ser5P) Pol II, as well as TBP (Figure 6). Furthermore, significant levels of histone modifications linked to early (H3K79me2) or late (H3K36me3) transcription elongation were observed in the region upstream of bidirectional promoters, confirming the fact that these regions undergo productive transcription (Figure 6A).

Strikingly, we also found evidences for chromatin features specific to LUAT-associated promoters. Although the three gene sets are similarly enriched for H3K36me3 within the coding gene body, the LUAT-associated genes display higher levels of H3K79me2 (Figure 6A). H3K79me2 is generally enriched at the 5′ end of expressed genes comprising the first exon and intron and mark the transition between early (immature) and late (productive) transcription elongation [46]. To have a more precise view of the differential enrichment in H3K79me2, we compared the H3K79me2 profiles within rescaled regions comprising the TSS to the end of 1$^{st}$ intron (Figure 6C and D). Indeed, we observed that H3K79me2 remains significantly higher throughout the first intron of LUAT-associated genes as compared to the control gene sets. As the genes from the three sets express similar amounts of productive transcription (as assessed by both Total and PolyA RNA-seq counting at exons; Figure 6A), our results most likely suggest an actual increased rate of early (immature) transcription elongation from LUAT-associated promoters.

In agreement with an increased rate of early transcriptional elongation at LUAT-associated genes, we observed aspecific enrichment of Pol II (either total or Ser5P) and TBP within the 5′ region of LUAT-associated genebodies (Figure 6B and Additional file 12: Figure S7), indicating extended accumulation of the initiating and/or early elongating transcription complex [47]. To further investigate the possibility of a transcriptional pause immediately downstream the TSS, we analyzed the average profiles in DP thymocytes of additional general transcription factors (GTFs), including TAF1, TFIIB, TFIIE and TFIIH (Additional file 13: Figure S8). In all cases, we observed a significant enrichment of the GTFs downstream the TSS. Finally, consistently with the transcriptional pause being directly linked to divergent transcription, we also observed a significant and specific enrichment of (Ser5P)Pol II and GTFs around a region 1 kb upstream the TSS of LUAT associated genes (Additional file 12: Figure S7). Of particular interest is the overall enrichment in TFIIH complex around the TSS of LUAT-associated genes, which play a key role in transcription initiation by phosphorylating Pol II at Ser5 [47]. Thus, our results indicate increased Pol II pausing at both sides of LUAT-associated promoters.

### Accumulation of immature transcripts at LUAT-associated genes

Early elongation and H3K79me2 enrichment are generally associated with the 5′ intronic sequences and splicing events [46]. Thus, to further explore the hypothesis of an accumulation of immature transcripts at LUAT-associated genes, we compared the average profiles around the TSS of Total and PolyA RNA-seq levels for the three set of equally expressed genes (Note that these sets of genes have equal distribution of exonic FPKM based on either Total or PolyA RNA-seq, data not shown). We found that Total-RNA signal downstream of the TSS is higher for LUAT-associated genes, while PolyA-RNA signal is similar among the three gene sets (Figure 7A, compare left and right panels). While the PolyA RNA-seq signals result only from complete (fully processed) transcripts, Total RNA-seq signals result from both immature (partial or unprocessed) and complete transcripts. Thus, a relative enrichment of Total RNA-seq, as compared to PolyA RNA-seq signal, is indicative of either increased rate of immature transcription or expression of less stable transcripts. However, the observed results could not be attributed to differences in transcript stability as the three gene sets display equivalent levels of H3K36me3, which is generally coupled to productive elongation (Figure 6A).

To directly assess whether LUAT-associated genes produce increased level of immature transcripts (i.e. more unspliced transcripts), we computed the splicing index across the three gene sets. As shown in Figure 7B, the splicing index is significantly reduced between the first two exons of LUAT-associated genes as compared to the other set of genes, while no differences are observed at the middle exons. This suggests an increased rate of immature transcription at the 5′ region of LUAT-associated genes. Consistently, we found that first introns of LUAT-associated genes display higher density of Total RNA-seq signal as compared to the control gene sets (Figure 7C),

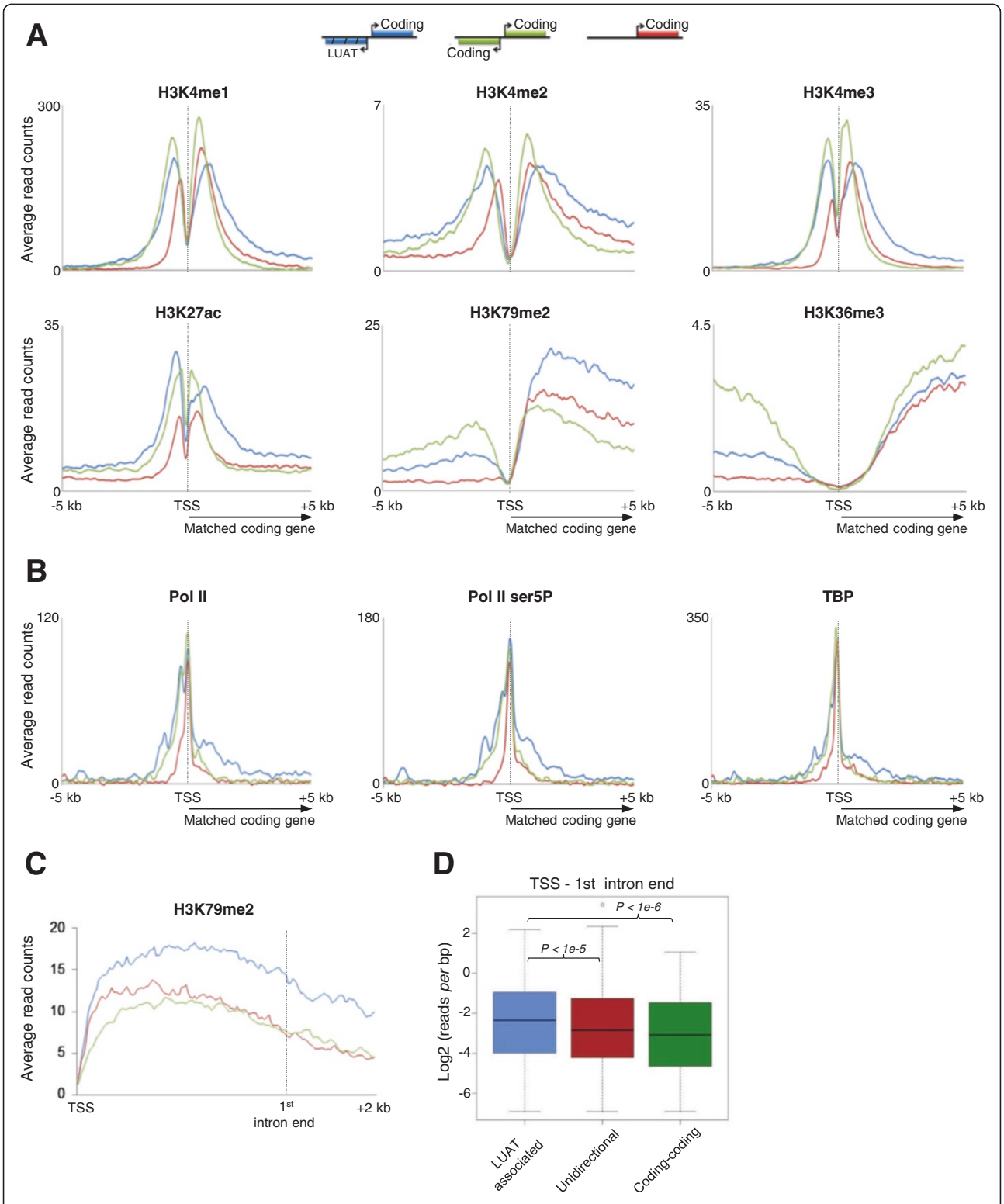



**Figure 6 Chromatin characteristics of LUAT-associated promoters in DP thymocytes.** Average profiles of ChIP-seq signals of the indicated histone modifications **(A)** and general transcription factors **(B)**. The 5 kb region around the TSS of the genes included in the indicated sets are shown: LUAT-associated genes, bidirectional coding-gene (coding-coding) and unidirectional genes. The three gene sets were selected ensuring a similar distribution of gene expression levels in DP thymocytes (see Methods). **C)** Average profile of ChIP-seq signals for H3K79me2 in rescaled region ranging from the TSS to the end of the first intron. **D)** Boxplots showing the distribution of reads per bp for H3K79me2 within the region comprising the TSS and the end of the first intron (*p*-values of the Wilcoxon rank sum test are shown).

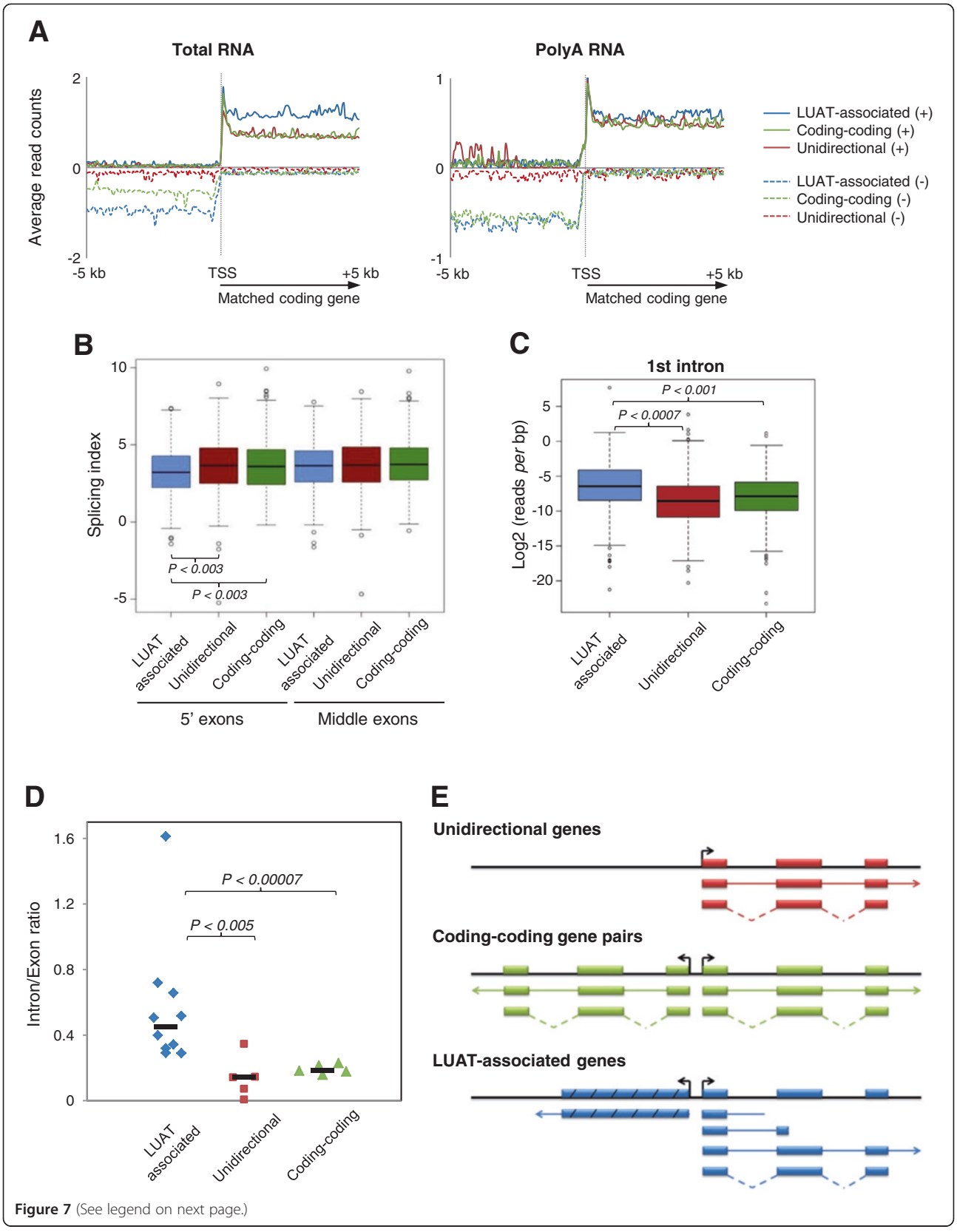



**Figure 7** (See legend on next page.)

(See figure on previous page.)
**Figure 7 LUAT-associated promoters are prone to pervasive transcription. A)** Average profiles of Total and PolyA RNA-seq signals in DP thymocytes, for the three set of similarly expressed genes. Signals coming from plus and minus strands are indicated by solid and dashed lines, respectively. **B)** Splicing index calculated for the 5′ and middle exons for the three set of similarly expressed genes in DP thymocytes. **C)** Boxplots showing the density of Total RNA-seq reads *per* bp in the same orientation as the matched coding genes and within the first intron of the three group of genes in DP thymocytes. Statistical significance was assessed by the Mann–Whitney U test. **D)** Intron/exon ratio of individual genes for the three gene sets in DP thymocytes assessed by reverse transcription quantitative PCR. Relative transcript levels at the first intron and the last exon of each gene was estimated based on a standard dilution of genomic DNA. Statistical significance were assessed using Wilcoxon rank sum tests. **E)** Schematic representation of RNA processing at the three different classes of gene loci. Exons are shown by rectangles (or stripped rectangles in the case of LUATs). Solid and dotted lines represent immature (unspliced) and processed (spliced) transcripts, respectively. Our results suggest that LUAT-associated genes display an increased rate of immature transcripts.

though there is no significant differences in first intron length between the three set of genes (data not shown). Significant enrichment of immature transcripts within the first intron of LUAT-associated genes was also confirmed by RT-qPCR analysis of individual genes (Figure 7D). All in all, LUAT-associated gene promoters are clearly more prone to induce immature transcription than other bidirectional or unidirectional gene promoters, indicating a less restricted control of Pol II pausing around the TSS, likely leading to divergent and pervasive transcription.

## Discussion

Here, we used directional RNA-seq from primary mouse tissues to directly and systematically characterize antisense transcripts. We have found that a significant fraction of gene promoters sustains expression of long non-coding antisense transcripts (here named LUATs). The LUAT/coding gene pairs are usually co-regulated throughout cell differentiation and development and generally function as transcriptional regulators. LUAT-associated promoters share several characteristics of promoters of developmentally regulated genes, including large CpG islands, high degree of conservation and epigenetic regulation during development. Moreover, the bidirectional transcription at these promoters appears to be linked to a specific GC skew profile and particularly longer first exons. Finally, LUAT-associated promoters display distinctive epigenetic features reflecting an intensified rate of early transcriptional elongation. Overall, our results support the view that promoters of a subclass of transcriptional regulators are characterized by a specialized mechanism of Pol II transcription, in which Pol II recruitment is directly coupled to relaxed bidirectional transcription.

Previous studies have shown that lincRNAs are preferentially located near protein-coding genes in divergent orientation and that their expression is often correlated [19,20,48]. However, the stringent criteria generally used to define lincRNAs (e.g., the presence of at least two exons) likely impaired a comprehensive identification of lncRNA transcripts, including those in divergent orientation from coding genes. A recent study has indeed described the abundance of divergently transcribed LncRNA/mRNA gene pairs in ES cells [21]. Although this and the present works likely described the same kind of antisense transcripts, our study largely complement and extend the previous study by using a more direct approach to identify upstream antisense lncRNAs solely based on the RNA-seq signal and by demonstrating their existence in many primary tissues and throughout T cell differentiation. Moreover, we show here that divergent transcription is clearly associated with a subset of genes coding for transcriptional regulators, and we propose a functional link between divergent transcription and gene expression regulation.

As suggested previously the presence of short bidirectional transcripts appears to be a hallmark of active promoters in mammals [12,13], generally associated with paused Pol II around the TSS. It has also been previously described that expression of upstream unstable transcripts (also called PROMPTs) are a common characteristic of Pol II transcribed genes [16,17]. Although some overlaps might exist between PROMPTs and LUATs, LUATs differ from exosome-sensitive PROMPTs transcripts. Firstly, LUATs are detected at significant levels without inhibition of the RNA exosome degradation machinery. Secondly, they are substantially longer than PROMPTs. Thirdly, they are associated with a specific category of genes. Hence, while many coding gene promoters, essentially those containing CpG islands [49], have the intrinsic property of bidirectional transcription [11], production of long antisense transcript is generally inhibited or are rapidly degraded at most loci.

### Divergent transcription is linked to intrinsic sequence properties shared with developmental gene promoters

We have found that LUAT-associated genes frequently encode TFs involved in cell differentiation and development. This is consistent with previous studies in mouse, human and zebra fish suggesting that large gene deserts flanking TF genes, with roles in embryonic development, preferentially harbor lincRNAs [50-54]. As such, these genes are expected to be subjected to fine tune regulation. Accordingly, we found that their promoters appears to be bound by lineage specific TFs (Figure 5G). Another striking characteristic of LUAT-associated promoters is the presence of very large CpG islands (Figure 5A and B),

a feature shared with developmental gene promoters [34], but not observed at bidirectional coding-gene promoters. As the promoters of these genes are enriched for CpG rich regions and are prone to pervasive transcription, it is expected that dedicated repressive mechanisms might be in place to control their expression. In line with this, we have found that LUAT-associated genes are specifically enriched for H3K27me3 (Figure 5E) and for bivalent domains (Figure 5F) in ES cells, thus suggesting Polycomb-mediated regulation of these genes. As suggested elsewhere [55], large CpG islands (as those found at LUAT-associated promoters) might represent a favored recruitment platform for Polycomb-associated complexes and thus play an important role in transcriptional regulation of key developmental genes. Whether large CpG islands are required for divergent transcription from LUAT-associated promoters will deserve further investigation.

#### A link between antisense and immature transcription?

Our results indicate an unexpected link between divergent lncRNA/mRNA transcription and premature termination of coding transcripts. Indeed, LUAT-associated genes are characterized by a significant accumulation of immature transcripts within the 5′ region of the genes (Figure 7). Our results most likely suggest that intrinsic properties of LUAT-associated promoters induce a specialized mechanism of Pol II transcription, in which recruitment of the enzyme is directly coupled to pervasive bidirectional transcription. Several arguments point to this direction. The presence of a TATA box is generally linked to strong directionality of transcription. Consistent with their bidirectional feature, LUAT-associated promoters are essentially depleted of TATA box (Figure 5C). Although the absence of a TATA box might be required for bidirectional transcription, as has been suggested for coding-coding promoters [56,57], it might not be sufficient as some TATA-less promoters still show strong directionality [57]. In addition, we have also observed that LUAT-associated promoters display two inverted, though moderated, GC skew profiles around the TSS (Figure 5E). It has been shown that a positive GC skew immediately upstream the TSS favors the formation of so-called R-loop structures [42], which are long, stable and three stranded RNA/DNA structure formed during transcription [58]. It has also been previously suggested that positive GC skew immediately after the TSS, and subsequent R loop formation, may serve to correct the lack of directionality in the initial steps of transcription [42]. This correction might be mediated by the ability of R loops to elicit transcriptional pausing [59-61]. Finally, 5′ splice sites and first exons have been recently shown to play a role in promoter directionality [43-45]. It is suggested that U1 snRNP binding at 5′ splice sites might help to stabilize Pol II recruitment at the promoter and enforce sense transcription. Moreover, the length of first exon appears to influence Pol II stability at promoters and transcription accuracy [45]. Indeed, genes with long first exon have Pol Il accumulation downstream the TSS and increased proportion of antisense transcripts [45], reminiscent of what we have observed at LUAT-associated loci. Strikingly, functional inhibition of U1 snRNP in ES cells resulted in accumulation of premature cleavage of coding gene transcripts [43], but also increased expression of antisense transcripts (Figure 5J). Consequently, spatial separation of promoter and 5′ splice sites might directly impact both directionality and transcription accuracy. It is plausible that, at LUAT-associated genes, 5′ splice sites are too distant from the TSS to ensure proper directionality of transcription, thus favoring bidirectional transcription. Taking all into account, we favor a hypothesis whereby both immature and bidirectional transcription at a specialized class of gene promoters are directly linked to intrinsic sequence properties, including TATA box depletion, unique GC skew profile and long first exon.

#### Functional relevance of bidirectional transcription

The specific association of antisense transcripts with genes related to transcription regulation and development strongly suggests that divergent transcription might be directly or indirectly involved in the tight regulation of these genes. In line with a direct role of LUATs, several examples have demonstrated a functional regulation by mammalian antisense transcripts in *cis* [11,62,63]. This regulation might act at multiple levels, from modifying local chromatin to enabling regional signal spreading [11,62] or controlling translation efficiency [10], yet in the latter case an overlapping stretch with coding exons is required. Alternatively, divergent transcription might reflect an intrinsic property of promoters of genes coding for developmental regulators. The finding that H3K79me2, a mark of early elongation, was found higher at LUAT-associated genes (Figure 6A) suggests the level of early elongating transcription is increased at these genes. This was further supported by a relative higher ratio of Total *versus* PolyA levels within the first intron, as well as decreased splicing efficiency (Figure 7A-C). Thus, our findings clearly indicate that LUAT-associated promoters are more prone to pervasive and/or immature transcription (Figure 7D).

What can be the functional relevance of this pervasive transcription? If divergent transcription and non-productive gene expression are linked, this will imply that the initiation step of transcription is less controlled at these promoters, likely leading to pervasive transcription in both directions. In this line, an appealing hypothesis will be that expression of LUAT-associated genes is regulated also at the level of transcript maturation and/or elongation efficiency. This might reflect a checkpoint for coupling

elongation and RNA processing, as previously suggested [64]. Although some genes are regulated by Pol II pausing in close proximity to the TSS [65], others are reported to be efficiently transcribed into precursor transcripts, while the efficiency of transcript processing is being regulated [66,67]. Generally, only a portion of the Pol II that assemble at the promoter enters into early elongation [68]. This entry is characterized by Pol II Ser5 phosphorylation and chromatin modifications that are specific to transcription initiation and early elongation (such as H3K79me2). Before transcribing further, the polymerase passes a 5′ checkpoint where it pauses, terminates, or commits to productive elongation. If the polymerase does not proceed through this checkpoint, transcription will be terminated producing an unstable transcript. If the polymerase proceeds through the checkpoint, it will enter into productive elongation that is associated with characteristic chromatin modifications (such as H3K36me3) and Pol II Ser2 phosphorylation [47]. In this context, rapid transcriptional induction might be facilitated by the active release of polymerase molecules that have initiated transcription, but are paused downstream the TSS. Thus, pausing during early elongation may provide both a kinetic 'window of opportunity', as well as an interaction surface, to facilitate additional levels of regulation of the nascent RNA before the transition to productive elongation.

Another related and not exclusive hypothesis would be that bidirectional promoter activity increase the stochasticity of gene expression, as suggested for antisense expression in yeast [69]. It is well known that expression of developmental regulators, including key transcription factors, is often regulated stochastically during cell differentiation, influencing cell and developmental decisions [70,71]. It is therefore plausible that LUAT expression might reflect a dedicated mechanism to induce stochastic expression of developmental regulators by modulating the rate of pervasive (i.e. non-productive) transcription. In any case, our observations might shed light on a new mechanism representing the outcome of an evolutionary pressure to control the expression of a subclass of genes coding for transcriptional regulators.

## Conclusions

We have found that divergent transcription of lncRNA/mRNA gene pairs is significantly associated with genes related to transcriptional regulation and development. Their promoters share several characteristics with those of developmental transcription factors, display a unique GC skew profile and are associated with genes harboring long first exons, reminiscent of their bidirectional activity. Unexpectedly, we also found that the 5′ region of the associated coding genes harbor a specific accumulation of H3K79me2 epigenetic mark, as well as initiating Pol II complexes, both of which are linked to an intensified rate of early transcriptional elongation. These results strongly suggest a functional link between divergent transcription and the regulation of genes coding for developmental transcription factors.

Altogether, our findings which indicate that, at a subset of transcription factor gene promoters, divergent and pervasive transcription are linked, might illustrate an additional mechanism for regulating the regulators, in a rather sophisticated system for fine-tuning mammalian gene expression.

## Methods

### Mice and cell preparations

Homozygous *Rag2*-deficient (ΔRag) mice [72] were housed under specific pathogen free conditions and handled in accordance with French and European directives. All mice were bred on a C57Bl/6 J background and were killed for analysis between 4 and 6 weeks of age. DN and DP thymocytes were purified as previously described [32,73].

### RNA extraction and library preparation

Total RNA from $10\times10^6$ thymocytes of ΔRag mice was extracted as previously described [32]. Strand-specific preparation, sequencing and processing of Total and PolyA RNA samples were carry out as previously described [32]. RNA quantity and quality were verified using RNA Pico chips on a 2100 Bioanalyzer (Agilent).

### RNA-seq datasets

Paired-end stranded RNA-seq was performed with ribo-depleted Total RNA obtained from ΔRag thymocytes using SOLiD platform version 4. Single-end stranded RNA-seq was also performed on ribo-depleted Total and PolyA RNA obtained from ΔRag thymocytes using Illumina GAII sequencer. Strand-specific Total and Poly(A) RNA-seq from double-positive T-cells were obtained from SRA ftp site (SRX063934, SRX063935 respectively). RNA-seq data from thymocytes across development (DN1, DN2a, DN2b, DN3 and DP) where obtained from SRA ftp site (SRP007822). RNA-seq data performed using fractionated chromatin-, nucleoplasmic-, and cytoplasmic-associated transcripts where obtained from SRA ftp site (SRX100837, SRX100832 and SRX100827 respectively). RNA-seq data from mouse tissues were obtained from SRA (SRP006787). For quality filtering, sequencing read obtained from Illumina and SoliD platforms were quality trimmed using sickle (–q10) or csfasta_quality_filter (–m 8 -l 25 -s 14 -n 2), respectively. Selected reads (minimum length 25 nt) were then aligned to mm9 genome using TopHat (version 2.0.4) [24]. Gene annotations (gtf file) and indexes (nucleotide or color space) were obtained from TopHat website (mm9 iGenome). Multireads were rejected from all experiments. Additional information

about mapping results for RNA-seq experiments is available in Additional file 1: Table S1.

### Quantitative RT-PCR

RNA was isolated from 6 weeks old C57BLK6 mice thymi using TRIzol® reagent (Invitrogen). RNA quality was monitored with RNA Nano-6000 Chips and 2100-BioAnalyzer (Agilent). Two-step quantitative reverse transcription PCR (RT-qPCR) was performed using the Stratagene Mx3000P Sequence Detection System. Random hexamers and the reverse transcriptase SuperScript II (Invitrogen) were used for RNA reverse transcription. Quantitative PCR was performed with Syber® Green PCR Mix (Applied Biosystem). Primers were designed in the first Intron and the last exon of selected unidirectional, coding-coding and LUAT-associated genes (primer sequences are provided in Additional file 14: Table S6). Relative transcript levels were estimated based on a standard dilution of genomic DNA and the intron/exon ratio was calculated for each gene.

### Identification of LUATs

We first selected the +/−5 kb regions around the TSS of all coding transcripts available from RefSeq database (mm9, UCSC). In case several TSSs originating from several isoforms of the same gene were distant from less than 100 bp, only one representative transcript was selected. To exclude coding gene whose promoter overlap any other coding transcript, we then filtered out transcripts whose upstream 5 kb region overlapped with any coding transcript from another gene both on positive and negative strand. Coverage was then measured (both on positive and negative strand) in binned regions (100 bp) around the TSS of the selected transcripts using coverageBed program (bedtools suite, version 2.13.3) [74] and expressed as log2(coverage +1). Transcripts were sorted according to the sum of bin coverage on the opposite strand of upstream 5 kb region. The subsequently obtained heatmap was visualized using treeview. In order to define a set of candidate coding genes displaying a significant signal in opposite strand within the upstream 5 kb regions, the same procedure (binning and coverage analysis) was applied to a set of 10,000 randomly selected intergenic regions. The distribution of the bin coverage sum obtained for all randomly chosen regions was used to define a threshold with $p$-value < 0.005. Cufflinks [24] was subsequently used to discover new isoforms and transcripts (using known transcript list as guide), and to perform assembly.

### Quantification of LUAT expression

We observed that inferred cufflinks transcripts appeared generally fragmented compared to the underlying RNA-seq signal, probably due to mapability issues (i.e. low complexity sequences), low expression levels or uneven coverage. Thus, to maximize the assembly of full length transcripts, cufflinks fragments closer than 800 bp were combined. We thus developed a python script that used novel Cufflinks transcripts located within 1.5 kb from the TSS of selected candidate coding genes as seed and extend them in 5′ and 3′ orientation while any novel cufflinks of length above 200 bp, present on the same strand and distant from less than 800 bp is found. A gtf file was subsequently produced containing coordinates of novel transcripts overlapping promoter regions. This file was merged with a gtf file containing coding genes and used for transcript abundance estimation (FPKM) using cufflinks (using the -G parameter). The list of LUAT/gene pairs, including genomic coordinates and FPKM, for the different data sets described in this study is provided in the Additional file 3: Table S2, Additional file 6: Table S3 and Additional file 8: Table S5.

### Definition of gene sets

Given the list of LUAT-associated genes found in DN or DP thymocytes, we first filtered out the genes having several alternative TSSs. We then generated two control sets of the same size. A set of coding genes with no overlapping transcript in their 5 kb upstream region (defined as unidirectional genes). A set of coding genes having another coding gene in their upstream region, oriented in the opposite direction, and with TSS separated less than 1.5 kb from each other (defined as coding-coding genes). The selection of the two control sets was then performed with a random sampling procedure implemented in R, ensuring similar distribution of gene expression (as measured by the exonic FPKM from Total RNA-seq data) in all 3 gene sets. In the case of coding-coding genes, the TSS used to anchor the average plots was the one corresponding to the matched expression. For multi-tissue analysis, where no unique reference expression level exists, the two control sets were chosen by selecting genes for which the maximum of expression across the 17 tissues matched those of the LUAT-associated gene list (Additional file 7: Table S4).

### Thymocytes-specific genes

We retrieved gene expression data from the GNF Gene Atlas [75], using samples from a large variety of tissues. For each gene, we computed a score of tissue-specificity $TS_{g,t} = (e_{g,t} - Q3_g)/(Q3_g - Q1_g)$, where $e_{g,t}$ is the expression of gene g in tissue t, $Q1_g$ and $Q3_g$ are the first and third quartiles in the distribution of expression values for gene g across all tissues. Genes with scores higher than 1 were considered as outliers of that distribution, so a gene g was called specific of a tissue t if $TS_{g,t} > 1$. For Figure 3A, we used scores associated to the thymus.

### ChIP-seq data and analysis

Chromatin preparation and immunoprecipation for ChIP was performed as described previously (Koch et al. [32]). H3K79me2 ChIP was performed from sonicated chromatin from 5 million DP thymocytes using 2 μg of antibody (ab3594, Abcam). Runx1 ChIP was performed from sonicated chromatin from 10 million DP thymocytes using 10 μg of antibody (ab3594, Abcam). ChIP samples were subsequently sequenced in either Genome Analyzer II (Illumina, USA; H3K79me2) or AB SOLiD V4.0 (Life Technologies; Runx1) according to the manufacturer's instructions. ChIP-seq data from mouse DP thymocytes for Pol II, TBP, General transcription factors TAF1, TFIIB, TFIIE and TFIIH, as well as other histone modifications have been previously published (Koch et al. [32]) and were analyzed as described in (Koch et al. [32]). ChIP-seq datasets for additional transcription factors in DP thymocytes were downloaded from Gene Expression Omnibus (IKZF1: GSE32311, ETS1: GSE29362, GATA3: GSE20898 merged with GSE31233, Input: GSE31233 and GSE32311). H3K4me3 and H3K27me3 ChIP-seq data from mouse ES cells were obtained from [37]. To generate average profiles, mm9 Refseq genes annotations were used to extract values from wiggle files associated with selected regions. The selected regions are defined in a region of 5 kb before and after TSSs of gene list selections. A bin scores from wiggle files were used to interpolate around the TSS and generates the average profiles. For assessing binding of lymphoid-specific transcription factors in Figure 5G, peak calling was performed by using the Hypergeometric Optimization of Motif EnRichment (HOMER) tool (v4.1) [76] with default settings (FDR: 0.001; local and input fold enrichment: 4.0). We computed the overlap between transcription factors peaks and regions around TSS (+/− 5 kb) for the indicated group of genes.

### Functional enrichment analysis

GO term analysis were performed with DAVID [77]. In the analyses shown in Figures 2A and B, we selected, for each category, the terms with a Benjamini-corrected *p* value below 0,001 using the "Functional annotation chart" and default options. For the analyses shown in Figure 2C, we used "Functional annotation clustering" and selected the top 10 clusters retrieved for each gene set. Assessment for functional enrichment of "developmental transcription factor" in the list of LUAT-associated coding genes was performed using R and TBrowser database [78]. GO data were first retrieved and genes associated both to terms "developmental process" (GO term GO:0032502, biological process ontology) and "sequence-specific DNA binding transcription factor activity" (GO term GO :0003700, molecular function ontology) were defined as "developmental transcription factor". A contingency table was then created using the list of genes annotated in both biological process and molecular function ontology as a reference. *p*-value was obtained using Fisher's exact test.

### Splicing index

For computation of splicing index, only coding RefSeq transcripts with at least 4 exons and FPKM above 0.1 were selected. The coverage of their exonic and intronic features was computed using coverageBed (from the bedtools suite) and a pseudo-count added to ensure non-zero values. FPKM values where then computed for each feature. The 5′ exonic signal was computed by averaging FPKM values corresponding to the first and second exons. For gene displaying an even number of exons the middle exonic RPKM was computed as the average signal between the two central exons whereas for genes displaying an odd number of exons the signal corresponding to the central exon was used. The splicing index corresponds to the log ratio between exonic FPKM value and intronic FPKM value.

### Availability of supporting data

Original ChIP-seq and RNA-seq data used in this study have been submitted to the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE44578.

## Additional files

**Additional file 1: Table S1.** Information about RNA-Seq datasets used in this study. The number of input reads and subsequent alignements are indicated.

**Additional file 2: Figure S1.** Identification of genes associated with long upstream antisense transcripts in embryonic kidney. A) Heatmap showing the Total RNA-seq signal from mouse embryonic Kidney (Thiagarajan et al. [23]) found in a [−5000;+5000] region around the TSS of all non-overlapping Refseq genes. Signal was computed based on number of reads per 100 bp binned regions originated from either antisense or sense strand with respect to gene annotation (left and right panels, respectively). The heatmap is ordered according to the antisense signal for the [−5000;0] region. B) Example of genes associated with LUAT in mouse kidney. Total RNA-seq signal from embryonic kidney (Thiagarajan et al. [23]) and ΔRag DN thymocytes (SOLiD platform, this study) are shown. Signals from plus and minus strands are displayed in blue and red respectively. The *Dntt* gene is shown as an example of T-cell specific LUAT-associated gene. The arrow highlights the presence of a LUAT.

**Additional file 3: Table S2.** LUAT and associated coding genes found in mouse DN thymocytes (SOLiD platform).

**Additional file 4: Figure S2.** Assessment of coding potential. PhyloCSF assessement of LUAT coding potential. The Galaxy web server (https://main.g2.bx.psu.edu/) was used to extract MAF blocks from 46-way multiZ alignments using cufflink transcript coordinates as input. Corresponding genomic sequences for human (hg19), Mus musculus (mm9), Rattus norvegicus (Rn4), Bos taurus (bosTau4) and Canis familiaris (canFam2) were retrieved for each block. Blocks shorter than 50 bp (95% of mouse exons) or missing one of the selected species were discarded. In order to create a positive control list, a set of blocks with same length distribution was randomly selected in exons from coding transcripts. The PhyloCSF program was used to assess coding potential of both sets. The resulting log-likelihood ratios are reported in units of decibans.

**Additional file 5: Figure S3.** LUATs are mainly found in the chromatin-associated fraction. To gain insight into the cellular localization of LUAT transcripts we analyzed recently published RNA-Seq data obtained from fractionated chromatin-associated, nucleoplasmic, and cytoplasmic transcripts from mouse macrophages (Bhatt et al. [27]; GEO serie: GSE32916). We performed assembly of divergent transcripts observed in these fractions (see "Identification of LUATs" in Methods section). A) boxplot displaying expression level as log2(FPKM + 1) for coding genes (blue) and LUATs (red) in the three different subcellular fractions. B) Representative examples of RNA-seq profiles from the three different subcellular fractions. Signal is provided for both plus and minus strands. The arrow highlights the presence of a LUAT.

**Additional file 6: Table S3.** LUAT and associated coding genes found in the multi-tissue analysis.

**Additional file 7: Table S4.** Coding-coding and unidirectional gene sets for which the maximum of expression across the 17 tissues matched those of LUAT-associated gene list.

**Additional file 8: Table S5.** LUAT and associated coding genes found in mouse DN and DP thymocytes (Illumina platforme).

**Additional file 9: Figure S4.** Dynamic regulation of LUAT and their associated genes through early T-cell differentiation. In order to define expression profiles of LUAT and associated coding-genes through discrete stages of thymocyte development we retrieved unstranded PolyA RNA-seq from GEO web site (GSE31234, Zhang at al. 2012). Unstranded RNA-seq signal is shown for DN1, DN2a, DN2b, DN3 and DP (black track). Signal obtained from ΔRag DN thymocytes (SOLiD platform, this study) is also shown to highlight the expected signals from the plus (blue) and minus (red) strands. The arrow highlights the presence of a LUAT.

**Additional file 10: Figure S5.** Tissue-specificity of LUATs. The histogram shows the number of tissues in which a given LUAT was found in the multi-tissue analysis.

**Additional file 11: Figure S6.** 5′ splice site distribution. Analysis for 5′ splice site motifs (Jaspar database; ID SD0001.1) in the 500 nt regions downstream of TSS for the three group of genes used for the multi-tissue analyses (A) or LUATs (B). The *y* axis shows the cumulative fraction of regions having at least one predicted site after traversal of a given number of nucleotides, as indicated on the *x* axis.

**Additional file 12: Figure S7.** Detailed view of TSS-centered ChIP-seq profiles for Pol II and TBP in DP thymocytes. Legends are as in Figure 6. The highlighted regions in pink correspond to the 500 nt regions analyzed in B. B) Number of reads within the indicated regions for the corresponding ChIP-seq experiments shown in A. The *p*-values of the Wilcoxon test are shown.

**Additional file 13: Figure S8.** Detailed view of TSS-centered ChIP-seq profiles for the indicated general transcription factors in DP thymocytes. Legends are as in Figure 6. The highlighted regions in pink correspond to the 500 nt regions analyzed in B. B) Number of reads within the indicated regions for the corresponding ChIP-seq experiments shown in A. The *p*-values of the Wilcoxon test are shown.

**Additional file 14: Table S6.** Primers used for the RT-qPCR assays.

### Competing interests
The authors declare that they have no competing interests.

### Authors' contributions
CL, MB, AB and DP performed the main bioinformatics analyses of the manuscript. AB, DP and RF processed deep sequencing data. AG analyzed ChIP-seq data from transcription factors. MY performed RT-qPCR assays, LV, JZC, FK, MAB performed RNA-seq and ChIP-seq experiments. BL, HL, JI, MG and IG performed deep sequencing. JCA, DP and SS conceived the study, and participated in its design and coordination. CL, DP and SS wrote the manuscript. All authors read and approved the final manuscript.

### Acknowledgements
Research in SS laboratory was supported by recurrent funding from the Inserm and Aix-Marseille University, and by specific grants from the European Union's FP7 Program (agreement n° 282510-BLUEPRINT), the ARC (project n° SFI20111203756) and the A*MIDEX project (n° ANR-11-IDEX-0001-02) funded by the « Investissements d'Avenir » French Government program. The work was also granted for sequencing costs by an ESGI Consortium grant of the EU (to JCA, program T-DynRegSeq) from the Seventh Framework Program (FP7/2007-2013) under grant agreement no. 262055. The TGML sequencing platform is supported by grants from GIS IBiSA and France Génomique ANR Investissements d'Avenir (to JI).

### Author details
[1]Technological Advances for Genomics and Clinics (TAGC), Case 928, 163 Avenue de Luminy, 13288, Marseille cedex 09, France. [2]Aix-Marseille Université, UMR1090 TAGC, Marseille F-13288, France. [3]INSERM, UMR1090 TAGC, Marseille F-13288, France. [4]TGML, IBiSA Platform, Parc Scientifique de Luminy, Marseille, France. [5]Centre d'Immunologie de Marseille-Luminy, Campus de Luminy, Case 906, 13288 Marseille cedex 9, France. [6]CNRS UMR6102, Marseille, France. [7]Inserm, U631 Marseille, France. [8]Centre Nacional D'Anàlisi Genómica, Parc Científic de Barcelona, Baldiri i Reixac, 08028 Barcelona, Spain. [9]Polytech Marseille, Parc Scientifique de Luminy, Marseille, France. [10]Present address: CNRS, Aix-Marseille Université, IGS UMR7256, Marseille, France.

Received: 10 December 2013 Accepted: 18 December 2013
Published: 23 December 2013

### References
1. Kapranov P, Willingham AT, Gingeras TR: **Genome-wide transcription and the implications for genomic organization.** *Nat Rev Genet* 2007, **8:**413–423.
2. Esteller M: **Non-coding RNAs in human disease.** *Nat Rev Genet* 2011, **12:**861–874.
3. Guttman M, Rinn JL: **Modular regulatory principles of large non-coding RNAs.** *Nature* 2012, **482:**339–346.
4. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, *et al*: **The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression.** *Genome Res* 2012, **22:**1775–1789.
5. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, *et al*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447:**799–816.
6. Tisseur M, Kwapisz M, Morillon A: **Pervasive transcription - Lessons from yeast.** *Biochimie* 2011, **93:**1889–1896.
7. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A, *et al*: **The reality of pervasive transcription.** *PLoS Biol* 2011, **9:**e1000625. discussion e1001102.
8. Magistri M, Faghihi MA, St Laurent G III, Wahlestedt C: **Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts.** *Trends Genet* 2012, **28:**389–396.
9. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, *et al*: **Antisense transcription in the mammalian transcriptome.** *Science* 2005, **309:**1564–1566.
10. Carrieri C, Cimatti L, Biagioli M, Beugnet A, Zucchelli S, Fedele S, Pesce E, Ferrer I, Collavin L, Santoro C, *et al*: **Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat.** *Nature* 2012, **491**(7424):454–457.
11. Wei W, Pelechano V, Jarvelin AI, Steinmetz LM: **Functional consequences of bidirectional promoters.** *Trends Genet* 2011, **27:**267–276.
12. Core LJ, Waterfall JJ, Lis JT: **Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters.** *Science* 2008, **322:**1845–1848.
13. Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA: **Divergent transcription from active promoters.** *Science* 2008, **322:**1849–1851.
14. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, *et al*: **RNA maps reveal new RNA classes and a possible function for pervasive transcription.** *Science* 2007, **316:**1484–1488.
15. Wyers F, Rougemaille M, Badis G, Rousselle JC, Dufour ME, Boulay J, Regnault B, Devaux F, Namane A, Seraphin B, *et al*: **Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly (A) polymerase.** *Cell* 2005, **121:**725–737.

16. Preker P, Almvig K, Christensen MS, Valen E, Mapendano CK, Sandelin A, Jensen TH: **PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters.** *Nucleic Acids Res* 2011, **39:**7179–7193.
17. Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, Schierup MH, Jensen TH: **RNA exosome depletion reveals transcription upstream of active human promoters.** *Science* 2008, **322:**1851–1854.
18. Neil H, Malabat C, d'Aubenton-Carafa Y, Xu Z, Steinmetz LM, Jacquier A: **Widespread bidirectional promoters are the major source of cryptic transcripts in yeast.** *Nature* 2009, **457:**1038–1042.
19. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, *et al*: **Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.** *Nat Biotechnol* 2010, **28:**503–510.
20. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL: **Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.** *Genes Dev* 2011, **25:**1915–1927.
21. Sigova AA, Mullen AC, Molinie B, Gupta S, Orlando DA, Guenther MG, Almada AE, Lin C, Sharp PA, Giallourakis CC, Young RA: **Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells.** *Proc Natl Acad Sci U S A* 2013, **110**(8):2876–2881.
22. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A: **Comprehensive comparative analysis of strand-specific RNA sequencing methods.** *Nat Methods* 2010, **7:**709–715.
23. Thiagarajan RD, Cloonan N, Gardiner BB, Mercer TR, Kolle G, Nourbakhsh E, Wani S, Tang D, Krishnan K, Georgas KM, *et al*: **Refining transcriptional programs in kidney development by integration of deep RNA-sequencing and array-based spatial profiling.** *BMC Genomics* 2011, **12:**441.
24. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nat Protoc* 2012, **7:**562–578.
25. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28:**511–515.
26. Lin MF, Jungreis I, Kellis M: **PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions.** *Bioinformatics* 2011, **27:**i275–i282.
27. Bhatt DM, Pandya-Jones A, Tong AJ, Barozzi I, Lissner MM, Natoli G, Black DL, Smale ST: **Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions.** *Cell* 2012, **150:**279–290.
28. Rothenberg EV, Moore JE, Yui MA: **Launching the T-cell-lineage developmental programme.** *NatRevImmunol* 2008, **8:**9–21.
29. Chen J, Bardes EE, Aronow BJ, Jegga AG: **ToppGene Suite for gene list enrichment analysis and candidate gene prioritization.** *Nucleic Acids Res* 2009, **37:**W305–W311.
30. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B: **A map of the cis-regulatory sequences in the mouse genome.** *Nature* 2012, **488**(7409):116–120.
31. Hayday AC, Pennington DJ: **Key factors in the organized chaos of early T cell development.** *Nat Immunol* 2007, **8:**137–144.
32. Koch F, Fenouil R, Gut M, Cauchy P, Albert TK, Zacarias-Cabeza J, Spicuglia S, de la Chapelle AL, Heidemann M, Hintermair C, *et al*: **Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters.** *Nat Struct Mol Biol* 2011, **18:**956–963.
33. Zhang JA, Mortazavi A, Williams BA, Wold BJ, Rothenberg EV: **Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish T cell identity.** *Cell* 2012, **149:**467–482.
34. Akalin A, Fredman D, Arner E, Dong X, Bryne JC, Suzuki H, Daub CO, Hayashizaki Y, Lenhard B: **Transcriptional features of genomic regulatory blocks.** *Genome Biol* 2009, **10:**R38.
35. Lenhard B, Sandelin A, Carninci P: **Metazoan promoters: emerging characteristics and insights into transcriptional regulation.** *Nat Rev Genet* 2012, **13:**233–245.
36. Zhou VW, Goren A, Bernstein BE: **Charting histone modifications and the functional organization of mammalian genomes.** *Nat Rev Genet* 2011, **12:**7–18.
37. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, *et al*: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.** *Nature* 2007, **448:**553–560.
38. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, *et al*: **A bivalent chromatin structure marks key developmental genes in embryonic stem cells.** *Cell* 2006, **125:**315–326.
39. Wei G, Abraham BJ, Yagi R, Jothi R, Cui K, Sharma S, Narlikar L, Northrup DL, Tang Q, Paul WE, *et al*: **Genome-wide analyses of transcription factor GATA3-mediated gene regulation in distinct T cell types.** *Immunity* 2011, **35:**299–311.
40. Zhang J, Jackson AF, Naito T, Dose M, Seavitt J, Liu F, Heller EJ, Kashiwagi M, Yoshida T, Gounari F, *et al*: **Harnessing of the nucleosome-remodeling-deacetylase complex controls lymphocyte development and prevents leukemogenesis.** *Nat Immunol* 2012, **13:**86–94.
41. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, *et al*: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nat Genet* 2006, **38:**626–635.
42. Ginno PA, Lott PL, Christensen HC, Korf I, Chedin F: **R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters.** *Mol Cell* 2012, **45:**814–825.
43. Almada AE, Wu X, Kriz AJ, Burge CB, Sharp PA: **Promoter directionality is controlled by U1 snRNP and polyadenylation signals.** *Nature* 2013, **499:**360–363.
44. Ntini E, Jarvelin AI, Bornholdt J, Chen Y, Boyd M, Jorgensen M, Andersson R, Hoof I, Schein A, Andersen PR, *et al*: **Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality.** *Nat Struct Mol Biol* 2013, **20:**923–928.
45. Bieberstein NI, Carrillo Oesterreich F, Straube K, Neugebauer KM: **First exon length controls active chromatin signatures and transcription.** *Cell Rep* 2012, **2:**62–68.
46. Huff JT, Plocik AM, Guthrie C, Yamamoto KR: **Reciprocal intronic and exonic histone modification regions in humans.** *Nat Struct Mol Biol* 2010, **17:**1495–1499.
47. Heidemann M, Hintermair C, Voss K, Eick D: **Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription.** *Biochim Biophys Acta* 1829, **2013:**55–62.
48. Nam JW, Bartel D: **Long non-coding RNAs in C. elegans.** *Genome Res* 2012, **22**(12):2529–2540.
49. Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, Ferrier P, Spicuglia S, Gut M, Gut I, Andrau JC: **CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters.** *Genome Res* 2012, **22:**2399–2408.
50. Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS: **Specific expression of long noncoding RNAs in the mouse brain.** *Proc Natl Acad Sci U S A* 2008, **105:**716–721.
51. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, *et al*: **Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.** *Nature* 2009, **458:**223–227.
52. Wamstad JA, Alexander JM, Truty RM, Shrikumar A, Li F, Eilertson KE, Ding H, Wylie JN, Pico AR, Capra JA, *et al*: **Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage.** *Cell* 2012, **151:**206–220.
53. Ulitsky I, Bartel DP: **lincRNAs: genomics, evolution, and mechanisms.** *Cell* 2013, **154:**26–46.
54. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP: **Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution.** *Cell* 2011, **147:**1537–1550.
55. Tanay A, O'Donnell AH, Damelin M, Bestor TH: **Hyperconserved CpG domains underlie Polycomb-binding sites.** *Proc Natl Acad Sci U S A* 2007, **104:**5521–5526.
56. Yang MQ, Elnitski LL: **Diversity of core promoter elements comprising human bidirectional promoters.** *BMC Genomics* 2008, **9**(Suppl 2):S3.
57. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM: **An abundance of bidirectional promoters in the human genome.** *Genome Res* 2004, **14:**62–66.
58. Li X, Manley JL: **Cotranscriptional processes and their influence on genome stability.** *Genes Dev* 2006, **20:**1838–1847.
59. Dominguez-Sanchez MS, Barroso S, Gomez-Gonzalez B, Luna R, Aguilera A: **Genome instability and transcription elongation impairment in human cells depleted of THO/TREX.** *PLoS Genet* 2011, **7:**e1002386.
60. El Hage A, French SL, Beyer AL, Tollervey D: **Loss of Topoisomerase I leads to R-loop-mediated transcriptional blocks during ribosomal RNA synthesis.** *Genes Dev* 2010, **24:**1546–1558.

61. Tous C, Aguilera A: **Impairment of transcription elongation by R-loops in vitro.** *Biochem Biophys Res Commun* 2007, **360:**428–432.
62. Lee JY, Khan AA, Min H, Wang X, Kim MH: **Identification and characterization of a noncoding RNA at the mouse Pcna locus.** *Mol Cells* 2012, **33:**111–116.
63. Grote P, Wittler L, Hendrix D, Koch F, Wahrisch S, Beisaw A, Macura K, Blass G, Kellis M, Werber M, Herrmann BG: **The Tissue-Specific lncRNA Fendrr Is an Essential Regulator of Heart and Body Wall Development in the Mouse.** *Dev Cell* 2013, **24:**206–214.
64. Adelman K, Lis JT: **Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans.** *Nat Rev Genet* 2012, **13:**720–731.
65. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA: **A chromatin landmark and transcription initiation at most promoters in human cells.** *Cell* 2007, **130:**77–88.
66. Adelman K, Kennedy MA, Nechaev S, Gilchrist DA, Muse GW, Chinenov Y, Rogatsky I: **Immediate mediators of the inflammatory response are poised for gene activation through RNA polymerase II stalling.** *Proc Natl Acad Sci U S A* 2009, **106:**18207–18212.
67. Hargreaves DC, Horng T, Medzhitov R: **Control of inducible gene expression by signal-dependent transcriptional elongation.** *Cell* 2009, **138:**129–145.
68. Koch F, Jourquin F, Ferrier P, Andrau JC: **Genome-wide RNA polymerase II: not genes only!** *Trends Biochem Sci* 2008, **33:**265–273.
69. Xu Z, Wei W, Gagneur J, Clauder-Munster S, Smolik M, Huber W, Steinmetz LM: **Antisense expression increases gene expression variability and locus interdependency.** *Mol Syst Biol* 2011, **7:**468.
70. Lopez-Maury L, Marguerat S, Bahler J: **Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation.** *Nat Rev Genet* 2008, **9:**583–593.
71. Arias AM, Hayward P: **Filtering transcriptional noise during development: concepts and mechanisms.** *Nat Rev Genet* 2006, **7:**34–44.
72. Shinkai Y, Rathbun G, Lam KP, Oltz EM, Stewart V, Mendelsohn M, Charron J, Datta M, Young K, Stall AL, Alt FW: **RAG-2 deficient mice lack mature lymphocytes owing to inability to initiate V(D)J rearrangement.** *Cell* 1992, **68:**855–867.
73. Pekowska A, Benoukraf T, Zacarias-Cabeza J, Belhocine M, Koch F, Holota H, Imbert J, Andrau JC, Ferrier P, Spicuglia S: **H3K4 tri-methylation provides an epigenetic signature of active enhancers.** *EMBO J* 2011, **30:**4198–4210.
74. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841–842.
75. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, *et al*: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci U S A* 2004, **101:**6062–6067.
76. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK: **Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.** *Mol Cell* 2010, **38:**576–589.
77. da Huang W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc* 2009, **4:**44–57.
78. Lepoivre C, Bergon A, Lopez F, Perumal NB, Nguyen C, Imbert J, Puthier D: **TranscriptomeBrowser 3.0: introducing a new compendium of molecular interactions and a new visualization tool for the study of gene regulatory networks.** *BMC Bioinf* 2012, **13:**19.

doi:10.1186/1471-2164-14-914
**Cite this article as:** Lepoivre *et al.*: Divergent transcription is associated with promoters of transcriptional regulators. *BMC Genomics* 2013 **14**:914.

# Article 4 :

## Identification et caractérisation des LncRNAs impliqués dans le développement des lymphocytes T et dans la leucémie lymphoblastique aiguë des cellules T.

### Contexte

Aifantis et coll. ont suggéré, en 2008, que la dérégulation des processus épigénétiques durant la différenciation des cellules T pourrait être à l'origine de la transformation oncogénique des précurseurs T (Aifantis et al., 2008b). Les LAL-T sont, en fait, un groupe hétérogène de cellules lymphoblastiques qui se sont arrêtées à différents stades de leur développement (Asnafi et al., 2004). Le progrès des dernières décennies a permis l'identification de mutations dites de « type A » qui bloquent la différenciation des cellules T à des stages spécifiques de leur développement et définissent des sous-groupes ayant des profils spécifiques d'expression génétique (Homminga et al., 2011, Ferrando et al., 2002).

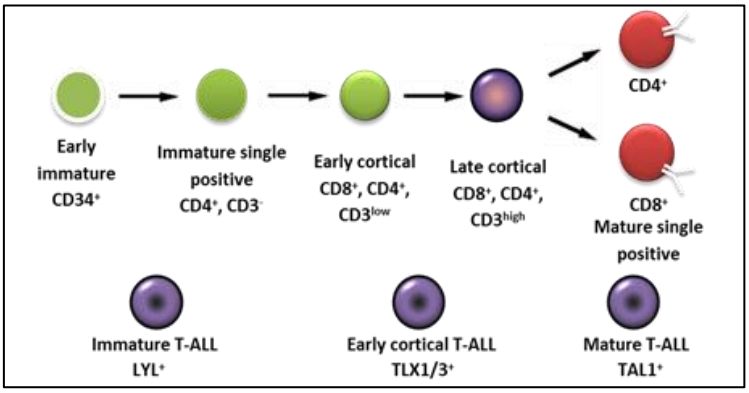



**Fig. 37 : schéma de la différenciation des lymphocytes humaines et les différents sous-types de LAL-T correspondant au stade d'arrêt de maturation.**

Étant donné leurs contributions aux processus physiologiques (*i.e.* la modulation de l'épigénome), une altération du niveau d'expression des LncRNAs pourrait mener au développement de pathologies, comme le cancer. Parmi les divers exemples de l'implication des LncRNAs dans le cancer, le rôle de HOTAIR dans les néoplasies humaines est le plus étudié (Esteller, 2011). Dans les cellules épithéliales cancéreuses, la surexpression de HOTAIR cause la redistribution des complexes *Polycomb* dans le génome. En conséquence, les complexes *Polycomb* ne répriment plus leurs gènes cibles usuels, comme les gènes du locus Hox, et mènent donc à l'augmentation de l'invasion cellulaire et à la propension à métastaser.

### Objectifs

Étant donné que les LncRNAs peuvent moduler l'épigénome par leurs interactions avec des régulateurs de la chromatine, comme les *Polycomb*, et que ces régulateurs de chromatine sont souvent mutés ou dérégulés dans les leucémies (Ntziachristos et al.,

2012), il est très probable que les LncRNAs puissent jouer un rôle important dans le développement de la leucémie. Dans cette étude nous avons eu comme objectif la caractérisation des LncRNAs qui sont importants pour la différenciation des cellules T et potentiellement impliqués dans les leucémies aiguës des lymphocytes T (LAL-T). Nous avons identifié et défini les LncRNAs exprimés à des stades clés de la différenciation des cellules T chez l'homme et sur un certain nombre de blastes LAL-T choisis et représentant les différentes classes de leucémies. Ceci a été accompli grâce à des données RNA-seq produites dans le cadre du projet *Blueprint*. L'originalité de ce projet est qu'il fournit une vue complète des LncRNAs régulés durant une différenciation cellulaire et sa contrepartie pathologique, le cancer. Ceci a le potentiel de révéler de nouvelles espèces de LncRNAs qui jouent un rôle dans la différenciation des lymphocytes T et/ou de la leucémie, ainsi que de découvrir de nouveaux mécanismes impliqués dans la différenciation cellulaire et/ou la leucémogenèse. En conséquence, durant ma thèse j'ai réalisé les différentes analyses suivantes :

1- Identification des nouveaux transcrits LncRNAs dans les lymphocytes T et LAL-T ;
2- Caractérisation des nouveaux transcrits LncRNAs ;
3- Etablissement d'un catalogue exhaustif de LncRNA connus et nouveaux ;
4- Quantification de l'expression des LncRNAs dans les lymphocytes T et LAL-T ;
5- Analyse différentielle entre condition normale et condition leucémique ;
6- Analyse de la corrélation entre l'expression des LncRNAs et ARNm ;
7- Identification des LncRNAs associés aux délétions récurrentes dans les LAL-T ;
8- Priorisation des LncRNAs associés aux domaines larges de la H3K4me3.

## Les données de séquençage

Les données que nous avons utilisées proviennent de trois sources différentes:

1- Des données de RNA-seq publiées par d'autres laboratoires. Au total 62 échantillons RNA-seq des LAL-T primaires et des lignées cellulaires ont été collectés ;
2- Des données générées dans le cadre du consortium européen Blueprint. Ce consortium, dont font partie mes deux laboratoires de tutelle, m'ont permis d'avoir un accès privilégié à des données RNA-seq et épigénomiques des 4 sous-populations thymiques humaines et de 9 LAL-T.
3- Des données RNA-seq pour 62 échantillons LAL-T collectés par le laboratoire de Vahid Asnafi (Institut Necker) et séquencés dans notre plateforme TGML de séquençage à haut débit au TAGC. Ces patients ont suivi le protocole européen GRAALL (http://www.graall-leukemia.fr) et nous possédons toutes les données clinico-biologiques nécessaires.

## Le protocole d'analyse bioinformatique

Les données obtenues après séquençage sont analysées selon le protocole de la Fig. 38. Toutes les analyses sont faites en ligne de commande sous Linux et le pipeline a été organisé grâce à *makefile* code.

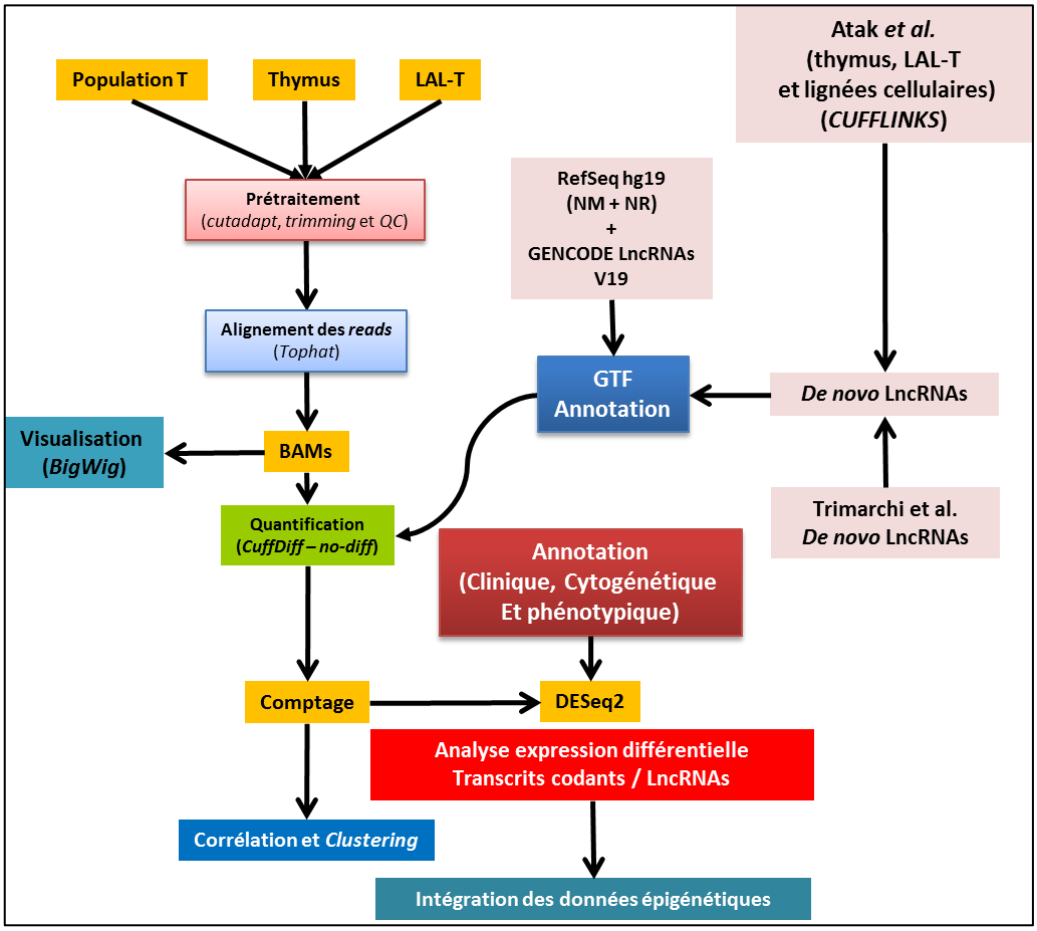



**Fig. 38 : Pipeline générale des analyses bioinformatiques**

### 1- Catalogue des LncRNAs et l'annotation de référence

Dans notre projet et afin de construire un catalogue robuste et représentatif, nous avons collecté le maximum de données de transcriptome humain. En effet, nous avons choisi d'utiliser comme référence une combinaison entre l'annotation *Refseq* (version du génome humain hg19) et la version 19 de l'annotation GENCODE (Harrow et al., 2012), produite par le projet ENCODE. Cette annotation GENCODE comporte trois types d'annotation: une annotation validée, une annotation manuelle et une automatique. Elle est basée sur les annotations *Ensembl* (automatique) et *HAVANA* (manuelle). Dans cette version 19 nous avons 57820 gènes annotés, dont 20345 gènes codant pour des protéines, 13870 codant des LncRNAs, 9013 codant des petits ARN et 14206 pseudogènes. Ensuite, nous avons complété cette annotation par deux ensembles de *De*

*Novo* transcrits générés par Trimarchi et Coll. (Trimarchi et al., 2014) et par nous-mêmes grâce à la découverte des nouveaux transcrits non codants par l'exploitation des données RNA-seq publiques (Atak et al., 2013) (voir prochain paragraphe).

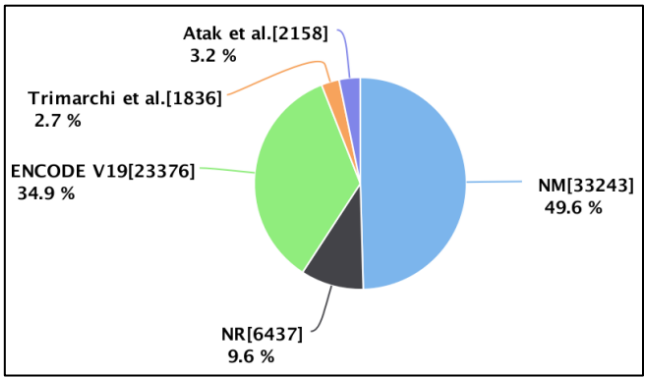



**Fig. 39 : Nombre des transcrits codants et LncRNAs de chaque catégorie d'annotation**

## 2- Identification *De Novo* de transcrits LncRNAs dans les lymphocytes et les leucémies

Nous avons utilisé pour cette recherche *de novo* des données RNA-seq publiques comportant 18 lignées cellulaires, 31 leucémies primaires et un thymus normal (Atak et al., 2013). Travail réalisé en collaboration avec Marc-Antoine Garibal (Cancéropôle PACA)

Nous sommes donc partis d'un ensemble de 50 fichiers d'alignement (BAM) sur lesquels nous avons réalisé une étape d'assemblage de transcrits avec l'outil ***cufflinks*** et sans fournir d'annotation de référence en conservant les paramètres par défaut, sauf pour les options « –j » et « –a » fixés respectivement à « 0.6 » et « 0.00001 ». Nous avons ensuite utilisé ***cuffmerge*** afin de réunir tous les fichiers (un par échantillon) de transcrits séparés en un seul en supprimant ceux considérés comme doublons ou composante d'un plus grand. Nous sommes passés ensuite par une étape de nettoyage par ***cuffcompare*** qui consistait à supprimer tous les transcrits ayant au moins une base en correspondance avec l'annotation de référence (RefSeq, GENCODE v19 et *de novo* Trimarchi) afin de sélectionner uniquement sur les transcrits encore non annotés. Ces transcrits non annotés ont ensuite été quantifiés en utilisant ***cufflink***s avec l'option '-G' dans chaque fichier BAM de chacun des échantillons à disposition afin de ne conserver que les transcrits ayant un FPKM supérieur à 1 dans au moins l'un des échantillons.

Trois étapes de nettoyage additionnel ont été réalisées : (1) conserver seulement les transcrits avec plus de 200 nucléotides de taille (2) conserver seulement des transcrits multi-exoniques (ayant au moins deux exons) et (3) conserver des transcrits sans un potentiel codant. Cette dernière étape a été réalisée grâce à l'outil **CAPT** où nous avons

considéré tout transcrit avec un score moins de 0.20 comme non codant. Nous avons obtenu au final 2158 transcrits LncRNAs *de novo*.

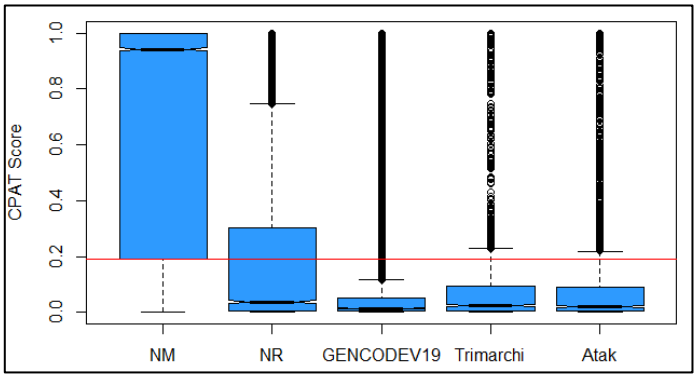



**Fig. 40: distribution des potentiels codants de différentes catégories d'ARN**

Dans le cadre du projet Tassili, une collaboration internationale entre l'Algérie et la France, je me suis occupé de l'encadrement d'une jeune doctorante en bioinformatique Mlle Yasmine Kermezli. Avec Yasmine nous avons réalisé une caractérisation de ces nouveaux LncRNAs :

### a) <u>Distribution de la taille des transcrits :</u>

La distribution de taille des *de novo* LncRNAs est comparée aux autres ARN annotés (NM, NR, GENCODEv19 and *De Novo* Trimarchi). Les ARN *de novo* d'Atak ont des tailles comparables à celles des autres LncRNAs annotés (Fig. 40). La distribution montre aussi que les ARN *de novo* d'Atak ont des tailles comprises entre 0.2 et 100 Kb (une moyenne de 5kb). Les tailles des ARN *de novo* restent dans la norme des LncRNAs annotés et inférieures aux tailles des ARNm.

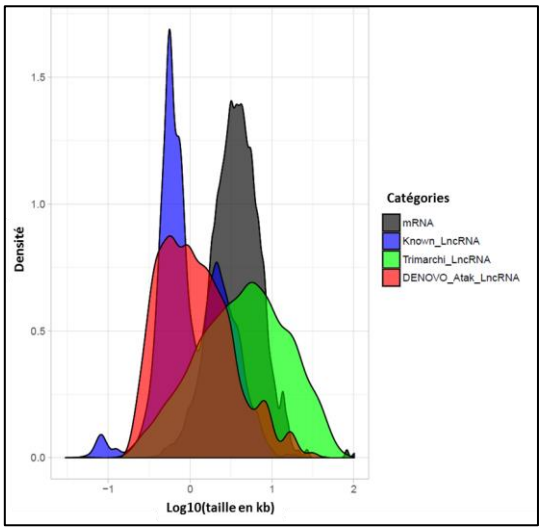



**Fig. 41 : Distribution et comparaison des tailles des LncRNAs *De novo* d'Atak versus les autres ARN annotés (ARNm, GENCODE V19 et LncRNAs *De novo* de Trimarchi)**

### b) Classification des LncRNAs :

L'une des façons de classifier les LncRNAs est de les subdiviser en catégories selon leurs positionnements relatifs avec les gènes codant pour les protéines. En conséquence, nous avons utilisé *GTFTOOLKIT*, un outil très pratique destiné à faciliter la manipulation des fichiers gtf, développé par Dr Puthier Denis. Ainsi, grâce à la fonction *annotate*, nous avons observé que la majorité (~ 60%) des LncRNAs *De novo* identifiés avec les données d'Atak et al. sont des LncRNAs intergénique (LincRNAs) ; et environ 30% sont des antisens exoniques/introniques ; alors que les divergents et les convergents ne dépassent pas les 8% ensemble.

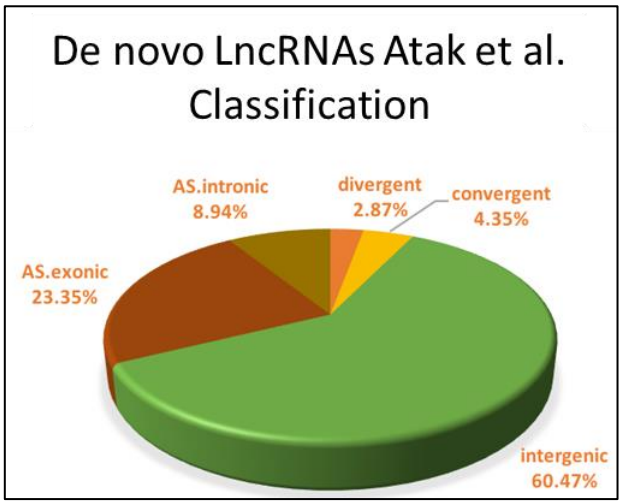



**Fig. 42 : Classification des LncRNAs *De novo* d'Atak en se basant sur l'intersection avec des gènes codant pour une protéine**

### c) Niveau de l'expression dans les LAL-T et les lignées cellulaires

L'expression des différents transcrits est normalisée par FPKM dans toutes les données d'expression du papier Atak et al. PloS One 2013, un niveau d'expression moyen est calculé pour chaque candidat: pour chaque transcrit, on calcule la moyenne de l'expression dans toutes les LAL-T ou les lignées cellulaires. Donc, pour chaque transcrit, on a une seule expression normalisée par FPKM. La comparaison des niveaux d'expression des différentes familles d'ARN : messagers, LncRNAs annotés, LncRNAs *de novo* Trimarchi et Atak a montré que ces derniers sont faiblement exprimés comparés aux messagers mais comparables aux LncRNAs annotés (Fig. 43) que ce soit chez les LAL-T ou les lignées cellulaires. Ce résultat s'explique par la caractéristique des LncRNAs d'avoir un faible niveau d'expression en comparaison avec les ARNm (Cabili et al., 2011; Derrien et al., 2012; Djebali et al., 2012).

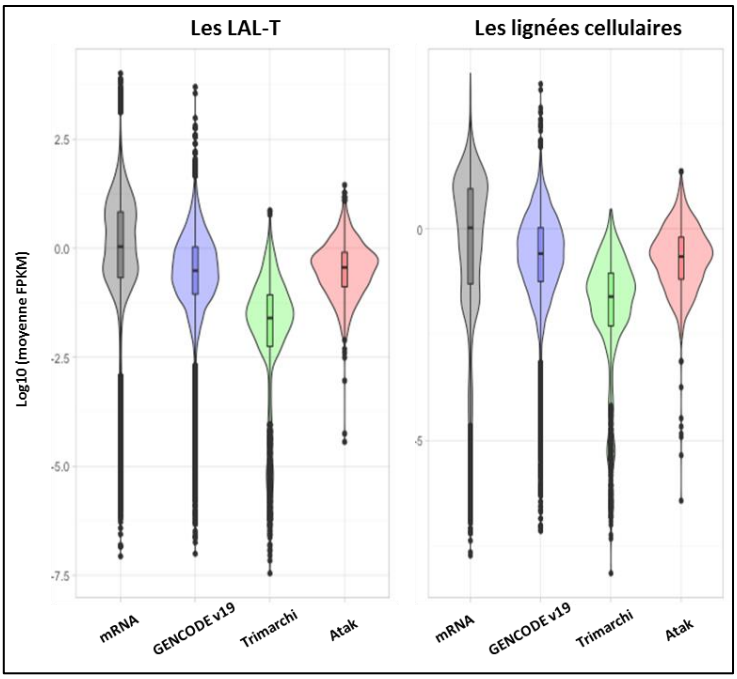



**Fig. 43 : Comparaison des niveaux d'expression de différentes catégories d'ARN (ARNm, LncRNAs annotés de GENCODE, les *de novo* de Trimarchi et d'Atak) les valeurs de l'expression moyenne chez les LAL-T et les lignées cellulaires sont transformées en log2.**

### d) Distribution du nombre des exons

Nous avons utilisé la fonction *count* de l'outil *GTFTOOLKIT* pour calculer le nombre d'exons dans les différents transcrits de chaque catégorie d'ARN. Nous avons observé que la majorité des transcrits LncRNAs annotés et LncRNAs *de novo* d'Atak et Trimarchi contiennent deux exons ; alors que plus de 70% des ARNm ont plus de 4 exons.

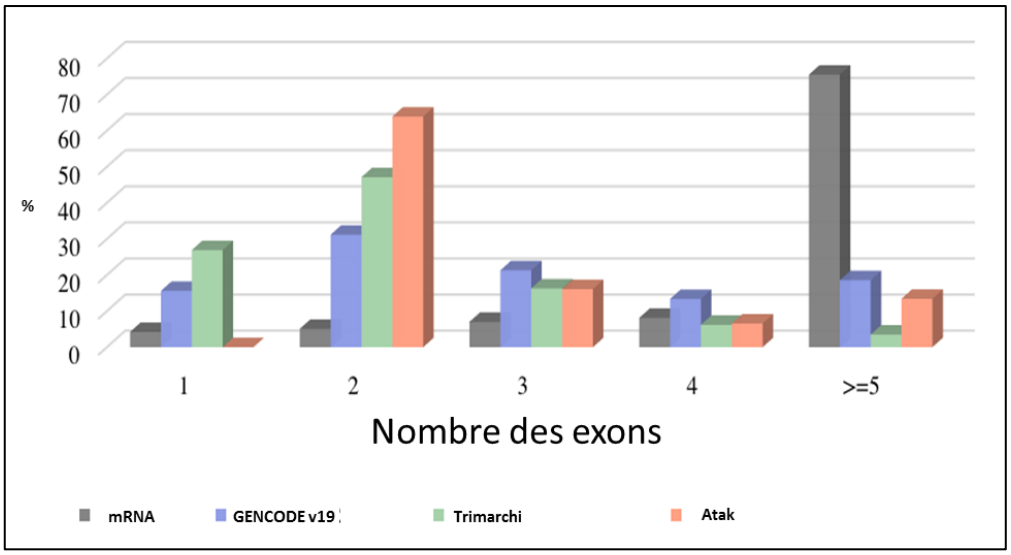



**Fig. 44 : Histogramme de pourcentage du nombre d'exons dans les différentes catégories d'ARNm.**

### 3- Contrôle de qualité

Tout d'abord, les données sont contrôlées pour leur qualité avant utilisation. Ainsi, les fichiers *fastq* sont analysés avec le logiciel ***FastQC.*** Ensuite, les adaptateurs ont été supprimés des *reads* à l'aide de l'outil ***CutAdapt*** (http://code.google.com/p/cutadapt/). Après, les *reads* de très mauvaise qualité, inférieure à 20 sur la totalité du *read*, sont éliminés grâce à l'outil ***sickle*** pour les données RNA-seq provenant d'illumina et grâce à un script *Perl*, fait maison, pour les données provenant de SOLiD. Aussi, les *reads* avec une extrémité 3' ou 5' de mauvaise qualité (qualité inférieure à 20) sont écourté. Enfin, la qualité des données prétraitées est contrôlée à nouveau avec ***FastQC.***

### 4- Alignement et quantification des *reads*

Les *reads* nettoyés ont été alignés sur la version du génome humain hg19 en utilisant l'outil ***Tophat2*** (Trapnell et al., 2009). Les *reads* alignés sur des positions chromosomiques uniques ont été retenus. Le nombre total de *reads* alignés pour toutes les expériences de cette étude est donné dans la Fig.45. Les *reads* alignés ont servi ensuite à générer des profils de densité en format *BigWig* séparément pour le brin positif et le brin négatif. Le nombre total de *reads* qui chevauchent seulement les régions exoniques des transcrits codant ou non codant ont été obtenus en utilisant la fonction ***cuffdiff*** de l'outil ***cufflinks*** (Trapnell et al., 2010). Enfin, Les comptages sont ensuite normalisés par RPKM (*Read Per Kilobase per Million*) pour une meilleure comparaison entre les différentes conditions.

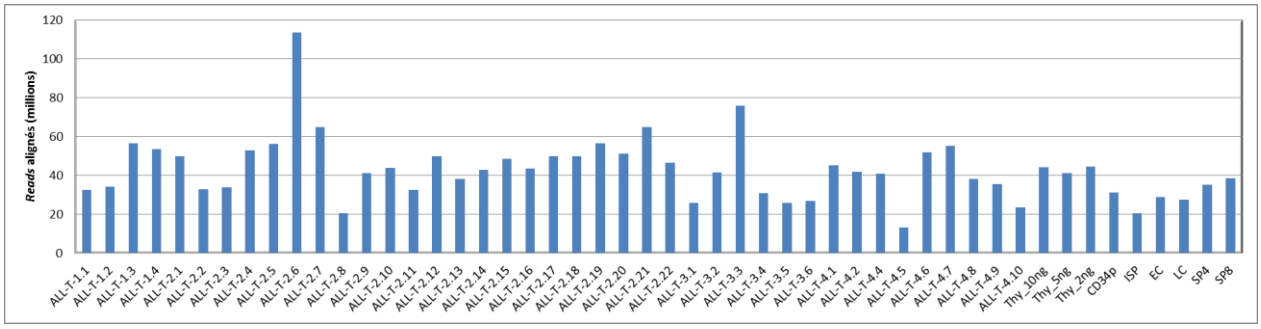



**Fig. 45 : Le nombre des *reads* alignées par *Tophat* pour chaque échantillon de la série Necker.**

### 5- Identification des profils d'expression dynamique

De nombreux LncRNAs montrent des profils d'expression temporaux et spatiaux spécifiques, qui peuvent nous orienter vers le contexte biologique dans lequel ils agissent. Ainsi, tous les transcrits ont été clustérisés en 20 modèles d'expression, en utilisant le regroupement par ***k-means*** avec la distance euclidienne et 100 itérations maximum comme paramètres (avec le logiciel R) et cela au cours des différentes transitions du développement des thymocytes. Les valeurs d'expression normalisées des transcrits dans les différents modèles d'expression ont été représentés sous forme de

boites à moustaches (Fig. 46). Le cluster N°20 n'est pas représenté, c'est le cluster contenant l'ensemble des gènes non exprimés nulle part durant la progression.

En conséquence, nous avons pu identifier des LncRNAs exprimés ou réprimés spécifiquement à un stade développemental bien précis. Par exemple, les clusters : C2, C4, C6, C7, C9 et C14 représentent les gènes exprimés spécifiquement pour chaque stade de développement T et donc les LncRNAs exprimés spécifiquement dans ces clusters peuvent être des LncRNAs fonctionnelles et interviennent dans la régulation de leurs stades d'expression respectif. Un autre exemple est le cluster N°13 qui semble contenir, entre autres, les gènes de la *beta sélection* puisque les gènes dans ce groupe sont régulés à la baisse progressivement. En plus, ce cluster contient au moins 90 facteurs de transcription important pour la régulation de la maturation tels que les gènes HOXA, donc les LncRNAs qui présentent un profil d'expression semblable peuvent, probablement, intervenir dans la *beta sélection* aussi.

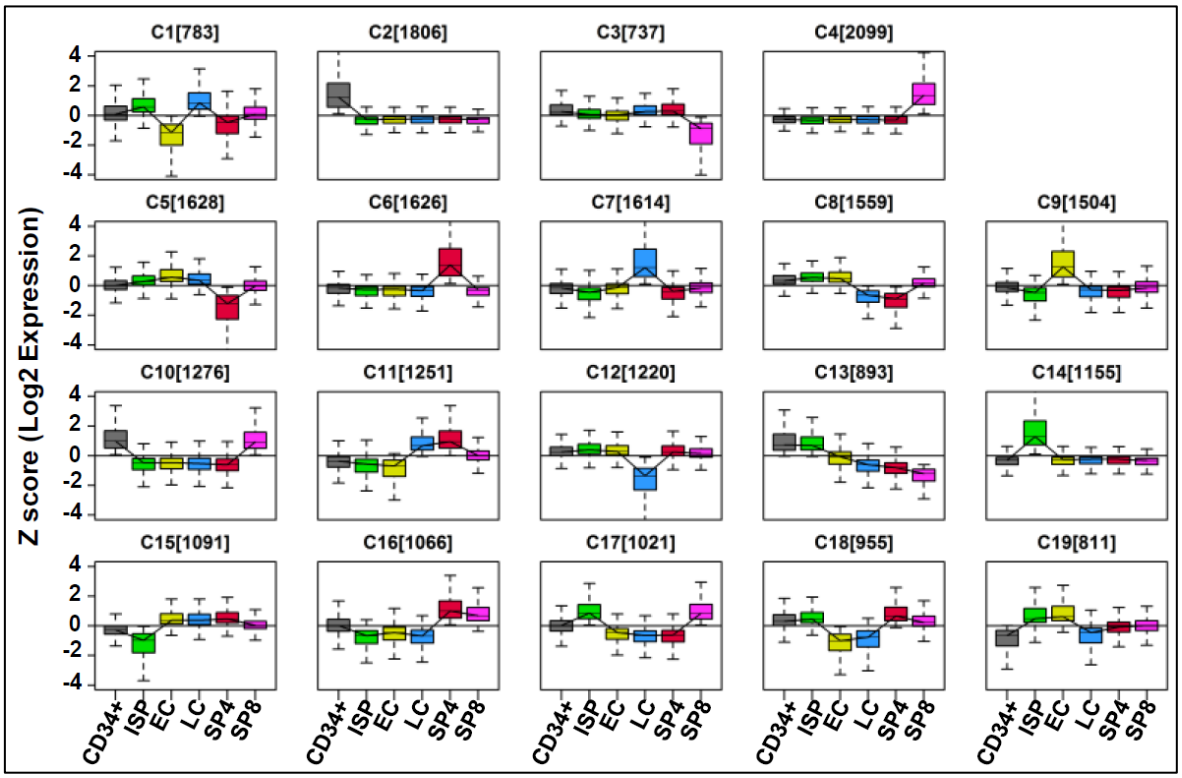



**Fig. 46 : Les différents profils d'expression des transcrits codants et LncRNAs durant la différenciation des lymphocytes T**

## 6- <u>Analyse différentielle entre condition normale et condition leucémique</u>

Etant donné que l'expression des LncRNAs est hautement spécifique, il y a un intérêt croissant dans l'utilisation de ces molécules comme des biomarqueurs ou des cibles thérapeutiques pour les leucémies. Les associations entre une maladie et des LncRNAs

peuvent être identifiées par les analyses de l'expression différentielle entre sujet malade et sain. En conséquence, nous avons réalisé une analyse de l'expression différentielle des transcrits codants pour des protéines et des LncRNAs avec l'outil ***DESeq2*** (Anders and Huber, 2010) en utilisant comme donnée d'entrée le fichier de comptage des transcrits généré auparavant par l'outil ***cuffdiff***. Les transcrits sont concédés comme différentiellement exprimés (DE) s'ils présentent une valeur absolue du *Log2Fold Change* plus de 1 entre le niveau d'expression dans les échantillons normaux et leucémiques et une *p-value* moins ou égale à 0.01. En effet, si le *Log2Fold Change* est positif donc le transcrit est surexprimé dans les échantillons leucémique alors que si le *Log2Fold Change* est négatif donc le transcrit est réprimé dans les échantillons leucémiques en comparaison aux échantillons normaux.

Cette analyse différentielle a été réalisée selon deux stratégies : (i) en comparant toutes les sous classes leucémiques (14 LAL-T immatures, 18 LAL-T corticales et 9 LAL-T matures) *versus* toutes les sous populations T normales afin d'identifier les LncRNAs leucémiques T spécifiques en général ; (ii) en comparant les classes leucémiques entre elles et puis soustraire la signature développementale normale de la contrepartie correspondante à son stade d'arrêt de maturation afin d'identifier les LncRNAs sous-type leucémiques spécifique (Fig.47).

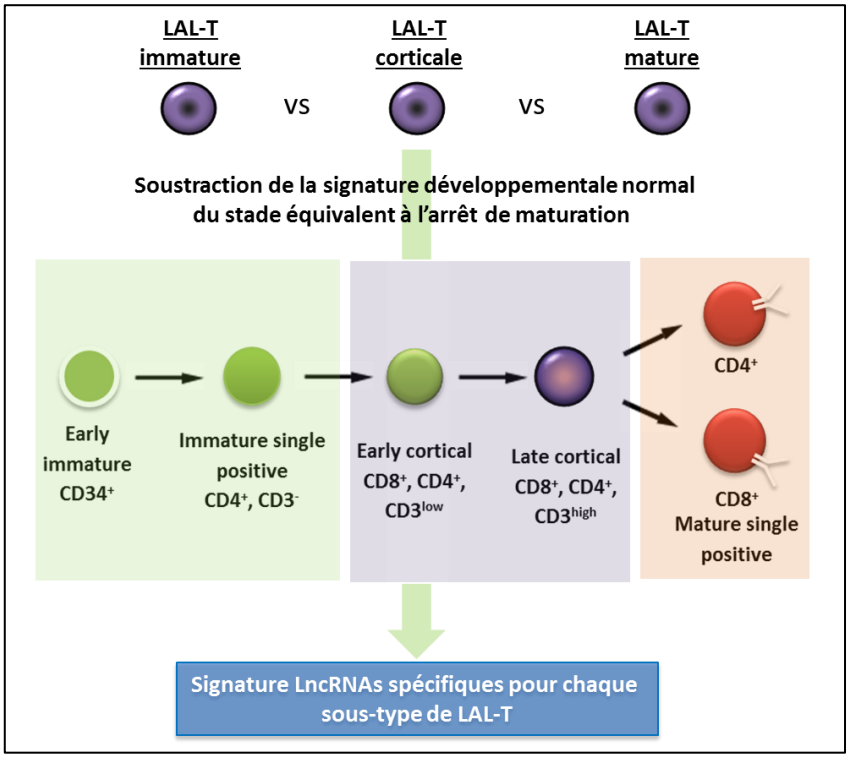



**Fig. 47: schéma de la stratégie d'identification des signatures sous-types LAL-T spécifiques.**

Dans un premier temps et grâce à la première stratégie nous avons obtenu, au final, une signature leucémique avec gènes codant et LncRNAs. Le résultat est résumé dans le tableau 7 suivant :

**Tableau 7 : Résultat de l'analyse différentielle entre normal T et leucémies.**

| | Transcrits codant | Gènes codant | Transcrits LncRNAs | Gènes LncRNAs |
|---|---|---|---|---|
| **Surexprimés dans LAL-T** | 3281 | 2842 | 2716 | 2201 |
| **Sous exprimés dans LAL-T** | 1445 | 1337 | 836 | 539 |
| **Total DE** | 4726 | 4179 | 3552 | 2740 |

Dans un deuxième temps, nous avons obtenu trois signatures sous-type spécifiques avec les différentes comparaisons réalisées grâce à la deuxième stratégie. Les résultats sont présentés dans le tableau 8 suivant :

**Tableau 8 : Résultat de l'analyse différentielle des LncRNAs entre les différents sous-types leucémies.**

| Signatures | | Immature | Corticale | Mature |
|---|---|---|---|---|
| **Avant soustraction** | **surexprimé** | 110 | 82 | 130 |
| | **réprimé** | 150 | 102 | 121 |
| **Après soustraction** | **surexprimé** | 55 | 31 | 59 |
| | **réprimé** | 73 | 47 | 51 |

En se focalisant sur les LncRNAs surexprimés des signatures et afin de sélectionner seulement les LncRNAs uniques de chaque sous-type, nous avons réalisé une intersection entre ces trois signatures. Ainsi, nous avons obtenu 12 LncRNAs spécifiques aux LAL-T immatures, 31 LncRNAs spécifiques aux LAL-T corticales et 16 LncRNAs spécifiques aux LAL-T matures (Fig. 48). Ces LncRNAs peuvent servir comme des biomarqueurs spécifiques aux différents sous-types des LAL-T.

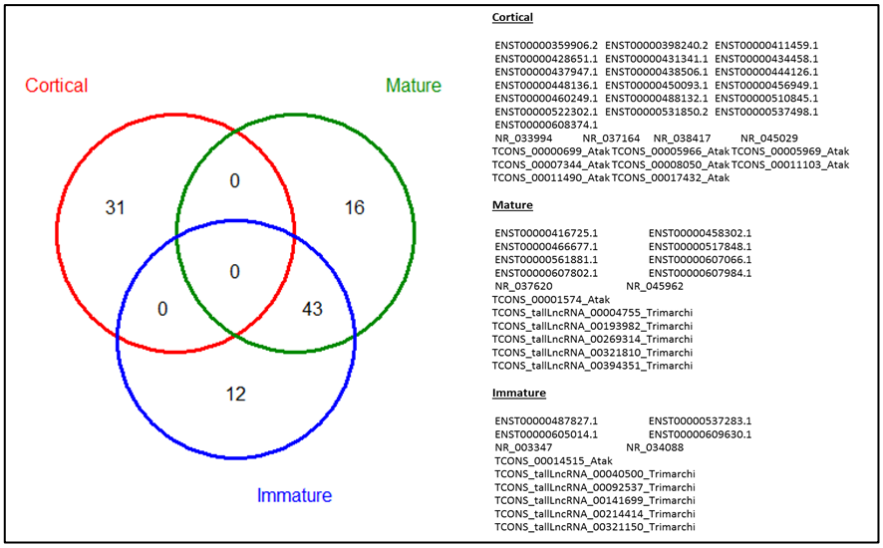



**Fig. 48: Les signatures des LncRNAs surexprimés sous-types spécifiques**

## 7- Classification des échantillons grâce aux transcrits différentiellement exprimés

Grâce à l'outil *Gene-E* (http://www.broadinstitute.org/cancer/software/GENE-E), Les FPKM transformés en Log2 (n+1) des transcrits DE de la première et deuxième (Fig. 50) stratégie ont été hiérarchiquement clustérisés en utilisant la corrélation *Pearson* et *average linkage* comme paramètres. Ensuite, avec le même outil, la matrice a été représentée sous forme d'un *Heatmap.* Ainsi, nous observons que, d'une part, les signatures leucémiques que ce soit des gènes codants ou LncRNAs séparent très bien les LAL-T des cellules normales (Fig. 49) ; et d'autre part, la signature des LncRNAs surexprimés sous-type spécifique sépare les différentes sous-types des LAL-T.

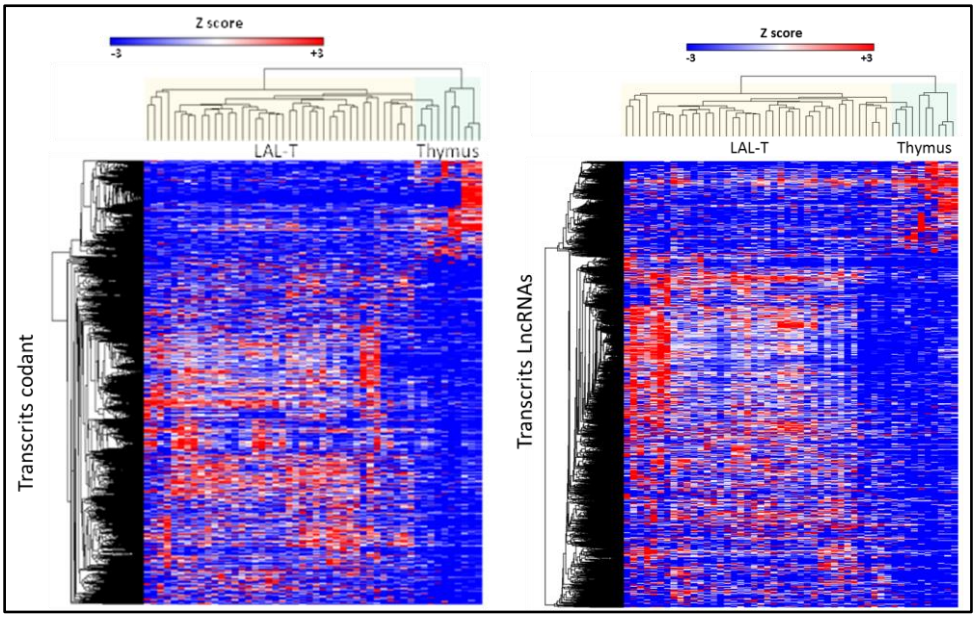



**Fig. 49 : *Clustering* hiérarchiques des signatures des transcrits codants (gauche) et LncRNAs (droite) différentiellement exprimés entre les LAL-T et les sous populations T normales.**

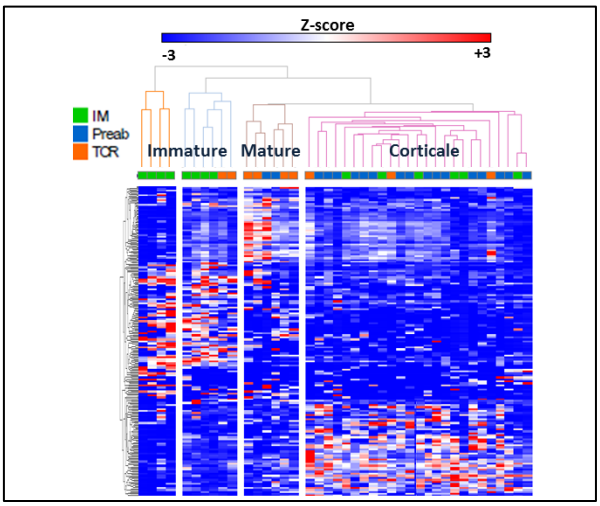



**Fig. 50 : *Clustering* hiérarchiques des signatures des LncRNAs surexprimés uniques de chaque sous-type de LAL-T**

## 8- Corrélation entre l'expression des LncRNAs exprimés et les ARNm codants

Nous avons analysé la corrélation entre l'expression des LncRNAs et les ARNm dans tous les échantillons. Le coefficient de corrélation de Pearson est calculé pour chaque LncRNA avec tous les ARNm qui ont au moins 1 RPKM d'expression dans un des échantillons. L'analyse a été faite sur 3552 LncRNAs et 9726 ARNm exprimés dans au moins un des échantillons. Nous avons gardé seulement les candidats ayant une corrélation supérieure à 0.7 ou inférieure à -0.7. Après ce filtre nous avons réalisé un *clustering* hiérarchique et identifié des 2 clusters de LncRNAs anti-corrélés entre eux. Le 1er cluster, avec 834 LncRNAs, corrèle avec deux clusters d'ARNm enrichis en voie de développement et de prolifération, alors que l'autre cluster, avec 2719 LncRNAs, corrèle avec 4 clusters enrichis en voie de régulation des miRNA et de la réponse immunitaire et de développement (Fig. 52). Ces résultats pourraient suggérer un rôle de certains LncRNAs dans la régulation ou la dérégulation de ces processus et probablement une implication dans la leucémogenèse.

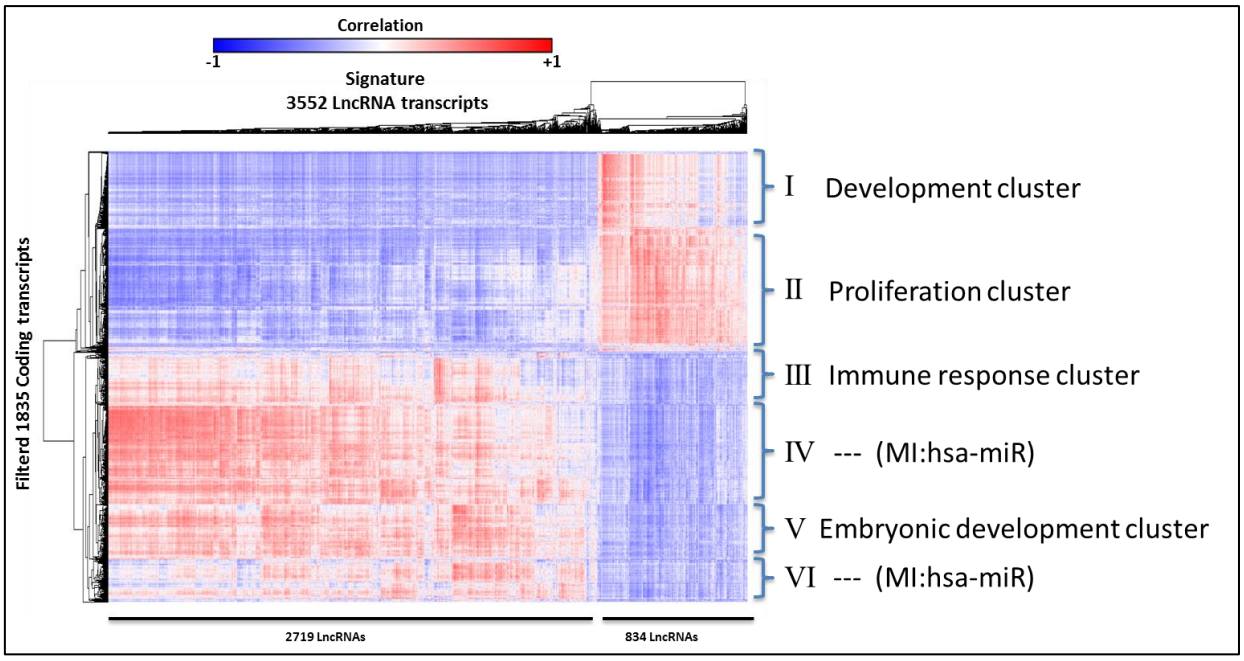



**Fig. 52 : *Clustering* de la matrice de corrélation de la signature LncRNAs et les ARNm avec l'enrichissement fonctionnel de chaque cluster.**

## 9- Identification des LncRNAs associés aux délétions récurrentes dans les LAL-T

Dans les cellules tumorales, les profils d'expression des pro-oncogènes ou des suppresseurs de tumeurs sont généralement dérégulés par des mutations somatiques ou des altérations de séquence telles que les insertions ou délétions. L'identification de telles perturbations dans le génome des patients est un objectif majeur de la génétique du cancer.

Ainsi, nous suggérons que la présence d'un LncRNA dans une région délétée récurrente dans les LAL-T peut être une indication d'un rôle probable dans la leucémogenèse. En effet, on peut imaginer qu'une délétion récurrente et proche d'un LncRNA surexprimé dans LAL-T suggère que cette région délétée participe à la création de la frontière (*Insulator*) entre le promoteur de ce LncRNA et une région *cis* régulatrice distale (*enhancer*) et donc la suppression de cette barrière permettrait l'activation du LncRNA qui joue un rôle d'un pro-oncogène. Un autre exemple peut être la délétion récurrente partielle ou complète d'un locus LncRNA et ainsi la perte de la fonction de ce transcrit dans les LAL-T, ce qui suggère probablement une fonction suppresseur de tumeur pour ce LncRNA délété.

Afin d'identifier les LncRNA pertinents dans la régulation et le développement de la leucémie, j'ai croisé la signature leucémique des LncRNAs avec la liste des délétions récurrentes dans les LAL-T décrite par (Hnisz et al., 2016). En conséquence, j'ai récupéré seulement les LncRNAs réprimés de la signature et associés à une délétion récurrente dans les LAL-T (Tableau 9).

**Tableau 9: Liste des LncRNAs réprimés (signature LAL-T versus tous les sous populations lymphocytaires) et délétés de façon récurrente dans les LAL-T**

| Délétion | ID Gène | ID Transcrit | Chr | Log2FC | *P*adj |
|---|---|---|---|---|---|
| chr6:6258985-6463233 | ENSG00000216863.5 | ENST00000435641.1 | chr6 | -1.60 | 0.000515198 |
| chr4:109035396-109084396 | ENSG00000232021.2 | ENST00000436413.1 | chr4 | -2.25 | 0.003730256 |
| chr1:1994731-2177270 | ENSG00000182873.4 | ENST00000444529.1 | chr1 | -1.41 | 0.019431217 |
| chr10:97889010-98067010 | ENSG00000177853.10 | ENST00000478086.1 | chr10 | -2.61 | 0.000209826 |
| chr4:152985058-153272297 | ENSG00000245954.2 | ENST00000499452.2 | chr4 | -3.38 | 2.63E-06 |
| chr12:92283532-92536532 | ENSG00000245904.2 | ENST00000499685.2 | chr12 | -1.83 | 0.000200637 |
| chr17:38629070-38974823 | ENSG00000266088.1 | ENST00000577557.1 | chr17 | -1.72 | 0.012101735 |
| chr21:35304300-35575358 | ENSG00000237945.3 | ENST00000593977.1 | chr21 | -1.46 | 0.00404629 |
| chr2:43425349-43712349 | LOC100129726 | NR_027251 | chr2 | -1.38 | 0.002073334 |
| chr10:1462813-1611788 | XLOC_00001842_Atak | TCONS_00001842_Atak | chr10 | -3.08 | 4.53E-06 |
| chr14:106969763-107285437 | XLOC_00005968_Atak | TCONS_00005968_Atak | chr14 | -1.62 | 0.010797128 |

## 10- Priorisation des LncRNAs pour la validation expérimentale

Comme discuté auparavant, dans le chapitre II – Article 2, les domaines larges de la méthylation de la H3K4 sont associés à des gènes tissu-spécifiques de l'identité cellulaire ou des suppresseurs de tumeur et sont souvent associés à des gènes dérégulés chez les leucémies. En conséquence, nous avons proposé l'utilisation de cette caractéristique épigénétique comme outil de priorisation des gènes codant ou LncRNAs fonctionnellement pertinents dans les leucémies. Ainsi, j'ai combiné la signature leucémique obtenue grâce aux données d'expression et les données épigénétiques au cours du développement des lymphocytes T et dans les LAL-T.

Dans un premier temps et pour des raisons techniques, nous nous sommes intéressés aux LncRNAs surexprimés dans les LAL-T. Ces analyses nous ont permis d'identifier 15 candidats LncRNAs qui ont acquis des larges domaines durant la transformation leucémique. Ces candidats font l'objet d'une validation expérimentale. Par exemple, dans le cas du *de novo* LncRNA « TCONS_00009269_Atak » (Fig. 53), nous avons une absence totale de pic d'enrichissement de la marque H3K4me3 dans les populations thymiques alors que nous observons l'apparition des domaines larges dans les leucémies et une lignée cellulaire leucémique (Jurkat).

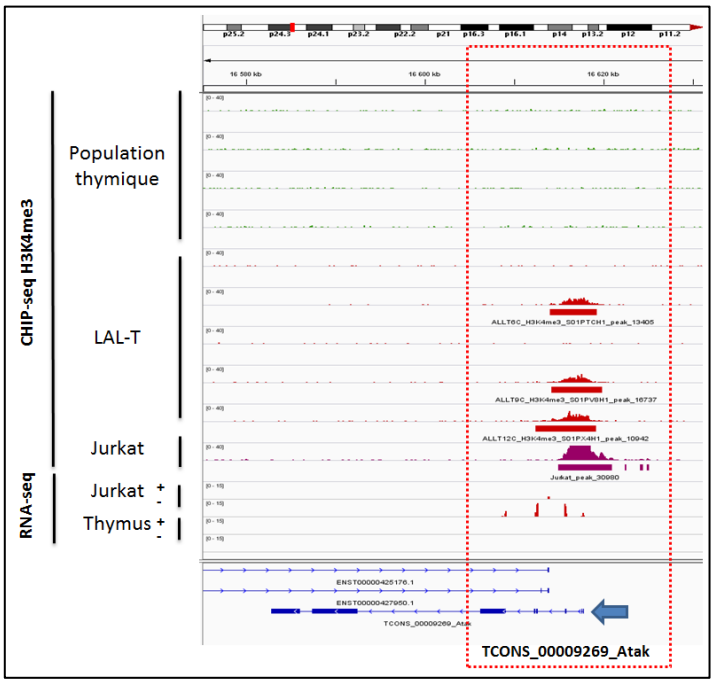



**Fig. 53: Exemple du LncRNA priorisé par stratégie de domaine large H3K4me3 : Le profil H3K4me3 du LncRNA *de novo* TCONS_00009269_Atak dans les populaires thymiques normales, Les LAL-T et la lignée cellulaire Jurkat. Le profil d'expression (RNA-seq) sur le brin plus et moins chez le thymus et la lignée cellulaire Jurkat.**

## Conclusion

Ce projet a permis d'avoir une vue d'ensemble de l'expression des LncRNAs exprimés au cours du développement des lymphocytes T et/ou la transformation oncogénique. Tout d'abord, nous avons pu identifier des profils d'expression dynamique d'un ensemble de LncRNAs durant la maturation des lymphocytes T. Ces profils vont permettre, ainsi, d'émettre des hypothèses fonctionnelles pour ces LncRNAs. Ensuite, nous avons sélectionné une signature leucémique de 2740 gènes LncRNAs (3552 transcrits) dérégulée entre les lymphocytes T et les leucémies. Ces LncRNA ont fait

ensuite l'objet d'une multitude de stratégies d'analyse bioinformatique afin de sélectionner certains candidats prometteurs. En effet, nous avons utilisé la corrélation générale de l'expression entre les paires LncRNAs/ARNm et l'association aux délétions récurrentes dans les LAL-T. Enfin, nous avons utilisé une nouvelle stratégie de priorisation grâce aux données épigénétiques et, ainsi, réduit le nombre des LncRNAs candidats pour les validations expérimentales. Notre projet a permis de mettre en évidence de nouveaux LncRNAs, qui constitueraient alors de biomarqueurs potentiels de diagnostic ou de pronostic et probablement des cibles thérapeutiques. Ces résultats seront très importants pour la suite de nos projets, mais aussi une ressource précieuse pour d'autres laboratoires.

# CHAPITRE IV : Discussion générale et conclusion

Les travaux accomplis lors de cette thèse ont permis une meilleure compréhension des mécanismes impliqués dans la régulation des gènes en analysant les données transcriptomiques et épigénomiques au cours du développement des lymphocytes T et dans les leucémies aigues lymphoblastiques T.

Dans la première partie de ma thèse, je me suis intéressé à un nouvel aspect de la régulation épigénétique qu'est l'information qualitative liée aux profils des modifications d'histones et à leur lien direct avec la régulation de la transcription. Ainsi, les domaines étendus de la marque H3K4me3 marquent les gènes de l'identité cellulaire et la tissu-spécificité (Pekowska et al., 2010, Benayoun et al., 2014). Plus intéressant encore, ces domaines semblent marquer les gènes suppresseurs de tumeurs (Chen et al., 2015a). Par ailleurs, la présence de ces domaines étendus semble conférer une stabilité accrue dans la qualité de la transcription, conduisant une moindre variabilité de l'expression entre les cellules ainsi qu'une stabilité accrue en réponse à des signaux externes (Pekowska et al., 2010, Chen et al., 2015a).

Un des exemples extrêmes de ces gènes associés aux domaines étendus est le locus TCRβ (article 1). En effet, le locus TCRβ présente un marquage très élevé et large de l'H3K4me3 qui peut ressembler à une plateforme transcriptionnelle. En effet, cette plateforme semble permettre l'établissement d'une structure chromatinienne spécifique afin d'assurer une régulation très fine du locus TCRβ. Nous avons pu démontrer l'importance du recrutement de la Pol II au niveau du locus TCRβ et son lien direct avec la phase d'élongation de la transcription.

Ensuite, nous avons analysé la dynamique de ces domaines au cours de la différenciation des lymphocytes T et dans les LAL-T (article 2). D'un côté, durant la différenciation des lymphocytes T et nous avons constaté que cette dynamique est étroitement liée au niveau d'expression, en particulier, les gènes principaux de la régulation et développement T. D'un autre côté, nous avons analysé ces domaines dans les leucémies et nous avons observé un nombre plus élevé des domaines chez les LAL-T en comparaison aux lymphocytes T. Un fait très intéressant est que les principaux oncogènes sont associés à un profil étendu de la H3K4me3 dans la leucémie où ils sont exprimés. En conséquence, nous avons suggéré que la perte des domaines dans LAL-T marque la répression des gènes de l'identité cellulaire T et les suppresseurs de tumeurs, alors que le gain des domaines dans LAL-T marque plutôt l'activation des oncogènes. Ainsi, afin de valider ces observations il faut envisager : (1) de tester cette stratégie dans d'autres modèles de cancer. Par contre, il faut prévoir d'avoir les données ChIP-seq H3K4me3 et RNA-seq sur les mêmes échantillons ; (2) de moduler le niveau de

H3K4me3 dans les lignées leucémiques et évaluer l'impact sur les profils étendus et sur l'expression des gènes qui y sont associés. Le fait que ces domaines soient associés aussi aux LncRNAs, nous a permis de les utiliser, dans la seconde partie de ma thèse, comme outil de priorisation des LncRNAs dans les LAL-T.

Quel est le mécanisme moléculaire impliqué dans l'établissement des domaines étendus de H3K4me3. Les travaux précédents (Chen et al., 2015a, Benayoun et al., 2014, Pekowska et al., 2010, Zacarias-Cabeza et al., 2015) ont suggéré que ce marquage chromatinien particulier est intimement lié au processus transcriptionnel lui-même. En effet, les gènes associés aux domaines étendus H3K4me3 possèdent un niveau de RNA Pol II accrue, indépendamment de leur niveau d'expression. Pour mieux comprendre les mécanismes sous-jacent, nous avons entrepris une analyse compréhensive des données de ChIP-seq des différentes formes de phosphorylation de l'ARN Pol II, de l'accessibilité de la chromatine (DNAseI et FAIRE) et des données de RNA-seq des différentes populations d'ARN (Totaux, PolyA, Petits ARNs). Bien que très préliminaires, nos analyses suggèrent qu'au niveau des gènes associés aux domaines étendus H3K4me3, les phases d'initiation et d'élongation son directement couplés. En particulier, nous avons observé un niveau relativement plus élevé des ARN totaux (transcrits immatures et matures) par rapport au niveau d'ARN polyadenylés (matures uniquement) (Fig. 55), ceci suggère que les gènes associés aux domaines H3K4me3 sont sujet à une transcription immature plus importante.

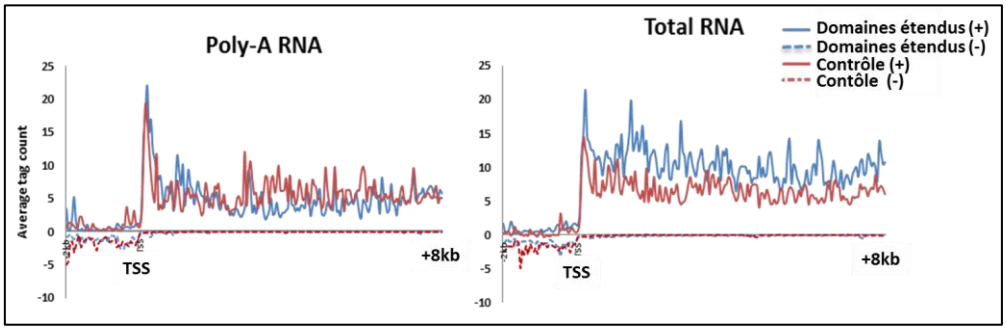



**Fig. 55 : Profils moyens (métagène) des ARN totaux (droite) et ARN polyadenylés (gauche) dans les gènes associés aux domaines étendus de H3K4me3 (bleu) et les gènes contrôles avec des profils canoniques de H3K4me3 (rouge). Les données utilisées proviennent des lymphocytes DP, les profils ont été réalisés avec deux groupes de gènes (associés aux domaines étendus et contrôle) exprimés au même niveau.**

Ce phénomène pourrait permettre une meilleure qualité de production des ARNm et d'expliquer les observations précédentes concernant la stabilité transcriptionnelle de ces gènes. Afin d'expliquer ces résultats, nous avons émis l'hypothèse que ces gènes présentent une vitesse de transcription plus rapide. Afin de vérifier cette hypothèse, nous avons établi une collaboration avec l'équipe du Dr JC Andreau (IGMM, Montpelier) afin de calculer la vitesse de transcription pour chaque catégorie des gènes à partir de données cinétiques de ChIP-seq et RNA-Seq.

Pendant la deuxième partie de ma thèse, je me suis intéressé à un autre nouvel acteur de la régulation, les LncRNAs. Bien que la pertinence biologique de nombreux LncRNAs ait été établie sans ambiguïté, ce nouveau niveau de complexité a conduit à la notion que la transcription n'est pas limitée à des locus fonctionnels bien définis, tels que les gènes. Mais peut être le reflet des régions régulatrices fonctionnelles comme les petits-ARN dans les promoteurs ou les *eRNA* dans les enhancers.

Cependant, une large proportion des LncRNAs est transcrite dans le sens inverse des gènes codants. Ces transcrits anti-sens peuvent être soit juste la conséquence d'une transcription bidirectionnelle non fonctionnelle ou soit des éléments importants avec une fonction régulatrice. En effet, les ARN anti-sens pourraient potentiellement exercer une fonction régulatrice sur l'ARNm transcrit dans l'autre sens. Des résultats récents, ont montré qu'ils peuvent agir comme régulateurs épigénétiques de l'expression tels que les *NATs* (Magistri et al., 2012). Par exemple, le gène *Kcnq1* est réprimé par son LncRNA anti-sens *Kcnq1ot1*, cela par l'interaction avec le complexe *Polycomb* grâce à une liaison directe avec *PCR2*. *Kcnq1ot1* induit ainsi la déposition des marques de répressions H3K27me3 et H3K9me3 (Pandey et al., 2008). D'autres études ont mis en évidence l'existence d'un mécanisme par lequel une transcription non codante divergente peut établir un environnement permissif de la chromatine. Par exemple, *Hand2* qui est un régulateur ancestral du développement cardiaque et l'un des quatre facteurs de transcription qui contrôle la reprogrammation des fibroblastes dans les cardiomyocytes (Srivastava et al., 1997), est strictement régulé par un *super-enhancer* enrichi en modifications d'histone H3K27ac (Charite et al., 2001). L'ARN non codant long associé à *Hand2* (nommé *Upperhand* : *Uph*) est nécessaire pour le maintien de la signature active du *super-enhancer.* Ainsi, le blocage de la transcription du transcrit *Uph* abolie complètement l'expression du gène *Hand2* et provoque une hypoplasie ventriculaire droite et la létalité embryonnaire chez les souris (Anderson et al., 2016). Un autre exemple très intéressant, celui de l'activation de l'expression d'un gène par une action en *cis* de son transcrit divergent. En effet, le gène EVX1, un gène très important pour la différenciation des cellules souches en mésendoderme, est régulé par son transcrit divergent EVX1as. La répression du transcrit EVX1as par *knock-down* induit une diminution drastique de l'expression du gène EVX1 et perturbe, ainsi, indirectement la différenciation (Luo et al., 2016).

Ainsi, durant ma thèse nous avons identifié de façon systématique les transcrits divergents (LUAT) et observé qu'ils sont associés d'une manière significative à des gènes développementaux (article 3). Les promoteurs de ces gènes développementaux présentent des caractéristiques génétiques et épigénétiques uniques. Ces résultats suggèrent fortement un lien fonctionnel entre la transcription divergente et la régulation des gènes codant pour des facteurs de transcription et de développement et ainsi pourraient illustrer un mécanisme supplémentaire pour affiner l'expression des gènes. Pour valider ces observations, nous prévoyions de moduler l'expression de ces LUAT

afin d'analyser leurs effets et puis étudier leur dérégulation dans des contextes pathologiques telle que la leucémie.

Dans la dernière partie de ma thèse, je me suis intéressé à l'identification et la caractérisation des LncRNAs impliqués dans le développement des lymphocytes T et dans la leucémie lymphoblastique aiguë des cellules T (article 4). Tout d'abord, nous avons identifié des nouveaux LncRNAs non annotés. Ces nouveaux LncRNAs ont des caractéristiques comparables avec ceux des LncRNAs annotés. Ces LncRNAs nouveaux ont été additionnés à toutes les autres annotations disponibles afin de construire un catalogue robuste et représentatif. Puis, nous avons analysé les données transcriptomiques et identifié des modules d'expression dynamique des LncRNAs durant la différenciation des sous populations T. Ensuite, nous avons sélectionné une signature leucémique de 3552 transcrits LncRNAs dérégulée entre les lymphocytes T et les leucémies et une autre signature de 59 transcrits LncRNAs qui distingue entre les différents sous-types d'arrêt de maturation. Enfin, nous avons testé deux stratégies d'analyse bioinformatique afin de sélectionner certains candidats prometteurs. La première stratégie utilisée est l'association aux altérations génomiques récurrentes. Ainsi, nous avons obtenu 11 LncRNAs réprimés et co-localisés avec des délétions récurrentes dont 2 sont des nouveaux LncRNAs non annotés. Cette analyse a été réalisée avec des données récupérées des bases de données publiques et ne provient pas directement de nos échantillons. Par contre, nous disposons actuellement des données *CGH array* qui sont en cours d'analyse afin de combiner les données des altérations génomiques (délétion et insertion) et les signatures transcriptomiques. La deuxième stratégie utilisée est la méthode de priorisation avec les domaines étendus de la H3K4me3 décrite au début de ma thèse. Cette méthode a permis l'identification de 15 LncRNAs surexprimés et associés à un domaine étendu dans les LAL-T dont 3 sont des nouveaux LncRNAs non annotés.

Deux autres stratégies sont envisagées à court terme. La première consiste à l'utilisation des données cliniques, telle que la survie ou le pronostic, afin d'associer les niveaux d'expression des LncRNAs avec des phénotypes d'intérêt. Cette stratégie nécessite l'utilisation d'un grand nombre d'échantillons afin que les analyses statistiques soient pertinentes. En effet, nous suggérons l'utilisation de la technique TLDA (O'Brien, 2017)(*TaqMan low density array* ; voir Annexe I) qui peut analyser un nombre réduit de candidats sur un très grand nombre d'échantillon. La deuxième consiste à chercher les LncRNAs très fortement exprimés dans un certain nombre d'échantillons seulement. Cette stratégie se base sur le modèle des oncogènes *masters* tel que les oncogène TLX1/3 ou TAL1.

Enfin, les LncRNAs sélectionnés vont servir à la caractérisation fonctionnelle qui incluera (1) leur *knock down* et/ou leur surexpression dans des thymocytes primaires, (2) ou dans des modèle leucémiques et (3) l'exploration des fonctions moléculaires et les mécanismes d'action.

# Bibliographie


ADOLFSSON, J., BORGE, O. J., BRYDER, D., THEILGAARD-MONCH, K., ASTRAND-GRUNDSTROM, I., SITNICKA, E., SASAKI, Y. & JACOBSEN, S. E. 2001. Upregulation of Flt3 expression within the bone marrow Lin(-)Sca1(+)c-kit(+) stem cell compartment is accompanied by loss of self-renewal capacity. *Immunity,* 15**,** 659-69.

AIFANTIS, I., RAETZ, E. & BUONAMICI, S. 2008a. Molecular pathogenesis of T-cell leukaemia and lymphoma. *Nat Rev Immunol,* 8**,** 380-90.

AIFANTIS, I., RAETZ, E. & BUONAMICI, S. 2008b. Molecular pathogenesis of T-cell leukaemia and lymphoma. *Nat.Rev.Immunol.,* 8**,** 380-390.

ALLIS, C. D., BERGER, S. L., COTE, J., DENT, S., JENUWIEN, T., KOUZARIDES, T., PILLUS, L., REINBERG, D., SHI, Y., SHIEKHATTAR, R., SHILATIFARD, A., WORKMAN, J. & ZHANG, Y. 2007. New nomenclature for chromatin-modifying enzymes. *Cell,* 131**,** 633-6.

ALVAREZ-DOMINGUEZ, J. R., HU, W., GROMATZKY, A. A. & LODISH, H. F. 2014. Long noncoding RNAs during normal and malignant hematopoiesis. *Int J Hematol,* 99**,** 531-41.

AMARAL, P. P., CLARK, M. B., GASCOIGNE, D. K., DINGER, M. E. & MATTICK, J. S. 2011. lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res,* 39**,** D146-51.

ANDERS, S. & HUBER, W. 2010. Differential expression analysis for sequence count data. *Genome Biol,* 11**,** R106.

ANDERSON, K. M., ANDERSON, D. M., MCANALLY, J. R., SHELTON, J. M., BASSEL-DUBY, R. & OLSON, E. N. 2016. Transcription of the non-coding RNA upperhand controls Hand2 expression and heart development. *Nature,* 539**,** 433-436.

ANDERSSON, R., GEBHARD, C., MIGUEL-ESCALADA, I., HOOF, I., BORNHOLDT, J., BOYD, M., CHEN, Y., ZHAO, X., SCHMIDL, C., SUZUKI, T., NTINI, E., ARNER, E., VALEN, E., LI, K., SCHWARZFISCHER, L., GLATZ, D., RAITHEL, J., LILJE, B., RAPIN, N., BAGGER, F. O., JORGENSEN, M., ANDERSEN, P. R., BERTIN, N., RACKHAM, O., BURROUGHS, A. M., BAILLIE, J. K., ISHIZU, Y., SHIMIZU, Y., FURUHATA, E., MAEDA, S., NEGISHI, Y., MUNGALL, C. J., MEEHAN, T. F., LASSMANN, T., ITOH, M., KAWAJI, H., KONDO, N., KAWAI, J., LENNARTSSON, A., DAUB, C. O., HEUTINK, P., HUME, D. A., JENSEN, T. H., SUZUKI, H., HAYASHIZAKI, Y., MULLER, F., FORREST, A. R., CARNINCI, P., REHLI, M. & SANDELIN, A. 2014. An atlas of active enhancers across human cell types and tissues. *Nature,* 507**,** 455-61.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics,* 25**,** 25-29.

ASNAFI, V., BELDJORD, K., BOULANGER, E., COMBA, B., LE TUTOUR, P., ESTIENNE, M. H., DAVI, F., LANDMAN-PARKER, J., QUARTIER, P., BUZYN, A., DELABESSE, E., VALENSI, F. & MACINTYRE, E. 2003a. Analysis of TCR, pT alpha, and RAG-1 in T-acute lymphoblastic leukemias improves understanding of early human T-lymphoid lineage commitment. *Blood,* 101**,** 2693-703.

ASNAFI, V., BELDJORD, K., LIBURA, M., VILLARESE, P., MILLIEN, C., BALLERINI, P., KUHLEIN, E., LAFAGE-POCHITALOFF, M., DELABESSE, E., BERNARD, O. & MACINTYRE, E. 2004. Age-related phenotypic and oncogenic differences in T-cell acute lymphoblastic leukemias may reflect thymic atrophy. *Blood,* 104**,** 4173-80.

ASNAFI, V., RADFORD-WEISS, I., DASTUGUE, N., BAYLE, C., LEBOEUF, D., CHARRIN, C., GARAND, R., LAFAGE-POCHITALOFF, M., DELABESSE, E., BUZYN, A., TROUSSARD, X. & MACINTYRE, E. 2003b. CALM-AF10 is a common fusion transcript in T-ALL and is specific to the TCRgammadelta lineage. *Blood,* 102**,** 1000-6.

ATAK, Z. K., GIANFELICI, V., HULSELMANS, G., DE KEERSMAECKER, K., DEVASIA, A. G., GEERDENS, E., MENTENS, N., CHIARETTI, S., DURINCK, K., UYTTEBROECK, A., VANDENBERGHE, P., WLODARSKA, I., CLOOS, J., FOA, R., SPELEMAN, F., COOLS, J. & AERTS, S. 2013. Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia. *PLoS Genet,* 9**,** e1003997.

AVERY, O. T., MACLEOD, C. M. & MCCARTY, M. 1944. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med,* 79**,** 137-58.

BAILEY, T. L., JOHNSON, J., GRANT, C. E. & NOBLE, W. S. 2015. The MEME Suite. *Nucleic Acids Res,* 43**,** W39-49.

BANERJI, J., OLSON, L. & SCHAFFNER, W. 1983. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell,* 33**,** 729-740.

BANNISTER, A. J. & KOUZARIDES, T. 2011. Regulation of chromatin by histone modifications. *Cell Res,* 21**,** 381-95.

BARSKI, A., CUDDAPAH, S., CUI, K., ROH, T. Y., SCHONES, D. E., WANG, Z., WEI, G., CHEPELEV, I. & ZHAO, K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell,* 129**,** 823-37.

BASSING, C. H., SWAT, W. & ALT, F. W. 2002. The mechanism and regulation of chromosomal V(D)J recombination. *Cell,* 109**,** S45-S55.

BAUMANN, M., MAMAIS, A., MCBLANE, F., XIAO, H. & BOYES, J. 2003. Regulation of V(D)J recombination by nucleosome positioning at recombination signal sequences. *EMBO J,* 22**,** 5197-5207.

BEDNAR, J., HOROWITZ, R. A., GRIGORYEV, S. A., CARRUTHERS, L. M., HANSEN, J. C., KOSTER, A. J. & WOODCOCK, C. L. 1998. Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. *Proc Natl Acad Sci U S A,* 95**,** 14173-8.

BEN ABDELALI, R., ASNAFI, V., PETIT, A., MICOL, J. B., CALLENS, C., VILLARESE, P., DELABESSE, E., REMAN, O., LEPRETRE, S., CAHN, J. Y., GUILLERM, G., BERTHON, C., GARDIN, C., CORRONT, B., LEGUAY, T., BENE, M. C., IFRAH, N., LEVERGER, G., DOMBRET, H. & MACINTYRE, E. 2013. The prognosis of CALM-AF10-positive adult T-cell acute lymphoblastic leukemias depends on the stage of maturation arrest. *Haematologica,* 98**,** 1711-7.

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. & BRUNET, A. 2014. H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell,* 158**,** 673-88.

BENE, M. C., CASTOLDI, G., KNAPP, W., LUDWIG, W. D., MATUTES, E., ORFAO, A. & VAN'T VEER, M. B. 1995. Proposals for the immunological classification of acute leukemias. European Group for the Immunological Characterization of Leukemias (EGIL). *Leukemia,* 9**,** 1783-6.

BERNSTEIN, B. E., BIRNEY, E., DUNHAM, I., GREEN, E. D., GUNTER, C. & SNYDER, M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

BERNSTEIN, B. E., STAMATOYANNOPOULOS, J. A., COSTELLO, J. F., REN, B., MILOSAVLJEVIC, A., MEISSNER, A., KELLIS, M., MARRA, M. A., BEAUDET, A. L., ECKER, J. R., FARNHAM, P. J., HIRST, M., LANDER, E. S., MIKKELSEN, T. S. & THOMSON, J. A. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol,* 28**,** 1045-8.

BERTANI, S., SAUER, S., BOLOTIN, E. & SAUER, F. 2011. The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin. *Mol Cell,* 43**,** 1040-6.

BIRD, A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev.,* 16**,** 6-21.

BIRD, A. P. 1986. CpG-rich islands and the function of DNA methylation. *Nature,* 321**,** 209-213.

BIRNEY, E., STAMATOYANNOPOULOS, J. A., DUTTA, A., GUIGO, R., GINGERAS, T. R., MARGULIES, E. H., WENG, Z., SNYDER, M., DERMITZAKIS, E. T., THURMAN, R. E., KUEHN, M. S., TAYLOR, C. M., NEPH, S., KOCH, C. M., ASTHANA, S., MALHOTRA, A., ADZHUBEI, I., GREENBAUM, J. A., ANDREWS, R. M., FLICEK, P., BOYLE, P. J., CAO, H., CARTER, N. P., CLELLAND, G. K., DAVIS, S., DAY, N., DHAMI, P., DILLON, S. C., DORSCHNER, M. O., FIEGLER, H., GIRESI, P. G., GOLDY, J., HAWRYLYCZ, M., HAYDOCK, A., HUMBERT, R., JAMES, K. D., JOHNSON, B. E., JOHNSON, E. M., FRUM, T. T., ROSENZWEIG, E. R., KARNANI, N., LEE, K., LEFEBVRE, G. C., NAVAS, P. A., NERI, F., PARKER, S. C., SABO, P. J., SANDSTROM, R., SHAFER, A., VETRIE, D., WEAVER, M., WILCOX, S., YU, M., COLLINS, F. S., DEKKER, J., LIEB, J. D., TULLIUS, T. D., CRAWFORD, G. E., SUNYAEV, S., NOBLE, W. S., DUNHAM, I., DENOEUD, F., REYMOND, A., KAPRANOV, P., ROZOWSKY, J., ZHENG, D., CASTELO, R., FRANKISH, A., HARROW, J., GHOSH, S., SANDELIN, A., HOFACKER, I. L., BAERTSCH, R., KEEFE, D., DIKE, S., CHENG, J., HIRSCH, H. A., SEKINGER, E. A., LAGARDE, J., ABRIL, J. F., SHAHAB, A., FLAMM, C., FRIED, C., HACKERMULLER, J., HERTEL, J., LINDEMEYER, M., MISSAL, K., TANZER, A., WASHIETL, S., KORBEL, J., EMANUELSSON, O., PEDERSEN, J. S., HOLROYD, N., TAYLOR, R., SWARBRECK, D., MATTHEWS, N., DICKSON, M. C., THOMAS, D. J., WEIRAUCH, M. T., GILBERT, J., et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature,* 447**,** 799-816.

BLOM, B., RES, P. C. & SPITS, H. 1998. T cell precursors in man and mice. *Crit Rev Immunol,* 18**,** 371-88.

BOLAND, M. J., NAZOR, K. L. & LORING, J. F. 2014. Epigenetic regulation of pluripotency and differentiation. *Circ Res,* 115**,** 311-24.

BONNET, M., FERRIER, P. & SPICUGLIA, S. 2009. Molecular genetics at the T-cell receptor beta locus: insights into the regulation of V(D)J recombination. *Adv Exp Med Biol,* 650**,** 116-32.

BROCKDORFF, N. 2002. X-chromosome inactivation: closing in on proteins that bind Xist RNA. *Trends Genet,* 18**,** 352-8.

BROWN, C. E., LECHNER, T., HOWE, L. & WORKMAN, J. L. 2000. The many HATs of transcription coactivators. *Trends Biochem Sci,* 25**,** 15-9.

BU, D., YU, K., SUN, S., XIE, C., SKOGERBO, G., MIAO, R., XIAO, H., LIAO, Q., LUO, H., ZHAO, G., ZHAO, H., LIU, Z., LIU, C., CHEN, R. & ZHAO, Y. 2012. NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res,* 40**,** D210-5.

BULGER, M. & GROUDINE, M. 2010. Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev Biol,* 339**,** 250-7.

BURATOWSKI, S. 2005. Connections between mRNA 3' end processing and transcription termination. *Curr Opin Cell Biol,* 17**,** 257-61.

BUSKE, F. A., MATTICK, J. S. & BAILEY, T. L. 2011. Potential in vivo roles of nucleic acid triple-helices. *RNA Biol,* 8**,** 427-39.

BUSSOTTI, G., LEONARDI, T., CLARK, M. B., MERCER, T. R., CRAWFORD, J., MALQUORI, L., NOTREDAME, C., DINGER, M. E., MATTICK, J. S. & ENRIGHT, A. J. 2016. Improved definition of the mouse transcriptome via targeted RNA sequencing. *Genome Res,* 26**,** 705-16.

CABILI, M. N., TRAPNELL, C., GOFF, L., KOZIOL, M., TAZON-VEGA, B., REGEV, A. & RINN, J. L. 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev,* 25**,** 1915-27.

CARRIERI, C., CIMATTI, L., BIAGIOLI, M., BEUGNET, A., ZUCCHELLI, S., FEDELE, S., PESCE, E., FERRER, I., COLLAVIN, L., SANTORO, C., FORREST, A. R., CARNINCI, P., BIFFO, S., STUPKA, E. & GUSTINCICH, S. 2012. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature,* 491**,** 454-457.

CESANA, M., CACCHIARELLI, D., LEGNINI, I., SANTINI, T., STHANDIER, O., CHINAPPI, M., TRAMONTANO, A. & BOZZONI, I. 2011. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell,* 147**,** 358-69.

CHARITE, J., MCFADDEN, D. G., MERLO, G., LEVI, G., CLOUTHIER, D. E., YANAGISAWA, M., RICHARDSON, J. A. & OLSON, E. N. 2001. Role of Dlx6 in regulation of an endothelin-1-dependent, dHAND branchial arch enhancer. *Genes Dev,* 15**,** 3039-49.

CHEN, G., WANG, Z., WANG, D., QIU, C., LIU, M., CHEN, X., ZHANG, Q., YAN, G. & CUI, Q. 2013. LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res,* 41**,** D983-6.

CHEN, K., CHEN, Z., WU, D., ZHANG, L., LIN, X., SU, J., RODRIGUEZ, B., XI, Y., XIA, Z., CHEN, X., SHI, X., WANG, Q. & LI, W. 2015a. Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nat Genet,* 47**,** 1149-57.

CHEN, L., KOSTADIMA, M., MARTENS, J. H., CANU, G., GARCIA, S. P., TURRO, E., DOWNES, K., MACAULAY, I. C., BIELCZYK-MACZYNSKA, E., COE, S., FARROW, S., POUDEL, P., BURDEN, F., JANSEN, S. B., ASTLE, W. J., ATTWOOD, A., BARIANA, T., DE BONO, B., BRESCHI, A., CHAMBERS, J. C., CHOUDRY, F. A., CLARKE, L., COUPLAND, P., VAN DER ENT, M., ERBER, W. N., JANSEN, J. H., FAVIER, R., FENECH, M. E., FOAD, N., FRESON, K., VAN GEET, C., GOMEZ, K., GUIGO, R., HAMPSHIRE, D., KELLY, A. M., KERSTENS, H. H., KOONER, J. S., LAFFAN, M., LENTAIGNE, C., LABALETTE, C., MARTIN, T., MEACHAM, S., MUMFORD, A., NURNBERG, S., PALUMBO, E., VAN DER REIJDEN, B. A., RICHARDSON, D., SAMMUT, S. J., SLODKOWICZ, G., TAMURI, A. U., VASQUEZ, L., VOSS, K., WATT, S., WESTBURY, S., FLICEK, P., LOOS, R., GOLDMAN, N., BERTONE, P., READ, R. J., RICHARDSON, S., CVEJIC, A., SORANZO, N., OUWEHAND, W. H., STUNNENBERG, H. G., FRONTINI, M. & RENDON, A. 2014. Transcriptional diversity during lineage commitment of human blood progenitors. *Science,* 345**,** 1251033.

CHEN, X., YAN, C. C., LUO, C., JI, W., ZHANG, Y. & DAI, Q. 2015b. Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci Rep,* 5**,** 11338.

CONNELLY, S. & MANLEY, J. L. 1988. A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. *Genes Dev,* 2**,** 440-52.

CONSORTIUM, E. P. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science,* 306**,** 636-40.

CONSORTIUM, F., THE, R. P., CLST, FORREST, A. R., KAWAJI, H., REHLI, M., BAILLIE, J. K., DE HOON, M. J., HABERLE, V., LASSMANN, T., KULAKOVSKIY, I. V., LIZIO, M., ITOH, M., ANDERSSON, R., MUNGALL, C. J., MEEHAN, T. F., SCHMEIER, S., BERTIN, N., JORGENSEN, M., DIMONT, E., ARNER, E., SCHMIDL, C., SCHAEFER, U., MEDVEDEVA, Y. A., PLESSY, C., VITEZIC, M., SEVERIN, J., SEMPLE, C., ISHIZU, Y., YOUNG, R. S., FRANCESCATTO, M., ALAM, I., ALBANESE, D., ALTSCHULER, G. M., ARAKAWA, T., ARCHER, J. A., ARNER, P., BABINA, M., RENNIE, S., BALWIERZ, P. J., BECKHOUSE, A. G., PRADHAN-BHATT, S., BLAKE, J. A., BLUMENTHAL, A., BODEGA, B., BONETTI, A., BRIGGS, J., BROMBACHER, F., BURROUGHS, A. M., CALIFANO, A., CANNISTRACI, C. V., CARBAJO, D., CHEN, Y., CHIERICI, M., CIANI, Y., CLEVERS, H. C., DALLA, E., DAVIS, C. A., DETMAR, M., DIEHL, A. D., DOHI, T., DRABLOS, F., EDGE, A. S., EDINGER, M., EKWALL, K., ENDOH, M., ENOMOTO, H., FAGIOLINI, M., FAIRBAIRN, L., FANG, H., FARACH-CARSON, M.

C., FAULKNER, G. J., FAVOROV, A. V., FISHER, M. E., FRITH, M. C., FUJITA, R., FUKUDA, S., FURLANELLO, C., FURINO, M., FURUSAWA, J., GEIJTENBEEK, T. B., GIBSON, A. P., GINGERAS, T., GOLDOWITZ, D., GOUGH, J., GUHL, S., GULER, R., GUSTINCICH, S., HA, T. J., HAMAGUCHI, M., HARA, M., HARBERS, M., HARSHBARGER, J., HASEGAWA, A., HASEGAWA, Y., HASHIMOTO, T., HERLYN, M., HITCHENS, K. J., HO SUI, S. J., HOFMANN, O. M., et al. 2014. A promoter-level mammalian expression atlas. *Nature,* 507**,** 462-70.

CORE, L. J., WATERFALL, J. J. & LIS, J. T. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science,* 322**,** 1845-8.

DADI, S., LE NOIR, S., PAYET-BORNET, D., LHERMITTE, L., ZACARIAS-CABEZA, J., BERGERON, J., VILLARESE, P., VACHEZ, E., DIK, W. A., MILLIEN, C., RADFORD, I., VERHOEYEN, E., COSSET, F. L., PETIT, A., IFRAH, N., DOMBRET, H., HERMINE, O., SPICUGLIA, S., LANGERAK, A. W., MACINTYRE, E. A., NADEL, B., FERRIER, P. & ASNAFI, V. 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**,** 563-76.

DAHL, J. A., JUNG, I., AANES, H., GREGGAINS, G. D., MANAF, A., LERDRUP, M., LI, G., KUAN, S., LI, B., LEE, A. Y., PREISSL, S., JERMSTAD, I., HAUGEN, M. H., SUGANTHAN, R., BJORAS, M., HANSEN, K., DALEN, K. T., FEDORCSAK, P., REN, B. & KLUNGLAND, A. 2016. Broad histone H3K4me3 domains in mouse oocytes modulate maternal-to-zygotic transition. *Nature,* 537**,** 548-552.

DEATON, A. M., WEBB, S., KERR, A. R., ILLINGWORTH, R. S., GUY, J., ANDREWS, R. & BIRD, A. 2011. Cell type-specific DNA methylation at intragenic CpG islands in the immune system. *Genome Res,* 21**,** 1074-86.

DELLINO, G. I., SCHWARTZ, Y. B., FARKAS, G., MCCABE, D., ELGIN, S. C. R. & PIRROTTA, V. 2004. Polycomb silencing blocks trascription initiation. *Mol Cell,* 13**,** 887-893.

DERRIEN, T., GUIGO, R. & JOHNSON, R. 2011. The Long Non-Coding RNAs: A New (P)layer in the "Dark Matter". *Front Genet,* 2**,** 107.

DERRIEN, T., JOHNSON, R., BUSSOTTI, G., TANZER, A., DJEBALI, S., TILGNER, H., GUERNEC, G., MARTIN, D., MERKEL, A., KNOWLES, D. G., LAGARDE, J., VEERAVALLI, L., RUAN, X., RUAN, Y., LASSMANN, T., CARNINCI, P., BROWN, J. B., LIPOVICH, L., GONZALEZ, J. M., THOMAS, M., DAVIS, C. A., SHIEKHATTAR, R., GINGERAS, T. R., HUBBARD, T. J., NOTREDAME, C., HARROW, J. & GUIGO, R. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res,* 22**,** 1775-89.

DIK, W. A., PIKE-OVERZET, K., WEERKAMP, F., DE RIDDER, D., DE HAAS, E. F., BAERT, M. R., VAN DER, S. P., KOSTER, E. E., REINDERS, M. J., VAN DONGEN, J. J., LANGERAK, A. W. & STAAL, F. J. 2005. 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.

DINGER, M. E., AMARAL, P. P., MERCER, T. R., PANG, K. C., BRUCE, S. J., GARDINER, B. B., ASKARIAN-AMIRI, M. E., RU, K., SOLDA, G., SIMONS, C., SUNKIN, S. M., CROWE, M. L., GRIMMOND, S. M., PERKINS, A. C. & MATTICK, J. S. 2008. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res,* 18**,** 1433-45.

DION, M. F., ALTSCHULER, S. J., WU, L. F. & RANDO, O. J. 2005. Genomic characterization reveals a simple histone H4 acetylation code. *Proc Natl Acad Sci U S A,* 102**,** 5501-6.

DJEBALI, S., DAVIS, C. A., MERKEL, A., DOBIN, A., LASSMANN, T., MORTAZAVI, A., TANZER, A., LAGARDE, J., LIN, W., SCHLESINGER, F., XUE, C., MARINOV, G. K., KHATUN, J., WILLIAMS, B. A., ZALESKI, C., ROZOWSKY, J., RODER, M., KOKOCINSKI, F., ABDELHAMID, R. F., ALIOTO, T., ANTOSHECHKIN, I., BAER, M. T., BAR, N. S., BATUT, P., BELL, K., BELL, I., CHAKRABORTTY, S., CHEN, X., CHRAST, J., CURADO, J., DERRIEN, T., DRENKOW, J., DUMAIS, E., DUMAIS, J., DUTTAGUPTA, R., FALCONNET, E., FASTUCA, M., FEJES-TOTH, K., FERREIRA, P., FOISSAC, S., FULLWOOD, M. J., GAO, H., GONZALEZ, D., GORDON, A., GUNAWARDENA, H., HOWALD, C., JHA, S., JOHNSON, R., KAPRANOV, P., KING, B., KINGSWOOD, C., LUO, O. J., PARK, E., PERSAUD, K., PREALL, J. B., RIBECA, P., RISK, B., ROBYR, D., SAMMETH, M., SCHAFFER, L., SEE, L. H., SHAHAB, A., SKANCKE, J., SUZUKI, A. M., TAKAHASHI, H., TILGNER, H., TROUT, D., WALTERS, N., WANG, H., WROBEL, J., YU, Y., RUAN, X., HAYASHIZAKI, Y., HARROW, J., GERSTEIN, M., HUBBARD, T., REYMOND, A., ANTONARAKIS, S. E., HANNON, G., GIDDINGS, M. C., RUAN, Y., WOLD, B., CARNINCI, P., GUIGO, R. & GINGERAS, T. R. 2012. Landscape of transcription in human cells. *Nature,* 489**,** 101-8.

DU, Z., FEI, T., VERHAAK, R. G., SU, Z., ZHANG, Y., BROWN, M., CHEN, Y. & LIU, X. S. 2013. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat Struct Mol Biol,* 20**,** 908-13.

DUDLEY, E. C., PETRIE, H. T., SHAH, L. M., OWEN, M. J. & HAYDAY, A. C. 1994. T cell receptor · chain gene rearrangement and selection during thymocyte development in adult mice. *Immunity,* 1**,** 83-93.

EBISUYA, M., YAMAMOTO, T., NAKAJIMA, M. & NISHIDA, E. 2008. Ripples from neighbouring transcription. *Nat Cell Biol,* 10**,** 1106-13.

EDWARDS, C. A. & FERGUSON-SMITH, A. C. 2007. Mechanisms regulating imprinted genes in clusters. *Curr Opin Cell Biol,* 19**,** 281-9.

EHRLICH, M., GAMA-SOSA, M. A., HUANG, L. H., MIDGETT, R. M., KUO, K. C., MCCUNE, R. A. & GEHRKE, C. 1982. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res,* 10**,** 2709-21.

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. & BERNSTEIN, B. E. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature,* 473**,** 43-9.

ESTELLER, M. 2011. Non-coding RNAs in human disease. *Nat Rev Genet,* 12**,** 861-74.

FEHLING, H. J. & VON BOEHMER, H. 1997. Early ‡· T cell development in the thymus of normal and genetically altered mice. *Curr.Opin.Immunol.,* 9**,** 263-275.

FEJES, A. P., ROBERTSON, G., BILENKY, M., VARHOL, R., BAINBRIDGE, M. & JONES, S. J. 2008. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics,* 24**,** 1729-30.

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., LANDER, E. S., GOLUB, T. R. & LOOK, A. T. 2002. Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer Cell,* 1**,** 75-87.

FISCHLE, W., TSENG, B. S., DORMANN, H. L., UEBERHEIDE, B. M., GARCIA, B. A., SHABANOWITZ, J., HUNT, D. F., FUNABIKI, H. & ALLIS, C. D. 2005. Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. *Nature,* 438**,** 1116-1122.

FOSS, D. L., DONSKOY, E. & GOLDSCHNEIDER, I. 2001. The importation of hematogenous precursors by the thymus is a gated phenomenon in normal adult mice. *J Exp Med,* 193**,** 365-74.

FRADET, Y., SAAD, F., APRIKIAN, A., DESSUREAULT, J., ELHILALI, M., TRUDEL, C., MASSE, B., PICHE, L. & CHYPRE, C. 2004. uPM3, a new molecular urine test for the detection of prostate cancer. *Urology,* 64**,** 311-5; discussion 315-6.

FUJINAGA, K., IRWIN, D., HUANG, Y., TAUBE, R., KUROSU, T. & PETERLIN, B. M. 2004. Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element. *Mol Cell Biol*, 24**,** 787-95.

GALE, R. P. 1987. Development of the immune system in human fetal liver. *Thymus,* 10**,** 45-56.

GALY, A., VERMA, S., BARCENA, A. & SPITS, H. 1993. Precursors of CD3+CD4+CD8+ cells in the human thymus are defined by expression of CD34. Delineation of early events in human thymic development. *J Exp Med,* 178**,** 391-401.

GARZON, R., VOLINIA, S., PAPAIOANNOU, D., NICOLET, D., KOHLSCHMIDT, J., YAN, P. S., MROZEK, K., BUCCI, D., CARROLL, A. J., BAER, M. R., WETZLER, M., CARTER, T. H., POWELL, B. L., KOLITZ, J. E., MOORE, J. O., EISFELD, A. K., BLACHLY, J. S., BLUM, W., CALIGIURI, M. A., STONE, R. M., MARCUCCI, G., CROCE, C. M., BYRD, J. C. & BLOOMFIELD, C. D. 2014. Expression and prognostic impact of lncRNAs in acute myeloid leukemia. *Proc Natl Acad Sci U S A,* 111**,** 18679-84.

GE, M., LI, A. & WANG, M. 2016. A Bipartite Network-based Method for Prediction of Long Non-coding RNA-protein Interactions. *Genomics Proteomics Bioinformatics,* 14**,** 62-71.

GERASIMOVA, T. I., BYRD, K. & CORCES, V. G. 2000. A chromatin insulator determines the nuclear localization of DNA. *Mol Cell,* 6**,** 1025-1035.

GERBER, M. & SHILATIFARD, A. 2003. Transcriptional elongation by RNA polymerase II and histone methylation. *J Biol Chem,* 278**,** 26303-6.

GERSTEIN, M. B., LU, Z. J., VAN NOSTRAND, E. L., CHENG, C., ARSHINOFF, B. I., LIU, T., YIP, K. Y., ROBILOTTO, R., RECHTSTEINER, A., IKEGAMI, K., ALVES, P., CHATEIGNER, A., PERRY, M., MORRIS, M., AUERBACH, R. K., FENG, X., LENG, J., VIELLE, A., NIU, W., RHRISSORRAKRAI, K., AGARWAL, A., ALEXANDER, R. P., BARBER, G., BRDLIK, C. M., BRENNAN, J., BROUILLET, J. J., CARR, A., CHEUNG, M. S., CLAWSON, H., CONTRINO, S., DANNENBERG, L. O., DERNBURG, A. F., DESAI, A., DICK, L., DOSE, A. C., DU, J., EGELHOFER, T., ERCAN, S., EUSKIRCHEN, G., EWING, B., FEINGOLD, E. A., GASSMANN, R., GOOD, P. J., GREEN, P., GULLIER, F., GUTWEIN, M., GUYER, M. S., HABEGGER, L., HAN, T., HENIKOFF, J. G., HENZ, S. R., HINRICHS, A., HOLSTER, H., HYMAN, T., INIGUEZ, A. L., JANETTE, J., JENSEN, M., KATO, M., KENT, W. J., KEPHART, E., KHIVANSARA, V., KHURANA, E., KIM, J. K., KOLASINSKA-ZWIERZ, P., LAI, E. C., LATORRE, I., LEAHEY, A., LEWIS, S., LLOYD, P., LOCHOVSKY, L., LOWDON, R. F., LUBLING, Y., LYNE, R., MACCOSS, M., MACKOWIAK, S. D., MANGONE, M., MCKAY, S., MECENAS, D., MERRIHEW, G., MILLER, D. M., 3RD, MUROYAMA, A., MURRAY, J. I., OOI, S. L., PHAM, H., PHIPPEN, T., PRESTON, E. A., RAJEWSKY, N., RATSCH, G., ROSENBAUM, H., ROZOWSKY, J., RUTHERFORD, K., RUZANOV, P., SAROV, M., SASIDHARAN, R., SBONER, A., SCHEID, P., SEGAL, E., SHIN, H., SHOU, C., SLACK, F. J., et al. 2010. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science,* 330**,** 1775-87.

GEYER, P. K. 1997. The role of insulator elements in defining domains of gene expression. *Curr.Opin.Genet.Dev.,* 7**,** 242-248.

GILLIES, S. D., MORRISON, S. L., OI, V. T. & TONEGAWA, S. 1983. A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell,* 33**,** 717-728.

GLAZKO, G. V., ZYBAILOV, B. L. & ROGOZIN, I. B. 2012. Computational prediction of polycomb-associated long non-coding RNAs. *PLoS One,* 7**,** e44878.

GOLDSCHNEIDER, I. 2006. Cyclical mobilization and gated importation of thymocyte progenitors in the adult mouse: evidence for a thymus-bone marrow feedback loop. *Immunol Rev,* 209**,** 58-75.

GONG, C. & MAQUAT, L. E. 2011. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature,* 470**,** 284-8.

GRIFFON, A., BARBIER, Q., DALINO, J., VAN HELDEN, J., SPICUGLIA, S. & BALLESTER, B. 2015. Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res,* 43**,** e27.

GROSVELD, F. G., BLOM VAN ASSENDELFT, G., GREAVES, D. R. & KOLLIAS, G. 1987. Position-independant , high level expression of the human ·-globin gene in transgenic mice. *Cell,* 51**,** 975-985.

GUO, G., KANG, Q., ZHU, X., CHEN, Q., WANG, X., CHEN, Y., OUYANG, J., ZHANG, L., TAN, H., CHEN, R., HUANG, S. & CHEN, J. L. 2014. A long noncoding RNA critically regulates Bcr-Abl-mediated cellular transformation by acting as a competitive endogenous RNA. *Oncogene*.

GUO, X., GAO, L., LIAO, Q., XIAO, H., MA, X., YANG, X., LUO, H., ZHAO, G., BU, D., JIAO, F., SHAO, Q., CHEN, R. & ZHAO, Y. 2013. Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. *Nucleic Acids Res,* 41**,** e35.

GUPTA, R. A., SHAH, N., WANG, K. C., KIM, J., HORLINGS, H. M., WONG, D. J., TSAI, M. C., HUNG, T., ARGANI, P., RINN, J. L., WANG, Y., BRZOSKA, P., KONG, B., LI, R., WEST, R. B., VAN DE VIJVER, M. J., SUKUMAR, S. & CHANG, H. Y. 2010. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature,* 464**,** 1071-6.

GUTTMAN, M., GARBER, M., LEVIN, J. Z., DONAGHEY, J., ROBINSON, J., ADICONIS, X., FAN, L., KOZIOL, M. J., GNIRKE, A., NUSBAUM, C., RINN, J. L., LANDER, E. S. & REGEV, A. 2010. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol,* 28**,** 503-10.

HAMPSEY, M. & REINBERG, D. 2003. Tails of intrigue: Phosphorylation of RNA Polymerase II mediates histone methylation. *Cell,* 113**,** 429-432.

HAO, Y., WU, W., LI, H., YUAN, J., LUO, J., ZHAO, Y. & CHEN, R. 2016. NPInter v3.0: an upgraded database of noncoding RNA-associated interactions. *Database (Oxford),* 2016.

HARROW, J., FRANKISH, A., GONZALEZ, J. M., TAPANARI, E., DIEKHANS, M., KOKOCINSKI, F., AKEN, B. L., BARRELL, D., ZADISSA, A., SEARLE, S., BARNES, I., BIGNELL, A., BOYCHENKO, V., HUNT, T., KAY, M., MUKHERJEE, G., RAJAN, J., DESPACIO-REYES, G., SAUNDERS, G., STEWARD, C., HARTE, R., LIN, M., HOWALD, C., TANZER, A., DERRIEN, T., CHRAST, J., WALTERS, N., BALASUBRAMANIAN, S., PEI, B., TRESS, M., RODRIGUEZ, J. M., EZKURDIA, I., VAN BAREN, J., BRENT, M., HAUSSLER, D., KELLIS, M., VALENCIA, A., REYMOND, A., GERSTEIN, M., GUIGO, R. & HUBBARD, T. J. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res,* 22**,** 1760-74.

HAYNES, B. F. & HEINLY, C. S. 1995. Early human T cell development: analysis of the human thymus at the time of initial entry of hematopoietic stem cells into the fetal thymic microenvironment. *J Exp Med,* 181**,** 1445-58.

HEINTZMAN, N. D., STUART, R. K., HON, G., FU, Y., CHING, C. W., HAWKINS, R. D., BARRERA, L. O., VAN CALCAR, S., QU, C., CHING, K. A., WANG, W., WENG, Z., GREEN, R. D., CRAWFORD, G. E. & REN, B. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet,* 39**,** 311-8.

HEINZ, S., BENNER, C., SPANN, N., BERTOLINO, E., LIN, Y. C., LASLO, P., CHENG, J. X., MURRE, C., SINGH, H. & GLASS, C. K. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell,* 38**,** 576-89.

HEZRONI, H., KOPPSTEIN, D., SCHWARTZ, M. G., AVRUTIN, A., BARTEL, D. P. & ULITSKY, I. 2015. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep,* 11**,** 1110-22.

HNISZ, D., ABRAHAM, B. J., LEE, T. I., LAU, A., SAINT-ANDRE, V., SIGOVA, A. A., HOKE, H. A. & YOUNG, R. A. 2013. Super-enhancers in the control of cell identity and disease. *Cell,* 155**,** 934-47.

HNISZ, D., WEINTRAUB, A. S., DAY, D. S., VALTON, A. L., BAK, R. O., LI, C. H., GOLDMANN, J., LAJOIE, B. R., FAN, Z. P., SIGOVA, A. A., REDDY, J., BORGES-RIVERA, D., LEE, T. I., JAENISCH, R., PORTEUS, M. H., DEKKER, J. & YOUNG, R. A. 2016. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science,* 351**,** 1454-8.

HOLSTEGE, F. C., FIEDLER, U. & TIMMERS, H. T. 1997. Three transitions in the RNA polymerase II transcription complex during initiation. *EMBO J,* 16**,** 7468-80.

HOLSTEGE, F. C., JENNINGS, E. G., WYRICK, J. J., LEE, T. I., HENGARTNER, C. J., GREEN, M. R., GOLUB, T. R., LANDER, E. S. & YOUNG, R. A. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell,* 95**,** 717-728.

HOMER, N., MERRIMAN, B. & NELSON, S. F. 2009. BFAST: an alignment tool for large scale genome resequencing. *PLoS One,* 4**,** e7767.

HOMMINGA, I., PIETERS, R., LANGERAK, A. W., DE ROOI, J. J., STUBBS, A., VERSTEGEN, M., VUERHARD, M., BUIJS-GLADDINES, J., KOOI, C., KLOUS, P., VAN VLIERBERGHE, P., FERRANDO, A. A., CAYUELA, J. M., VERHAAF, B., BEVERLOO, H. B., HORSTMANN, M., DE HAAS, V., WIEKMEIJER, A. S., PIKE-OVERZET, K., STAAL, F. J., DE LAAT, W., SOULIER, J., SIGAUX, F. & MEIJERINK, J. P. 2011. Integrated transcript and genome analyses reveal NKX2-1 and MEF2C as potential oncogenes in T cell acute lymphoblastic leukemia. *Cancer Cell,* 19**,** 484-97.

HU, W., ALVAREZ-DOMINGUEZ, J. R. & LODISH, H. F. 2012. Regulation of mammalian cell differentiation by long non-coding RNAs. *EMBO Rep,* 13**,** 971-83.

HUANG, M. D., CHEN, W. M., QI, F. Z., SUN, M., XU, T. P., MA, P. & SHU, Y. Q. 2015. Long non-coding RNA TUG1 is up-regulated in hepatocellular carcinoma and promotes cell growth and apoptosis by epigenetically silencing of KLF2. *Mol Cancer,* 14**,** 165.

HUARTE, M., GUTTMAN, M., FELDSER, D., GARBER, M., KOZIOL, M. J., KENZELMANN-BROZ, D., KHALIL, A. M., ZUK, O., AMIT, I., RABANI, M., ATTARDI, L. D., REGEV, A., LANDER, E. S., JACKS, T. & RINN, J. L. 2010. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell,* 142**,** 409-19.

HUGUET, F., LEGUAY, T., RAFFOUX, E., THOMAS, X., BELDJORD, K., DELABESSE, E., CHEVALLIER, P., BUZYN, A., DELANNOY, A., CHALANDON, Y., VERNANT, J. P., LAFAGE-POCHITALOFF, M., CHASSEVENT, A., LHERITIER, V., MACINTYRE, E., BENE, M. C., IFRAH, N. & DOMBRET, H. 2009. Pediatric-inspired therapy in adults with Philadelphia chromosome-negative acute lymphoblastic leukemia: the GRAALL-2003 study. *J Clin Oncol,* 27**,** 911-8.

INTERNATIONAL HUMAN GENOME SEQUENCING, C. 2004. Finishing the euchromatic sequence of the human genome. *Nature,* 431**,** 931-45.

IYER, M. K., NIKNAFS, Y. S., MALIK, R., SINGHAL, U., SAHU, A., HOSONO, Y., BARRETTE, T. R., PRENSNER, J. R., EVANS, J. R., ZHAO, S., POLIAKOV, A., CAO, X., DHANASEKARAN, S. M., WU, Y. M., ROBINSON, D. R., BEER, D. G., FENG, F. Y., IYER, H. K. & CHINNAIYAN, A. M. 2015. The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet*.

JAENISCH, R. & BIRD, A. 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet,* 33 Suppl**,** 245-54.

JENJAROENPUN, P., WONGSURAWAT, T., YENAMANDRA, S. P. & KUZNETSOV, V. A. 2015. QmRLFS-finder: a model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences. *Nucleic Acids Res,* 43**,** 10081.

JHUNJHUNWALA, S., VAN ZELM, M. C., PEAK, M. M. & MURRE, C. 2009. Chromatin architecture and the generation of antigen receptor diversity. *Cell,* 138**,** 435-48.

JOACHIMS, M. L., CHAIN, J. L., HOOKER, S. W., KNOTT-CRAIG, C. J. & THOMPSON, L. F. 2006. Human ‡· and g Î Thymocyte Development: TCR Gene Rearrangements, Intracellular TCR· Expression, and g Î Developmental Potential--Differences between Men and Mice. *J Immunol,* 176**,** 1543-1552.

JONES, P. A. & BAYLIN, S. B. 2007. The epigenomics of cancer. *Cell,* 128**,** 683-92.

JOTHI, R., CUDDAPAH, S., BARSKI, A., CUI, K. & ZHAO, K. 2008. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res,* 36**,** 5221-31.

KANEHISA, M., GOTO, S., SATO, Y., KAWASHIMA, M., FURUMICHI, M. & TANABE, M. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res,* 42**,** D199-205.

KAPUSTA, A., KRONENBERG, Z., LYNCH, V. J., ZHUO, X., RAMSAY, L., BOURQUE, G., YANDELL, M. & FESCHOTTE, C. 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet,* 9**,** e1003470.

KAWAI, J., SHINAGAWA, A., SHIBATA, K., YOSHINO, M., ITOH, M., ISHII, Y., ARAKAWA, T., HARA, A., FUKUNISHI, Y., KONNO, H., ADACHI, J., FUKUDA, S., AIZAWA, K., IZAWA, M., NISHI, K., KIYOSAWA, H., KONDO, S., YAMANAKA, I., SAITO, T., OKAZAKI, Y., GOJOBORI, T., BONO, H., KASUKAWA, T., SAITO, R., KADOTA, K., MATSUDA, H., ASHBURNER, M., BATALOV, S., CASAVANT, T., FLEISCHMANN, W., GAASTERLAND, T., GISSI, C., KING, B., KOCHIWA, H., KUEHL, P., LEWIS, S., MATSUO, Y., NIKAIDO, I., PESOLE, G., QUACKENBUSH, J., SCHRIML, L. M., STAUBLI, F., SUZUKI, R., TOMITA, M., WAGNER, L., WASHIO, T., SAKAI, K., OKIDO, T., FURUNO, M., AONO, H., BALDARELLI, R., BARSH, G., BLAKE, J., BOFFELLI, D., BOJUNGA, N., CARNINCI, P., DE BONALDO, M. F., BROWNSTEIN, M. J., BULT, C., FLETCHER, C., FUJITA, M., GARIBOLDI, M., GUSTINCICH, S., HILL, D., HOFMANN, M., HUME, D. A., KAMIYA, M., LEE, N. H., LYONS, P., MARCHIONNI, L., MASHIMA, J., MAZZARELLI, J., MOMBAERTS, P., NORDONE, P., RING, B., RINGWALD, M., RODRIGUEZ, I., SAKAMOTO, N., SASAKI, H., SATO, K., SCHONBACH, C., SEYA, T., SHIBATA, Y., STORCH, K. F., SUZUKI, H., TOYO-OKA, K., WANG, K. H., WEITZ, C., WHITTAKER, C., WILMING, L., WYNSHAW-BORIS, A., YOSHIDA, K., HASEGAWA, Y., KAWAJI, H., KOHTSUKI, S., HAYASHIZAKI, Y., TEAM, R. G. E. R. G. P. I. & THE, F. C. 2001. Functional annotation of a full-length mouse cDNA collection. *Nature,* 409**,** 685-90.

KELLEY, D. R. & RINN, J. L. 2012. Transposable elements reveal a stem cell specific class of long noncoding RNAs. *Genome Biol,* 13**,** R107.

KHAN, A. & ZHANG, X. 2016. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res,* 44**,** D164-71.

KIELBASA, S. M., WAN, R., SATO, K., HORTON, P. & FRITH, M. C. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Res,* 21**,** 487-93.

KIM, D., PERTEA, G., TRAPNELL, C., PIMENTEL, H., KELLEY, R. & SALZBERG, S. L. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol,* 14**,** R36.

KIM, D. H., MARINOV, G. K., PEPKE, S., SINGER, Z. S., HE, P., WILLIAMS, B., SCHROTH, G. P., ELOWITZ, M. B. & WOLD, B. J. 2015. Single-cell transcriptome analysis reveals dynamic changes in lncRNA expression during reprogramming. *Cell Stem Cell,* 16**,** 88-101.

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. & GREENBERG, M. E. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature,* 465**,** 182-7.

KLOSE, R. J. & BIRD, A. P. 2006. Genomic DNA methylation: the mark and its mediators. *Trends Biochem Sci.*, 31**,** 89-97.
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. & ANDRAU, J. C. 2011. Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat Struct Mol Biol,* 18**,** 956-63.
KOERNER, M. V., PAULER, F. M., HUANG, R. & BARLOW, D. P. 2009. The function of non-coding RNAs in genomic imprinting. *Development,* 136**,** 1771-83.
KOGO, R., SHIMAMURA, T., MIMORI, K., KAWAHARA, K., IMOTO, S., SUDO, T., TANAKA, F., SHIBATA, K., SUZUKI, A., KOMUNE, S., MIYANO, S. & MORI, M. 2011. Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res,* 71**,** 6320-6.
KONDO, M., WEISSMAN, I. L. & AKASHI, K. 1997. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell,* 91**,** 661-672.
KORNIENKO, A. E., DOTTER, C. P., GUENZL, P. M., GISSLINGER, H., GISSLINGER, B., CLEARY, C., KRALOVICS, R., PAULER, F. M. & BARLOW, D. P. 2016. Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biol,* 17**,** 14.
KOUADJO, K. E., NISHIDA, Y., CADRIN-GIRARD, J. F., YOSHIOKA, M. & ST-AMAND, J. 2007. Housekeeping and tissue-specific genes in mouse tissues. *BMC Genomics,* 8**,** 127.
KOUZARIDES, T. 2007. Chromatin modifications and their function. *Cell,* 128**,** 693-705.
KRETZ, M., SIPRASHVILI, Z., CHU, C., WEBSTER, D. E., ZEHNDER, A., QU, K., LEE, C. S., FLOCKHART, R. J., GROFF, A. F., CHOW, J., JOHNSTON, D., KIM, G. E., SPITALE, R. C., FLYNN, R. A., ZHENG, G. X., AIYER, S., RAJ, A., RINN, J. L., CHANG, H. Y. & KHAVARI, P. A. 2013. Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature,* 493**,** 231-5.
LAI, F. & SHIEKHATTAR, R. 2014. Enhancer RNAs: the new molecules of transcription. *Curr Opin Genet Dev,* 25**,** 38-42.
LANCTOT, C., CHEUTIN, T., CREMER, M., CAVALLI, G. & CREMER, T. 2007. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat Rev Genet,* 8**,** 104-15.
LANGFELDER, P. & HORVATH, S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics,* 9**,** 559.
LANGMEAD, B. & SALZBERG, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods,* 9**,** 357-9.
LANGMEAD, B., TRAPNELL, C., POP, M. & SALZBERG, S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol,* 10**,** R25.
LEE, J. T. 2011. Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control. *Nat Rev Mol Cell Biol,* 12**,** 815-26.
LEPOIVRE, C., BELHOCINE, M., BERGON, A., GRIFFON, A., YAMMINE, M., VANHILLE, L., ZACARIAS-CABEZA, J., GARIBAL, M. A., KOCH, F., MAQBOOL, M. A., FENOUIL, R., LORIOD, B., HOLOTA, H., GUT, M., GUT, I., IMBERT, J., ANDRAU, J. C., PUTHIER, D. & SPICUGLIA, S. 2013. Divergent transcription is associated with promoters of transcriptional regulators. *BMC Genomics,* 14**,** 914.
LI, B., CAREY, M. & WORKMAN, J. L. 2007. The Role of Chromatin during Transcription. *Cell,* 128**,** 707-719.
LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.
LI, Q., PETERSON, K. R., FANG, X. & STAMATOYANNOPOULOS, G. 2002. Locus control regions. *Blood,* 100**,** 3077-3086.
LIAO, Q., LIU, C., YUAN, X., KANG, S., MIAO, R., XIAO, H., ZHAO, G., LUO, H., BU, D., ZHAO, H., SKOGERBO, G., WU, Z. & ZHAO, Y. 2011. Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic Acids Res,* 39**,** 3864-78.
LIBBRECHT, M. W. & NOBLE, W. S. 2015. Machine learning applications in genetics and genomics. *Nat Rev Genet,* 16**,** 321-32.
LIU, B. A., ENGELMANN, B. W. & NASH, P. D. 2012. High-throughput analysis of peptide-binding modules. *Proteomics,* 12**,** 1527-46.
LIU, G. Y., ZHAO, G. N., CHEN, X. F., HAO, D. L., ZHAO, X., LV, X. & LIU, D. P. 2016a. The long noncoding RNA Gm15055 represses Hoxa gene expression by recruiting PRC2 to the gene cluster. *Nucleic Acids Res,* 44**,** 2613-27.
LIU, X., WANG, C., LIU, W., LI, J., LI, C., KOU, X., CHEN, J., ZHAO, Y., GAO, H., WANG, H., ZHANG, Y., GAO, Y. & GAO, S. 2016b. Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos. *Nature,* 537**,** 558-562.
LIU, Y., SANOFF, H. K., CHO, H., BURD, C. E., TORRICE, C., MOHLKE, K. L., IBRAHIM, J. G., THOMAS, N. E. & SHARPLESS, N. E. 2009. INK4/ARF transcript expression is associated with chromosome 9p21 variants linked to atherosclerosis. *PLoS One,* 4**,** e5027.
LIU, Y., SUBRAHMANYAM, R., CHAKRABORTY, T., SEN, R. & DESIDERIO, S. 2007. A plant homeodomain in RAG-2 that binds hypermethylated lysine 4 of histone H3 is necessary for efficient antigen-receptor-gene rearrangement. *Immunity.,* 27**,** 561-571.
LOGAN, J., FALCK-PEDERSEN, E., DARNELL, J. E., JR. & SHENK, T. 1987. A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. *Proc Natl Acad Sci U S A,* 84**,** 8306-10.
LOVE, P. E. & BHANDOOLA, A. 2011. Signal integration and crosstalk during thymocyte migration and emigration. *Nat Rev Immunol,* 11**,** 469-77.
LUGER, K., MADER, A. W., RICHMOND, R. K., SARGENT, D. F. & RICHMOND, T. J. 1997. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature,* 389**,** 251-260.
LUO, S., LU, J. Y., LIU, L., YIN, Y., CHEN, C., HAN, X., WU, B., XU, R., LIU, W., YAN, P., SHAO, W., LU, Z., LI, H., NA, J., TANG, F., WANG, J., ZHANG, Y. E. & SHEN, X. 2016. Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells. *Cell Stem Cell,* 18**,** 637-52.
MAGISTRI, M., FAGHIHI, M. A., ST LAURENT, G., 3RD & WAHLESTEDT, C. 2012. Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts. *Trends Genet,* 28**,** 389-96.
MALLICK, C. A., DUDLEY, E. C., VINEY, J. L., OWEN, M. J. & HAYDAY, A. C. 1993. Rearrangement and diversity of T cell receptor · chain genes in thymocytes: a critical role for the · chain in development. *Cell,* 73**,** 513-519.
MANSOUR, M. R., ABRAHAM, B. J., ANDERS, L., BEREZOVSKAYA, A., GUTIERREZ, A., DURBIN, A. D., ETCHIN, J., LAWTON, L., SALLAN, S. E., SILVERMAN, L. B., LOH, M. L., HUNGER, S. P., SANDA, T., YOUNG, R. A. & LOOK, A. T. 2014. Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science,* 346**,** 1373-7.
MARGARITIS, T. & HOLSTEGE, F. C. 2008. Poised RNA polymerase II gives pause for thought. *Cell,* 133**,** 581-4.
MARKS, P., RIFKIND, R. A., RICHON, V. M., BRESLOW, R., MILLER, T. & KELLY, W. K. 2001. Histone deacetylases and cancer: causes and therapies. *Nat Rev Cancer,* 1**,** 194-202.
MARQUES, A. C., HUGHES, J., GRAHAM, B., KOWALCZYK, M. S., HIGGS, D. R. & PONTING, C. P. 2013. Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs. *Genome Biol,* 14**,** R131.
MARSHALL, N. F. & PRICE, D. H. 1995. Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *J Biol Chem,* 270**,** 12335-8.

MATHELIER, A., FORNES, O., ARENILLAS, D. J., CHEN, C. Y., DENAY, G., LEE, J., SHI, W., SHYR, C., TAN, G., WORSLEY-HUNT, R., ZHANG, A. W., PARCY, F., LENHARD, B., SANDELIN, A. & WASSERMAN, W. W. 2016. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res,* 44**,** D110-5.

MATHIEU, E. L., BELHOCINE, M., DAO, L. T., PUTHIER, D. & SPICUGLIA, S. 2014. [Functions of lncRNA in development and diseases]. *Med Sci (Paris),* 30**,** 790-6.

MATTHEWS, A. G., KUO, A. J., RAMON-MAIQUES, S., HAN, S., CHAMPAGNE, K. S., IVANOV, D., GALLARDO, M., CARNEY, D., CHEUNG, P., CICCONE, D. N., WALTER, K. L., UTZ, P. J., SHI, Y., KUTATELADZE, T. G., YANG, W., GOZANI, O. & OETTINGER, M. A. 2007. RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination. *Nature,* 450**,** 1106-1110.

MEMCZAK, S., JENS, M., ELEFSINIOTI, A., TORTI, F., KRUEGER, J., RYBAK, A., MAIER, L., MACKOWIAK, S. D., GREGERSEN, L. H., MUNSCHAUER, M., LOEWER, A., ZIEBOLD, U., LANDTHALER, M., KOCKS, C., LE NOBLE, F. & RAJEWSKY, N. 2013. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature,* 495**,** 333-8.

MERCER, T. R., DINGER, M. E., SUNKIN, S. M., MEHLER, M. F. & MATTICK, J. S. 2008. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A,* 105**,** 716-21.

MESHORER, E. & MISTELI, T. 2006. Chromatin in pluripotent embryonic stem cells and differentiation. *Nat Rev Mol Cell Biol,* 7**,** 540-6.

MILACIC, M., HAW, R., ROTHFELS, K., WU, G., CROFT, D., HERMJAKOB, H., D'EUSTACHIO, P. & STEIN, L. 2012. Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers (Basel),* 4**,** 1180-211.

MILLER, J. F. & MITCHELL, G. F. 1967. The thymus and the precursors of antigen reactive cells. *Nature,* 216**,** 659-63.

MOD, E. C., ROY, S., ERNST, J., KHARCHENKO, P. V., KHERADPOUR, P., NEGRE, N., EATON, M. L., LANDOLIN, J. M., BRISTOW, C. A., MA, L., LIN, M. F., WASHIETL, S., ARSHINOFF, B. I., AY, F., MEYER, P. E., ROBINE, N., WASHINGTON, N. L., DI STEFANO, L., BEREZIKOV, E., BROWN, C. D., CANDEIAS, R., CARLSON, J. W., CARR, A., JUNGREIS, I., MARBACH, D., SEALFON, R., TOLSTORUKOV, M. Y., WILL, S., ALEKSEYENKO, A. A., ARTIERI, C., BOOTH, B. W., BROOKS, A. N., DAI, Q., DAVIS, C. A., DUFF, M. O., FENG, X., GORCHAKOV, A. A., GU, T., HENIKOFF, J. G., KAPRANOV, P., LI, R., MACALPINE, H. K., MALONE, J., MINODA, A., NORDMAN, J., OKAMURA, K., PERRY, M., POWELL, S. K., RIDDLE, N. C., SAKAI, A., SAMSONOVA, A., SANDLER, J. E., SCHWARTZ, Y. B., SHER, N., SPOKONY, R., STURGILL, D., VAN BAREN, M., WAN, K. H., YANG, L., YU, C., FEINGOLD, E., GOOD, P., GUYER, M., LOWDON, R., AHMAD, K., ANDREWS, J., BERGER, B., BRENNER, S. E., BRENT, M. R., CHERBAS, L., ELGIN, S. C., GINGERAS, T. R., GROSSMAN, R., HOSKINS, R. A., KAUFMAN, T. C., KENT, W., KURODA, M. I., ORR-WEAVER, T., PERRIMON, N., PIRROTTA, V., POSAKONY, J. W., REN, B., RUSSELL, S., CHERBAS, P., GRAVELEY, B. R., LEWIS, S., MICKLEM, G., OLIVER, B., PARK, P. J., CELNIKER, S. E., HENIKOFF, S., KARPEN, G. H., LAI, E. C., MACALPINE, D. M., STEIN, L. D., WHITE, K. P. & KELLIS, M. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science,* 330**,** 1787-97.

MODARRESI, F., FAGHIHI, M. A., LOPEZ-TOLEDANO, M. A., FATEMI, R. P., MAGISTRI, M., BROTHERS, S. P., VAN DER BRUG, M. P. & WAHLESTEDT, C. 2012. Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation. *Nat Biotechnol,* 30**,** 453-9.

MONDAL, T., SUBHASH, S., VAID, R., ENROTH, S., UDAY, S., REINIUS, B., MITRA, S., MOHAMMED, A., JAMES, A. R., HOBERG, E., MOUSTAKAS, A., GYLLENSTEN, U., JONES, S. J., GUSTAFSSON, C. M., SIMS, A. H., WESTERLUND, F., GORAB, E. & KANDURI, C. 2015. MEG3 long noncoding RNA regulates the TGF-beta pathway genes through formation of RNA-DNA triplex structures. *Nat Commun,* 6**,** 7743.

MORSHEAD, K. B., CICCONE, D. N., TAVERNA, S. D., ALLIS, C. D. & OETTINGER, M. A. 2003. Antigen receptor loci poised for V(D)J rearrangement are broadly associated with BRG1 and flanked by peaks of histone H3 dimethylated at lysine 4. *Proc Natl Acad Sci USA,* 100**,** 11577-11582.

MOUSE, E. C., STAMATOYANNOPOULOS, J. A., SNYDER, M., HARDISON, R., REN, B., GINGERAS, T., GILBERT, D. M., GROUDINE, M., BENDER, M., KAUL, R., CANFIELD, T., GISTE, E., JOHNSON, A., ZHANG, M., BALASUNDARAM, G., BYRON, R., ROACH, V., SABO, P. J., SANDSTROM, R., STEHLING, A. S., THURMAN, R. E., WEISSMAN, S. M., CAYTING, P., HARIHARAN, M., LIAN, J., CHENG, Y., LANDT, S. G., MA, Z., WOLD, B. J., DEKKER, J., CRAWFORD, G. E., KELLER, C. A., WU, W., MORRISSEY, C., KUMAR, S. A., MISHRA, T., JAIN, D., BYRSKA-BISHOP, M., BLANKENBERG, D., LAJOIE, B. R., JAIN, G., SANYAL, A., CHEN, K. B., DENAS, O., TAYLOR, J., BLOBEL, G. A., WEISS, M. J., PIMKIN, M., DENG, W., MARINOV, G. K., WILLIAMS, B. A., FISHER-AYLOR, K. I., DESALVO, G., KIRALUSHA, A., TROUT, D., AMRHEIN, H., MORTAZAVI, A., EDSALL, L., MCCLEARY, D., KUAN, S., SHEN, Y., YUE, F., YE, Z., DAVIS, C. A., ZALESKI, C., JHA, S., XUE, C., DOBIN, A., LIN, W., FASTUCA, M., WANG, H., GUIGO, R., DJEBALI, S., LAGARDE, J., RYBA, T., SASAKI, T., MALLADI, V. S., CLINE, M. S., KIRKUP, V. M., LEARNED, K., ROSENBLOOM, K. R., KENT, W. J., FEINGOLD, E. A., GOOD, P. J., PAZIN, M., LOWDON, R. F. & ADAMS, L. B. 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biol,* 13**,** 418.

MUPPIRALA, U. K., HONAVAR, V. G. & DOBBS, D. 2011. Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics,* 12**,** 489.

NAVARRO, J. M., TOUZART, A., PRADEL, L. C., LOOSVELD, M., KOUBI, M., FENOUIL, R., LE NOIR, S., MAQBOOL, M. A., MORGADO, E., GREGOIRE, C., JAEGER, S., MAMESSIER, E., PIGNON, C., HACEIN-BEY-ABINA, S., MALISSEN, B., GUT, M., GUT, I. G., DOMBRET, H., MACINTYRE, E. A., HOWE, S. J., GASPAR, H. B., THRASHER, A. J., IFRAH, N., PAYET-BORNET, D., DUPREZ, E., ANDRAU, J. C., ASNAFI, V. & NADEL, B. 2015. Site- and allele-specific polycomb dysregulation in T-cell leukaemia. *Nat Commun,* 6**,** 6094.

NECSULEA, A., SOUMILLON, M., WARNEFORS, M., LIECHTI, A., DAISH, T., ZELLER, U., BAKER, J. C., GRUTZNER, F. & KAESSMANN, H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature,* 505**,** 635-40.

NEPH, S., VIERSTRA, J., STERGACHIS, A. B., REYNOLDS, A. P., HAUGEN, E., VERNOT, B., THURMAN, R. E., JOHN, S., SANDSTROM, R., JOHNSON, A. K., MAURANO, M. T., HUMBERT, R., RYNES, E., WANG, H., VONG, S., LEE, K., BATES, D., DIEGEL, M., ROACH, V., DUNN, D., NERI, J., SCHAFER, A., HANSEN, R. S., KUTYAVIN, T., GISTE, E., WEAVER, M., CANFIELD, T., SABO, P., ZHANG, M., BALASUNDARAM, G., BYRON, R., MACCOSS, M. J., AKEY, J. M., BENDER, M. A., GROUDINE, M., KAUL, R. & STAMATOYANNOPOULOS, J. A. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature,* 489**,** 83-90.

NIAZI, F. & VALADKHAN, S. 2012. Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3' UTRs. *RNA,* 18**,** 825-43.

NIRENBERG, M. & LEDER, P. 1964. Rna Codewords and Protein Synthesis. The Effect of Trinucleotides Upon the Binding of Srna to Ribosomes. *Science,* 145**,** 1399-407.

NOHATA, N., ABBA, M. C. & GUTKIND, J. S. 2016. Unraveling the oral cancer lncRNAome: Identification of novel lncRNAs associated with malignant progression and HPV infection. *Oral Oncol,* 59**,** 58-66.

NTZIACHRISTOS, P., TSIRIGOS, A., VAN VLIERBERGHE, P., NEDJIC, J., TRIMARCHI, T., FLAHERTY, M. S., FERRES-MARCO, D., DA ROS, V., TANG, Z., SIEGLE, J., ASP, P., HADLER, M., RIGO, I., DE KEERSMAECKER, K., PATEL, J., HUYNH, T., UTRO, F., POGLIO, S., SAMON, J. B., PAIETTA, E., RACEVSKIS, J., ROWE, J. M., RABADAN, R., LEVINE, R. L., BROWN, S., PFLUMIO, F., DOMINGUEZ, M., FERRANDO, A. & AIFANTIS, I. 2012. Genetic inactivation of the polycomb repressive complex 2 in T cell acute lymphoblastic leukemia. *Nat Med,* 18**,** 298-301.

O'BRIEN, K. 2017. TaqMan Low Density Array: MicroRNA Profiling for Biomarker and Oncosuppressor Discovery. *Methods Mol Biol,* 1509**,** 71-84.

O'LEARY, V. B., OVSEPIAN, S. V., CARRASCOSA, L. G., BUSKE, F. A., RADULOVIC, V., NIYAZI, M., MOERTL, S., TRAU, M., ATKINSON, M. J. & ANASTASOV, N. 2015. PARTICLE, a Triplex-Forming Long ncRNA, Regulates Locus-Specific Methylation in Response to Low-Dose Irradiation. *Cell Rep,* 11**,** 474-85.

OLINS, A. L. & OLINS, D. E. 1974. Spheroid chromatin units (v bodies). *Science,* 183**,** 330-2.

OROM, U. A., DERRIEN, T., BERINGER, M., GUMIREDDY, K., GARDINI, A., BUSSOTTI, G., LAI, F., ZYTNICKI, M., NOTREDAME, C., HUANG, Q., GUIGO, R. & SHIEKHATTAR, R. 2010. Long noncoding RNAs with enhancer-like function in human cells. *Cell,* 143**,** 46-58.

OROM, U. A. & SHIEKHATTAR, R. 2013. Long noncoding RNAs usher in a new era in the biology of enhancers. *Cell,* 154**,** 1190-3.

OVCHARENKO, I., LOOTS, G. G., GIARDINE, B. M., HOU, M., MA, J., HARDISON, R. C., STUBBS, L. & MILLER, W. 2005. Mulan: multiple-sequence local alignment and visualization for studying function and evolution. *Genome Res.,* 15**,** 184-194.

PANDEY, R. R., MONDAL, T., MOHAMMAD, F., ENROTH, S., REDRUP, L., KOMOROWSKI, J., NAGANO, T., MANCINI-DINARDO, D. & KANDURI, C. 2008. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol Cell,* 32**,** 232-46.

PANZITT, K., TSCHERNATSCH, M. M., GUELLY, C., MOUSTAFA, T., STRADNER, M., STROHMAIER, H. M., BUCK, C. R., DENK, H., SCHROEDER, R., TRAUNER, M. & ZATLOUKAL, K. 2007. Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA. *Gastroenterology,* 132**,** 330-42.

PARASKEVOPOULOU, M. D., GEORGAKILAS, G., KOSTOULAS, N., RECZKO, M., MARAGKAKIS, M., DALAMAGAS, T. M. & HATZIGEORGIOU, A. G. 2013. DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. *Nucleic Acids Res,* 41**,** D239-45.

PARK, P. J. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet,* 10**,** 669-80.

PASSEGUE, E., JAMIESON, C. H., AILLES, L. E. & WEISSMAN, I. L. 2003. Normal and leukemic hematopoiesis: are leukemias a stem cell disorder or a reacquisition of stem cell characteristics? *Proc Natl Acad Sci U S A,* 100 Suppl 1**,** 11842-9.

PATERSON, D. J., JEFFERIES, W. A., GREEN, J. R., BRANDON, M. R., CORTHESY, P., PUKLAVEC, M. & WILLIAMS, A. F. 1987. Antigens of activated rat T lymphocytes including a molecule of 50,000 Mr detected only on CD4 positive T blasts. *Mol Immunol,* 24**,** 1281-90.

PEKOWSKA, A., BENOUKRAF, T., FERRIER, P. & SPICUGLIA, S. 2010. A unique H3K4me2 profile marks tissue-specific gene regulation. *Genome Res,* 20**,** 1493-502.

PEKOWSKA, A., BENOUKRAF, T., ZACARIAS-CABEZA, J., BELHOCINE, M., KOCH, F., HOLOTA, H., IMBERT, J., ANDRAU, J. C., FERRIER, P. & SPICUGLIA, S. 2011. H3K4 tri-methylation provides an epigenetic signature of active enhancers. *EMBO J,* 30**,** 4198–4210.

PERLOT, T. & ALT, F. W. 2008. Cis-regulatory elements and epigenetic changes control genomic rearrangements of the IgH locus. *Adv Immunol,* 99**,** 1-32.

PETERLIN, B. M. & PRICE, D. H. 2006. Controlling the elongation phase of transcription with P-TEFb. *Mol Cell,* 23**,** 297-305.

POKHOLOK, D. K., HARBISON, C. T., LEVINE, S., COLE, M., HANNETT, N. M., LEE, T. I., BELL, G. W., WALKER, K., ROLFE, P. A., HERBOLSHEIMER, E., ZEITLINGER, J., LEWITTER, F., GIFFORD, D. K. & YOUNG, R. A. 2005. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell,* 122**,** 517-527.

POTT, S. & LIEB, J. D. 2015. What are super-enhancers? *Nat Genet,* 47**,** 8-12.

PREKER, P., NIELSEN, J., KAMMLER, S., LYKKE-ANDERSEN, S., CHRISTENSEN, M. S., MAPENDANO, C. K., SCHIERUP, M. H. & JENSEN, T. H. 2008. RNA exosome depletion reveals transcription upstream of active human promoters. *Science,* 322**,** 1851-4.

PROUDFOOT, N. 2000. Connecting transcription to messenger RNA processing. *Trends Biochem Sci,* 25**,** 290-3.

PROUDFOOT, N. 2004. New perspectives on connecting messenger RNA 3' end formation to transcription. *Curr Opin Cell Biol,* 16**,** 272-8.

PROUDFOOT, N. J., FURGER, A. & DYE, M. J. 2002. Integrating mRNA processing with transcription. *Cell,* 108**,** 501-12.

PUI, C. H. & EVANS, W. E. 2006. Treatment of acute lymphoblastic leukemia. *N Engl J Med,* 354**,** 166-78.

PUI, C. H., RELLING, M. V. & DOWNING, J. R. 2004. Acute lymphoblastic leukemia. *N Engl J Med,* 350**,** 1535-48.

RADA-IGLESIAS, A., BAJPAI, R., SWIGUT, T., BRUGMANN, S. A., FLYNN, R. A. & WYSOCKA, J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature,* 470**,** 279-83.

RAMON-MAIQUES, S., KUO, A. J., CARNEY, D., MATTHEWS, A. G., OETTINGER, M. A., GOZANI, O. & YANG, W. 2007. The plant homeodomain finger of RAG2 recognizes histone H3 methylated at both lysine-4 and arginine-2. *Proc Natl Acad Sci USA,* 104**,** 18993-18998.

RAMOS, A. D., DIAZ, A., NELLORE, A., DELGADO, R. N., PARK, K. Y., GONZALES-ROYBAL, G., OLDHAM, M. C., SONG, J. S. & LIM, D. A. 2013. Integration of genome-wide approaches identifies lncRNAs of adult neural stem cells and their progeny in vivo. *Cell Stem Cell,* 12**,** 616-28.

REIMAND, J., ARAK, T., ADLER, P., KOLBERG, L., REISBERG, S., PETERSON, H. & VILO, J. 2016. g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.*

RHEE, H. S. & PUGH, B. F. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell,* 147**,** 1408-19.

RINN, J. L., KERTESZ, M., WANG, J. K., SQUAZZO, S. L., XU, X., BRUGMANN, S. A., GOODNOUGH, L. H., HELMS, J. A., FARNHAM, P. J., SEGAL, E. & CHANG, H. Y. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell,* 129**,** 1311-23.

ROADMAP EPIGENOMICS, C., KUNDAJE, A., MEULEMAN, W., ERNST, J., BILENKY, M., YEN, A., HERAVI-MOUSSAVI, A., KHERADPOUR, P., ZHANG, Z., WANG, J., ZILLER, M. J., AMIN, V., WHITAKER, J. W., SCHULTZ, M. D., WARD, L. D., SARKAR, A., QUON, G., SANDSTROM, R. S., EATON, M. L., WU, Y. C., PFENNING, A. R., WANG, X., CLAUSSNITZER, M., LIU, Y., COARFA, C., HARRIS, R. A., SHORESH, N., EPSTEIN, C. B., GJONESKA, E., LEUNG, D., XIE, W., HAWKINS, R. D., LISTER, R., HONG, C., GASCARD, P., MUNGALL, A. J., MOORE, R., CHUAH, E., TAM, A., CANFIELD, T. K., HANSEN, R. S., KAUL, R., SABO, P. J., BANSAL, M. S., CARLES, A., DIXON, J. R., FARH, K. H., FEIZI, S., KARLIC, R., KIM, A. R., KULKARNI, A., LI, D., LOWDON, R., ELLIOTT, G., MERCER, T. R., NEPH, S. J., ONUCHIC, V., POLAK, P., RAJAGOPAL, N., RAY, P., SALLARI, R. C., SIEBENTHALL, K. T., SINNOTT-ARMSTRONG, N. A., STEVENS, M., THURMAN, R. E., WU, J., ZHANG, B., ZHOU, X., BEAUDET, A. E., BOYER, L. A., DE JAGER, P. L., FARNHAM, P. J., FISHER, S. J., HAUSSLER, D.,

JONES, S. J., LI, W., MARRA, M. A., MCMANUS, M. T., SUNYAEV, S., THOMSON, J. A., TLSTY, T. D., TSAI, L. H., WANG, W., WATERLAND, R. A., ZHANG, M. Q., CHADWICK, L. H., BERNSTEIN, B. E., COSTELLO, J. F., ECKER, J. R., HIRST, M., MEISSNER, A., MILOSAVLJEVIC, A., REN, B., STAMATOYANNOPOULOS, J. A., WANG, T. & KELLIS, M. 2015. Integrative analysis of 111 reference human epigenomes. *Nature*, 518**,** 317-30.

ROBINSON, P. J. & RHODES, D. 2006. Structure of the '30 nm' chromatin fibre: a key role for the linker histone. *Curr Opin Struct Biol*, 16**,** 336-43.

ROEDER, R. G. 1996. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem Sci*, 21**,** 327-35.

ROMANOV, S., MEDVEDEV, A., GAMBARIAN, M., POLTORATSKAYA, N., MOESER, M., MEDVEDEVA, L., GAMBARIAN, M., DIATCHENKO, L. & MAKAROV, S. 2008. Homogeneous reporter system enables quantitative functional assessment of multiple transcription factors. *Nat Methods*, 5**,** 253-60.

ROY, A. L. & SINGER, D. S. 2015. Core promoters in transcription: old problem, new insights. *Trends Biochem Sci*, 40**,** 165-71.

SAEED, S., QUINTIN, J., KERSTENS, H. H., RAO, N. A., AGHAJANIREFAH, A., MATARESE, F., CHENG, S. C., RATTER, J., BERENTSEN, K., VAN DER ENT, M. A., SHARIFI, N., JANSSEN-MEGENS, E. M., TER HUURNE, M., MANDOLI, A., VAN SCHAIK, T., NG, A., BURDEN, F., DOWNES, K., FRONTINI, M., KUMAR, V., GIAMARELLOS-BOURBOULIS, E. J., OUWEHAND, W. H., VAN DER MEER, J. W., JOOSTEN, L. A., WIJMENGA, C., MARTENS, J. H., XAVIER, R. J., LOGIE, C., NETEA, M. G. & STUNNENBERG, H. G. 2014. Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science*, 345**,** 1251086.

SALAMEH, A., LEE, A. K., CARDO-VILA, M., NUNES, D. N., EFSTATHIOU, E., STAQUICINI, F. I., DOBROFF, A. S., MARCHIO, S., NAVONE, N. M., HOSOYA, H., LAUER, R. C., WEN, S., SALMERON, C. C., HOANG, A., NEWSHAM, I., LIMA, L. A., CARRARO, D. M., OLIVIERO, S., KOLONIN, M. G., SIDMAN, R. L., DO, K. A., TRONCOSO, P., LOGOTHETIS, C. J., BRENTANI, R. R., CALIN, G. A., CAVENEE, W. K., DIAS-NETO, E., PASQUALINI, R. & ARAP, W. 2015. PRUNE2 is a human prostate cancer suppressor regulated by the intronic long noncoding RNA PCA3. *Proc Natl Acad Sci U S A*, 112**,** 8403-8.

SEBZDA, E., MARIATHASAN, S., OHTEKI, T., JONES, R., BACHMANN, M. F. & OHASHI, P. S. 1999. Selection of the T cell repertoire. *Annu.Rev.Immunol.*, 17**,** 829-874.

SHEIK MOHAMED, J., GAUGHWIN, P. M., LIM, B., ROBSON, P. & LIPOVICH, L. 2010. Conserved long noncoding RNAs transcriptionally regulated by Oct4 and Nanog modulate pluripotency in mouse embryonic stem cells. *RNA*, 16**,** 324-37.

SHI, X., SUN, M., LIU, H., YAO, Y. & SONG, Y. 2013. Long non-coding RNAs: a new frontier in the study of human diseases. *Cancer Lett*, 339**,** 159-66.

SHIMAZAKI, N., TSAI, A. G. & LIEBER, M. R. 2009. H3K4me3 stimulates the V(D)J RAG complex for both nicking and hairpinning in trans in addition to tethering in cis: implications for translocations. *Mol Cell*, 34**,** 535-544.

SHIN, S. Y., FAUMAN, E. B., PETERSEN, A. K., KRUMSIEK, J., SANTOS, R., HUANG, J., ARNOLD, M., ERTE, I., FORGETTA, V., YANG, T. P., WALTER, K., MENNI, C., CHEN, L., VASQUEZ, L., VALDES, A. M., HYDE, C. L., WANG, V., ZIEMEK, D., ROBERTS, P., XI, L., GRUNDBERG, E., MULTIPLE TISSUE HUMAN EXPRESSION RESOURCE, C., WALDENBERGER, M., RICHARDS, J. B., MOHNEY, R. P., MILBURN, M. V., JOHN, S. L., TRIMMER, J., THEIS, F. J., OVERINGTON, J. P., SUHRE, K., BROSNAN, M. J., GIEGER, C., KASTENMULLER, G., SPECTOR, T. D. & SORANZO, N. 2014. An atlas of genetic influences on human blood metabolites. *Nat Genet*, 46**,** 543-50.

SHLYUEVA, D., STAMPFEL, G. & STARK, A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet*, 15**,** 272-86.

SIGNAL, B., GLOSS, B. S. & DINGER, M. E. 2016. Computational Approaches for Functional Prediction and Characterisation of Long Noncoding RNAs. *Trends Genet*, 32**,** 620-37.

SIGOVA, A. A., MULLEN, A. C., MOLINIE, B., GUPTA, S., ORLANDO, D. A., GUENTHER, M. G., ALMADA, A. E., LIN, C., SHARP, P. A., GIALLOURAKIS, C. C. & YOUNG, R. A. 2013. Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proc Natl Acad Sci U S A*.

SILVERMAN, L. B., GELBER, R. D., DALTON, V. K., ASSELIN, B. L., BARR, R. D., CLAVELL, L. A., HURWITZ, C. A., MOGHRABI, A., SAMSON, Y., SCHORIN, M. A., ARKIN, S., DECLERCK, L., COHEN, H. J. & SALLAN, S. E. 2001. Improved outcome for children with acute lymphoblastic leukemia: results of Dana-Farber Consortium Protocol 91-01. *Blood*, 97**,** 1211-8.

SINGER, A. L., BUNNELL, S. C., OBSTFELD, A. E., JORDAN, M. S., WU, J. N., MYUNG, P. S., SAMELSON, L. E. & KORETZKY, G. A. 2004. Roles of the proline-rich domain in SLP-76 subcellular localization and T cell function. *J Biol Chem*, 279**,** 15481-90.

SMITH, M. A., GESELL, T., STADLER, P. F. & MATTICK, J. S. 2013. Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res*, 41**,** 8220-36.

SMITH, Z. D. & MEISSNER, A. 2013. DNA methylation: roles in mammalian development. *Nat Rev Genet*, 14**,** 204-20.

SOULIER, J., CLAPPIER, E., CAYUELA, J. M., REGNAULT, A., GARCIA-PEYDRO, M., DOMBRET, H., BARUCHEL, A., TORIBIO, M. L. & SIGAUX, F. 2005. HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). *Blood*, 106**,** 274-86.

SPENCER, V. A. & DAVIE, J. R. 1999. Role of covalent modifications of histones in regulating gene expression. *Gene*, 240**,** 1-12.

SPICUGLIA, S. & VANHILLE, L. 2012. Chromatin signatures of active enhancers. *Nucleus*, 3**,** 126-31.

SPICUGLIA, S., ZACARIAS-CABEZA, J., PEKOWSKA, P. & FERRIER, P. 2010. Epigenetic regulation of antigen receptor gene rearrangement. *F1000 Biol Rep*, 2.

SPITS, H. 2002. Development of alphabeta T cells in the human thymus. *Nat Rev Immunol*, 2**,** 760-72.

SRIVASTAVA, D., THOMAS, T., LIN, Q., KIRBY, M. L., BROWN, D. & OLSON, E. N. 1997. Regulation of cardiac mesodermal and neural crest development by the bHLH transcription factor, dHAND. *Nat Genet*, 16**,** 154-60.

ST LAURENT, G., SHTOKALO, D., TACKETT, M. R., YANG, Z., EREMINA, T., WAHLESTEDT, C., URCUQUI-INCHIMA, S., SEILHEIMER, B., MCCAFFREY, T. A. & KAPRANOV, P. 2012. Intronic RNAs constitute the major fraction of the non-coding RNA in mammalian cells. *BMC Genomics*, 13**,** 504.

STUART, J. M., SEGAL, E., KOLLER, D. & KIM, S. K. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302**,** 249-55.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. & MESIROV, J. P. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102**,** 15545-50.

SUN, J., SHI, H., WANG, Z., ZHANG, C., LIU, L., WANG, L., HE, W., HAO, D., LIU, S. & ZHOU, M. 2014. Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. *Mol Biosyst*, 10**,** 2074-81.

TAGHON, T., YUI, M. A., PANT, R., DIAMOND, R. A. & ROTHENBERG, E. V. 2006. Developmental and molecular characterization of emerging beta- and gammadelta-selected pre-T cells in the adult mouse thymus. *Immunity.*, 24**,** 53-64.

THOMAS-CHOLLIER, M., SAND, O., TURATSINZE, J. V., JANKY, R., DEFRANCE, M., VERVISCH, E., BROHEE, S. & VAN HELDEN, J. 2008. RSAT: regulatory sequence analysis tools. *Nucleic Acids Res*, 36**,** W119-27.

THOMAS, L. R., COBB, R. M. & OLTZ, E. M. 2009. Dynamic regulation of antigen receptor gene assembly. *Adv Exp Med Biol,* 650**,** 103-15.

THURMAN, R. E., RYNES, E., HUMBERT, R., VIERSTRA, J., MAURANO, M. T., HAUGEN, E., SHEFFIELD, N. C., STERGACHIS, A. B., WANG, H., VERNOT, B., GARG, K., JOHN, S., SANDSTROM, R., BATES, D., BOATMAN, L., CANFIELD, T. K., DIEGEL, M., DUNN, D., EBERSOL, A. K., FRUM, T., GISTE, E., JOHNSON, A. K., JOHNSON, E. M., KUTYAVIN, T., LAJOIE, B., LEE, B. K., LEE, K., LONDON, D., LOTAKIS, D., NEPH, S., NERI, F., NGUYEN, E. D., QU, H., REYNOLDS, A. P., ROACH, V., SAFI, A., SANCHEZ, M. E., SANYAL, A., SHAFER, A., SIMON, J. M., SONG, L., VONG, S., WEAVER, M., YAN, Y., ZHANG, Z., ZHANG, Z., LENHARD, B., TEWARI, M., DORSCHNER, M. O., HANSEN, R. S., NAVAS, P. A., STAMATOYANNOPOULOS, G., IYER, V. R., LIEB, J. D., SUNYAEV, S. R., AKEY, J. M., SABO, P. J., KAUL, R., FUREY, T. S., DEKKER, J., CRAWFORD, G. E. & STAMATOYANNOPOULOS, J. A. 2012. The accessible chromatin landscape of the human genome. *Nature,* 489**,** 75-82.

TOMAZOU, E. M., SHEFFIELD, N. C., SCHMIDL, C., SCHUSTER, M., SCHONEGGER, A., DATLINGER, P., KUBICEK, S., BOCK, C. & KOVAR, H. 2015. Epigenome mapping reveals distinct modes of gene regulation and widespread enhancer reprogramming by the oncogenic fusion protein EWS-FLI1. *Cell Rep,* 10**,** 1082-95.

TRAPNELL, C., PACHTER, L. & SALZBERG, S. L. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics,* 25**,** 1105-11.

TRAPNELL, C., ROBERTS, A., GOFF, L., PERTEA, G., KIM, D., KELLEY, D. R., PIMENTEL, H., SALZBERG, S. L., RINN, J. L. & PACHTER, L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc,* 7**,** 562-78.

TRAPNELL, C., WILLIAMS, B. A., PERTEA, G., MORTAZAVI, A., KWAN, G., VAN BAREN, M. J., SALZBERG, S. L., WOLD, B. J. & PACHTER, L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol,* 28**,** 511-5.

TRIMARCHI, T., BILAL, E., NTZIACHRISTOS, P., FABBRI, G., DALLA-FAVERA, R., TSIRIGOS, A. & AIFANTIS, I. 2014. Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia. *Cell,* 158**,** 593-606.

TSAI, M. C., MANOR, O., WAN, Y., MOSAMMAPARAST, N., WANG, J. K., LAN, F., SHI, Y., SEGAL, E. & CHANG, H. Y. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science,* 329**,** 689-93.

TSUKADA, Y., FANG, J., ERDJUMENT-BROMAGE, H., WARREN, M. E., BORCHERS, C. H., TEMPST, P. & ZHANG, Y. 2006. Histone demethylation by a family of JmjC domain-containing proteins. *Nature,* 439**,** 811-6.

ULITSKY, I. & BARTEL, D. P. 2013. lincRNAs: genomics, evolution, and mechanisms. *Cell,* 154**,** 26-46.

VAN VLIERBERGHE, P. & FERRANDO, A. 2012. The molecular basis of T cell acute lymphoblastic leukemia. *J Clin Invest,* 122**,** 3398-406.

VAN VLIERBERGHE, P., PIETERS, R., BEVERLOO, H. B. & MEIJERINK, J. P. 2008. Molecular-genetic insights in paediatric T-cell acute lymphoblastic leukaemia. *Br J Haematol,* 143**,** 153-68.

VELCULESCU, V. E., MADDEN, S. L., ZHANG, L., LASH, A. E., YU, J., RAGO, C., LAL, A., WANG, C. J., BEAUDRY, G. A., CIRIELLO, K. M., COOK, B. P., DUFAULT, M. R., FERGUSON, A. T., GAO, Y., HE, T. C., HERMEKING, H., HIRALDO, S. K., HWANG, P. M., LOPEZ, M. A., LUDERER, H. F., MATHEWS, B., PETROZIELLO, J. M., POLYAK, K., ZAWEL, L., KINZLER, K. W. & ET AL. 1999. Analysis of human transcriptomes. *Nat Genet,* 23**,** 387-8.

VICKARYOUS, M. K. & HALL, B. K. 2006. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol Rev Camb Philos Soc,* 81**,** 425-55.

VISEL, A., RUBIN, E. M. & PENNACCHIO, L. A. 2009. Genomic views of distant-acting enhancers. *Nature,* 461**,** 199-205.

VOLDERS, P. J., HELSENS, K., WANG, X., MENTEN, B., MARTENS, L., GEVAERT, K., VANDESOMPELE, J. & MESTDAGH, P. 2012. LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.*

WANG, K. C., YANG, Y. W., LIU, B., SANYAL, A., CORCES-ZIMMERMAN, R., CHEN, Y., LAJOIE, B. R., PROTACIO, A., FLYNN, R. A., GUPTA, R. A., WYSOCKA, J., LEI, M., DEKKER, J., HELMS, J. A. & CHANG, H. Y. 2011. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature,* 472**,** 120-4.

WANG, Y., ZHANG, W., JIN, Y., JOHANSEN, J. & JOHANSEN, K. M. 2001. The JIL-1 tandem kinase mediates histone H3 phosphorylation and is required for maintenance of chromatin structure in Drosophila *Cell,* 105**,** 433-443.

WEST, K. L., SINGHA, N. C., DE IOANNES, P., LACOMIS, L., ERDJUMENT-BROMAGE, H., TEMPST, P. & CORTES, P. 2005. A Direct Interaction between the RAG2 C Terminus and the Core Histones Is Required for Efficient V(D)J Recombination. *Immunity,* 23**,** 203-212.

WHETSTINE, J. R., NOTTKE, A., LAN, F., HUARTE, M., SMOLIKOV, S., CHEN, Z., SPOONER, E., LI, E., ZHANG, G., COLAIACOVO, M. & SHI, Y. 2006. Reversal of histone lysine trimethylation by the JMJD2 family of histone demethylases. *Cell,* 125**,** 467-81.

WHYTE, W. A., ORLANDO, D. A., HNISZ, D., ABRAHAM, B. J., LIN, C. Y., KAGEY, M. H., RAHL, P. B., LEE, T. I. & YOUNG, R. A. 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell,* 153**,** 307-19.

WINGENDER, E. 2013. Criteria for an updated classification of human transcription factor DNA-binding domains. *J Bioinform Comput Biol,* 11**,** 1340007.

WOLFFE, A. P. 1998. Packaging principle: how DNA methylation and histone acetylation control the transcriptional activity of chromatin. *J Exp Zool,* 282**,** 239-44.

WONGSURAWAT, T., JENJAROENPUN, P., KWOH, C. K. & KUZNETSOV, V. 2012. Quantitative model of R-loop forming structures reveals a novel level of RNA-DNA interactome complexity. *Nucleic Acids Res,* 40**,** e16.

WOODCOCK, C. L. 2005. A milestone in the odyssey of higher-order chromatin structure. *Nat Struct Mol Biol,* 12**,** 639-40.

WU, J., HUANG, B., CHEN, H., YIN, Q., LIU, Y., XIANG, Y., ZHANG, B., LIU, B., WANG, Q., XIA, W., LI, W., LI, Y., MA, J., PENG, X., ZHENG, H., MING, J., ZHANG, W., ZHANG, J., TIAN, G., XU, F., CHANG, Z., NA, J., YANG, X. & XIE, W. 2016. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature,* 534**,** 652-7.

XIAO, Y., LV, Y., ZHAO, H., GONG, Y., HU, J., LI, F., XU, J., BAI, J., YU, F. & LI, X. 2015. Predicting the functions of long noncoding RNAs using RNA-seq based on Bayesian network. *Biomed Res Int,* 2015**,** 839590.

XIE, W., SCHULTZ, M. D., LISTER, R., HOU, Z., RAJAGOPAL, N., RAY, P., WHITAKER, J. W., TIAN, S., HAWKINS, R. D., LEUNG, D., YANG, H., WANG, T., LEE, A. Y., SWANSON, S. A., ZHANG, J., ZHU, Y., KIM, A., NERY, J. R., URICH, M. A., KUAN, S., YEN, C. A., KLUGMAN, S., YU, P., SUKNUNTHA, K., PROPSON, N. E., CHEN, H., EDSALL, L. E., WAGNER, U., LI, Y., YE, Z., KULKARNI, A., XUAN, Z., CHUNG, W. Y., CHI, N. C., ANTOSIEWICZ-BOURGET, J. E., SLUKVIN, I., STEWART, R., ZHANG, M. Q., WANG, W., THOMSON, J. A., ECKER, J. R. & REN, B. 2013. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell,* 153**,** 1134-48.

XU, C. R. & FEENEY, A. J. 2009. The epigenetic profile of Ig genes is dynamically regulated during B cell differentiation and is modulated by pre-B cell receptor signaling. *J Immunol,* 182**,** 1362-9.

YAN, W., YOUNG, A. Z., SOARES, V. C., KELLEY, R., BENEZRA, R. & ZHUANG, Y. 1997. High incidence of T-cell tumors in E2A-null mice and E2A/Id1 double- knockout mice. *Mol Cell Biol,* 17**,** 7317-7327.

YANG, J. H., LI, J. H., JIANG, S., ZHOU, H. & QU, L. H. 2013. ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. *Nucleic Acids Res,* 41**,** D177-87.

YANG, L., FROBERG, J. E. & LEE, J. T. 2014a. Long noncoding RNAs: fresh perspectives into the RNA world. *Trends Biochem Sci,* 39**,** 35-43.

YANG, X., GAO, L., GUO, X., SHI, X., WU, H., SONG, F. & WANG, B. 2014b. A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. *PLoS One,* 9**,** e87797.

YAO, P., LIN, P., GOKOOLPARSADH, A., ASSAREH, A., THANG, M. W. & VOINEAGU, I. 2015. Coexpression networks identify brain region-specific enhancer RNAs in the human brain. *Nat Neurosci,* 18**,** 1168-74.

YU, W., GIUS, D., ONYANGO, P., MULDOON-JACOBS, K., KARP, J., FEINBERG, A. P. & CUI, H. 2008. Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. *Nature,* 451**,** 202-6.

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. & SPICUGLIA, S. 2015. Transcription-Dependent Generation of a Specialized Chromatin Structure at the TCRbeta Locus. *J Immunol,* 194**,** 3432-43.

ZAWEL, L. & REINBERG, D. 1995. Common themes in assembly and function of eukaryotic transcription complexes. *Annu Rev Biochem,* 64**,** 533-61.

ZEITLINGER, J., STARK, A., KELLIS, M., HONG, J. W., NECHAEV, S., ADELMAN, K., LEVINE, M. & YOUNG, R. A. 2007. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet,* 39**,** 1512-6.

ZHANG, B., ZHENG, H., HUANG, B., LI, W., XIANG, Y., PENG, X., MING, J., WU, X., ZHANG, Y., XU, Q., LIU, W., KOU, X., ZHAO, Y., HE, W., LI, C., CHEN, B., LI, Y., WANG, Q., MA, J., YIN, Q., KEE, K., MENG, A., GAO, S., XU, F., NA, J. & XIE, W. 2016. Allelic reprogramming of the histone modification H3K4me3 in early mammalian development. *Nature,* 537**,** 553-557.

ZHANG, K., SHI, Z. M., CHANG, Y. N., HU, Z. M., QI, H. X. & HONG, W. 2014. The ways of action of long non-coding RNAs in cytoplasm and nucleus. *Gene,* 547**,** 1-9.

ZHANG, Y., LIU, T., MEYER, C. A., EECKHOUTE, J., JOHNSON, D. S., BERNSTEIN, B. E., NUSBAUM, C., MYERS, R. M., BROWN, M., LI, W. & LIU, X. S. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol,* 9**,** R137.

ZHOU, M., WANG, X., LI, J., HAO, D., WANG, Z., SHI, H., HAN, L., ZHOU, H. & SUN, J. 2015. Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network. *Mol Biosyst,* 11**,** 760-9.

# Notes

# Annexe I : Technique *TaqMan Low Density Array* (TLDA)

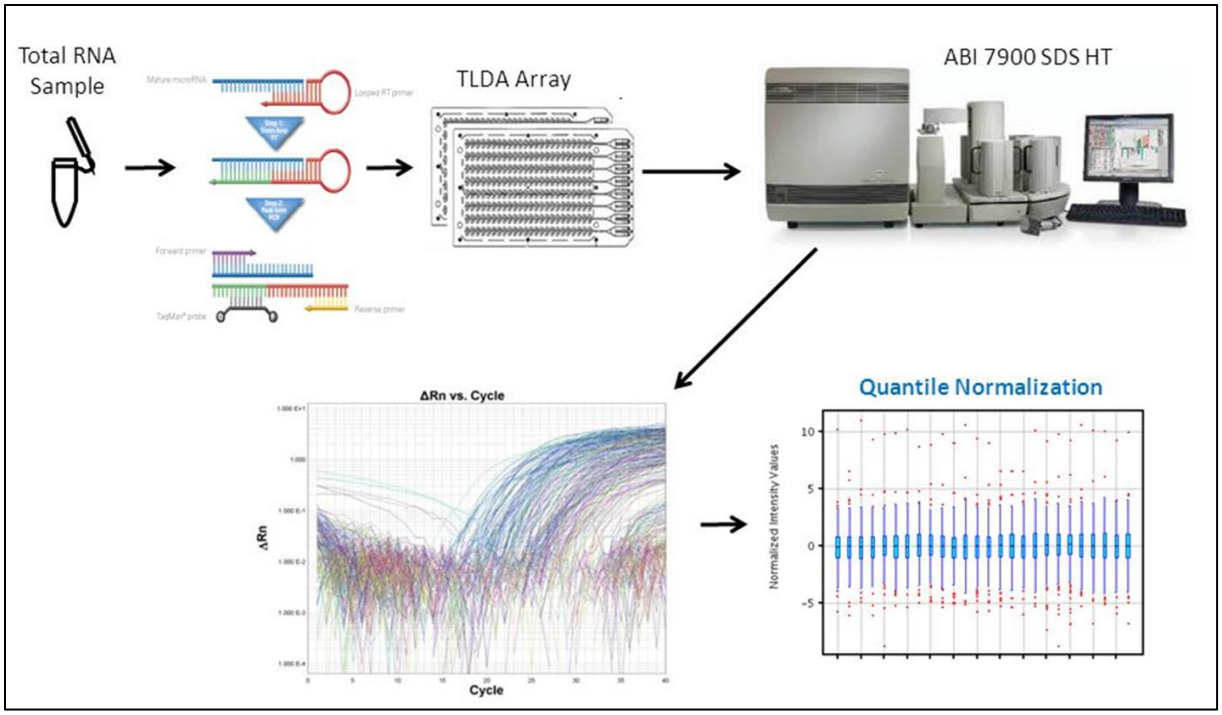



**Fig 56 : Les différents étapes de la technologie *TaqMan Low Density Array* .**

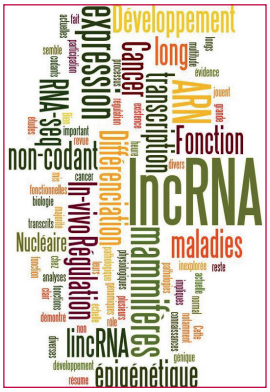
# Annexe II : Rôle des longs ARN non codants dans le développement normal et pathologique. (Review)

médecine/sciences 2014 ; 30 : 790-6

# Rôle des longs ARN non codants dans le développement normal et pathologique

Ève-Lyne Mathieu[1,2], Mohamed Belhocine[1,2], Lan T.M. Dao[1,2], Denis Puthier[1-3], Salvatore Spicuglia[1,2]

> Des analyses génomiques à grande échelle ont mis en évidence l'existence d'une multitude de longs ARN non codants (lncRNA). Bien que la fonction de la majorité de ces lncRNA reste à l'heure actuelle inexplorée, il semble clair que plusieurs de ces transcrits jouent un rôle important dans la régulation de l'expression génique et sont impliqués dans diverses pathologies, notamment le cancer. De fait, des études fonctionnelles ont démontré la participation des lncRNA à divers processus physiologiques. Cette revue résume les connaissances actuelles sur la biologie des lncRNA et sur leurs fonctions dans le développement normal et pathologique chez les mammifères. <




[1]Inserm UMR1090 TAGC (*technological advances for genomics and clinics*), 13288 Marseille, France ; [2]Université Aix-Marseille, UMR1090 TAGC, 13288 Marseille, France ; [3]Polytech Marseille, parc scientifique de Luminy, Marseille, France.
salvatore.spicuglia@inserm.fr

L'avènement des analyses génomiques à grande échelle a profondément modifié notre compréhension de l'organisation fonctionnelle du génome. En particulier, des études récentes ont mis en évidence qu'au moins 70 % du génome humain peut être transcrit [1]. De ce fait, il est maintenant admis que la complexité du transcriptome s'étend bien au-delà des transcrits codant pour des protéines, puisqu'elle comprend aussi une myriade d'ARN non codants ayant des profils d'expression et de régulation complexes. Ces ARN non codants comprennent, d'une part, des types d'ARN régulateurs dont les fonctions et les mécanismes d'action impliqués sont relativement bien définis comme les miARN (micro-ARN) et les piARN (*piwi-interacting RNA*) et, d'autre part, des classes d'ARN moins bien caractérisées (comme les longs ARN non codants, lncRNA). Bien que la fonction de la grande majorité des lncRNA reste, à l'heure actuelle, largement inexplorée, il semble que ces transcrits jouent un rôle important dans plusieurs processus physiologiques, tels que la régulation épigénétique de l'expression génique ou la maturation des ARN messagers [2-6]. De plus, plusieurs travaux récents convergent pour indiquer que certains d'entre eux seraient impliqués dans le développement de pathologies humaines. Dans cette synthèse, nous nous proposons de passer en revue la littérature récente traitant des lncRNA et de leurs fonctions dans le développement normal et pathologique chez les mammifères.

## Une panoplie d'ARN longs non codants

Les lncRNA sont des transcrits de plus de 200 nucléotides qui ne codent pas pour des protéines. Bien que cette définition soit en apparence arbitraire, elle permet de distinguer les lncRNA des petits ARN régulateurs, tels que les miARN, les piARN et les autres petits ARN nucléaires. Les lncRNA, transcrits généralement par l'ARN polymérase II, constituent un groupe de taille très hétérogène, certains pouvant s'étendre sur plusieurs dizaines de kilobases (kb). Les gènes de lncRNA partagent plusieurs caractéristiques avec les gènes codant pour des protéines, comme des profils épigénétiques similaires, la présence de signaux d'épissage et de polyadénylation, ainsi que la taille des exons et des introns [3]. Toutefois, et comparativement aux ARNm, les lncRNA sont généralement plus enrichis dans le noyau et montrent une plus faible conservation de séquence, bien que certains d'entre eux soient fortement conservés [3]. De plus, les gènes de lncRNA s'expriment plus faiblement que les gènes codants et leur expression est remarquablement spécifique de certains tissus.

En fonction de leur position par rapport aux gènes codants, les gènes de lncRNA peuvent être divisés en deux grandes catégories : les lncRNA intergéniques et les lncRNA intragéniques *(Figure 1)*. Les


m/s n° 8-9, vol. 30, août-septembre 2014
DOI : 10.1051/medsci/20143008018

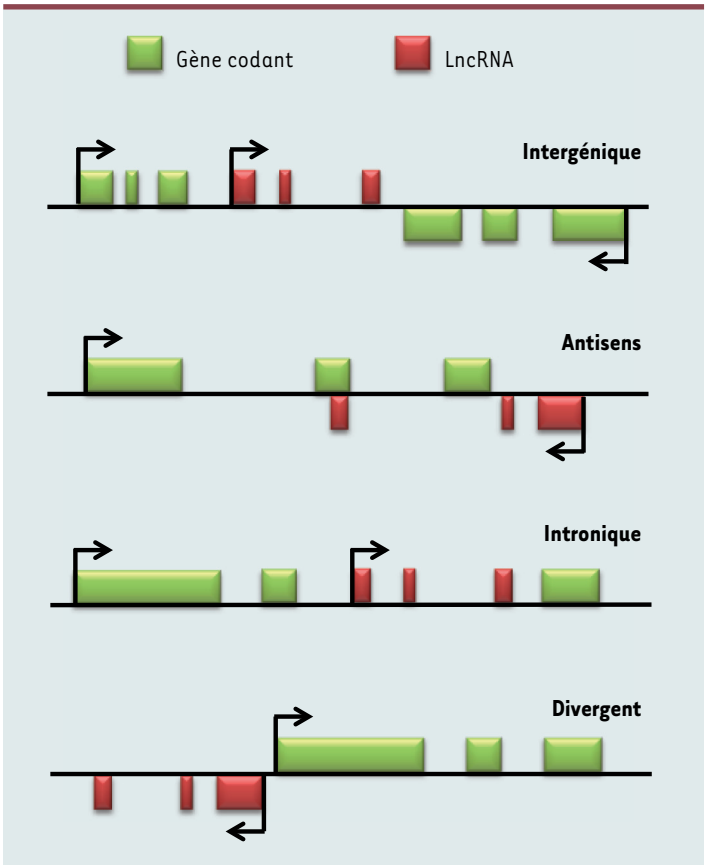



***Figure 1.*** ***Représentation schématique des principales catégories de lncRNA.*** Les gènes de lncRNA sont définis par leur position relative aux gènes codants qui sont situés à leur proximité. Les gènes *lncRNA* intergéniques (lincRNA) sont des lncRNA localisés dans une région non annotée du génome. Les gènes *lncRNA* antisens sont des lncRNA transcrits dans la direction opposée d'un gène codant et dont la séquence chevauche en partie ou totalement le gène codant lui étant associé. Les lncRNA introniques sont des lncRNA contenus dans l'intron d'un gène codant. Les lncRNA divergents sont transcrits de façon divergente au promoteur d'un gène codant. Les exons de gènes sont représentés par des boîtes ; l'orientation et le site d'initiation de la transcription sont représentés par des flèches. Les gènes codants sont représentés en vert, les lncRNA sont en rouge.

lncRNA intergéniques, localisés par définition dans les régions non annotées du génome, sont généralement dénommés lincRNA. Ils représentent, à l'heure actuelle, la classe de lncRNA la mieux étudiée [3]. Les lncRNA intragéniques, quant à eux, peuvent être subdivisés en fonction de la façon dont ils chevauchent les gènes codants ou de leur orientation par rapport à ceux-ci (antisens, intronique, etc.). Il est à noter que de nombreux gènes de lncRNA intergéniques ont un site d'initiation de la transcription proche de celui d'un gène codant, la transcription se faisant sur le brin opposé (transcription divergente). Nous avons récemment démontré que les gènes associés à ces transcrits divergents codent fréquemment pour des régulateurs transcriptionnels impliqués dans le développement et la différenciation cellulaire [7]. Enfin, certains lncRNA chevauchent de petits ARN, tels que des petits ARN nucléaires (snoRNA) [46] ou des miARN, avec des liens fonctionnels potentiels, comme dans le cas de régions soumises à l'empreinte génomique [6]. De multiples lncRNA contiennent des éléments répétés, comme des LINE (*long interspersed nuclear elements*) ou des SINE (*short interspersed nuclear elements*), avec des implications fonctionnelles potentielles [8, 9]. Cependant, pris dans leur ensemble, les lncRNA n'ont pas de séquence ou de structure conservée qui pourrait être indicative d'une fonction particulière [3]. De ce fait, la plupart des études visant à identifier des lncRNA potentiellement pertinents dans un contexte physiologique ou pathologique donné sont basées sur des analyses de coexpression ou de corégulation. On transpose alors la fonction des gènes voisins ayant un profil similaire aux lncRNA dont la fonction est à définir [10]. Ces approches sont aussi connues sous le terme « culpabilité par association » *(Figure 2)*.

## L'étude à grande échelle des lncRNA

Au cours des deux dernières décennies, le séquençage systématique des banques d'ADNc a permis l'identification de divers transcrits, dont un certain nombre sont aujourd'hui considérés comme des lncRNA. De même, des études pionnières basées sur des puces à ADN ou sur l'analyse de profils épigénétiques ont révélé la présence de régions transcrites produisant un certain nombre de lncRNA intergéniques (lincRNA) [3]. En particulier, des analyses basées sur la combinaison de certaines marques de chromatine ont permis les premières caractérisations à haut débit des lincRNA (voir par exemple [11]). Il a toutefois fallu attendre le développement généralisé du séquençage à haut débit de l'ARN (*RNA-sequencing*, RNA-seq), pour que des analyses systématiques du transcriptome révèlent l'existence de lncRNA de manière non ambiguë et permettent d'apprécier leur complexité chez les mammifères [12, 13]. Cependant, étant donné que la plupart des lncRNA sont faiblement exprimés et/ou ont une structure exon/intron très complexe, il est parfois difficile d'identifier les différents transcrits produits à partir d'un gène lncRNA. Il est alors nécessaire de combiner une approche de RNA-seq classique avec d'autres techniques, telles que des analyses épigénétiques (certaines marques indiquent la présence de la région promotrice ou s'étendent sur l'ensemble du gène), du séquençage ciblé des transcrits d'intérêt après leur capture sur une puce d'ADN, et des analyses d'identification précise des extrémités 5' ou 3' des transcrits, de façon à pouvoir déterminer sans ambiguïté la structure des différents variants découlant de la transcription d'un gène lncRNA [12].

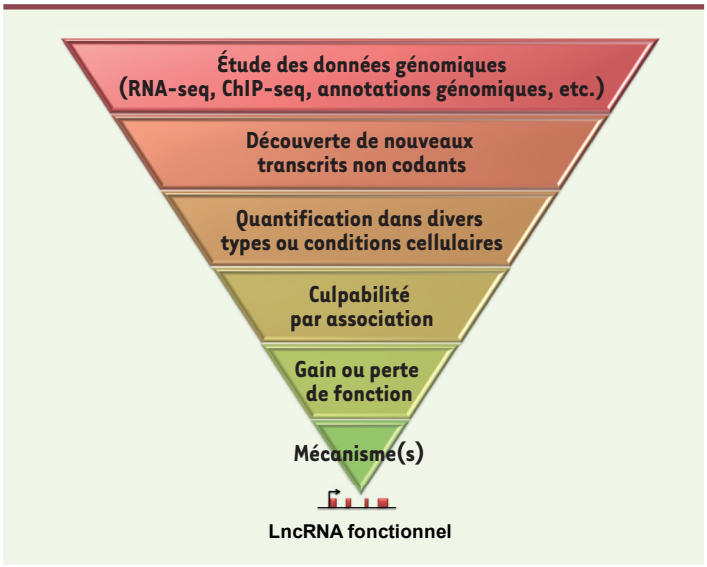



***Figure 2.*** ***Approches expérimentales menant à la caractérisation fonctionnelle des lncRNA.*** Les méthodes actuelles combinent des approches expérimentales et bio-informatiques afin de construire un pipeline robuste pour la découverte, l'annotation et la caractérisation fonctionnelle des lncRNA. RNA-seq : *RNA-sequencing* ; ChIP-seq : *chromatin immunoprecipitation-sequencing*.

Ces diverses approches, couplées à des analyses bio-informatiques poussées, ont permis l'identification d'un très grand nombre de lncRNA exprimés dans une variété de tissus et de lignées cellulaires. Notamment, le projet ENCODE (*encyclopedia of DNA elements*), à travers la base de donnée GENCODE (http://www.gencodegenes.org), maintient à jour la liste la plus exhaustive de lncRNA exprimés chez l'homme [14]. Actuellement, la version V19 de GENCODE compte 13 870 lncRNA détectés chez l'humain, dont 7 114 correspondent à des lincRNA. De plus, les lncRNA connus sont regroupés dans plusieurs catalogues publics, tels que « lncrnadb » (http://www.lncrnadb.org) et « noncode » (http://www.noncode.org), lesquels comportent une description détaillée de leurs structures génomiques.

## Fonctions des lncRNA dans les processus physiologiques

Comme mentionné précédemment, contrairement aux protéines qui possèdent souvent des domaines fonctionnels bien définis, il est pour l'instant impossible de prédire la fonction des lncRNA à partir de leur séquence. Toutefois, il apparaît que les lncRNA agissent essentiellement en modulant l'expression des gènes [15] *(Tableau I)*. Cette fonction peut s'exercer localement, lorsque les lncRNA agissent en *cis* sur les gènes voisins, ou de façon distale, lorsque leurs fonctions se produisent indépendamment de la localisation des gènes cibles *(Figure 3)*. En particulier, il existe une classe de lncRNA ayant une activité de type « *enhancer* » et qui peuvent activer de façon transcriptionnelle des gènes voisins [16]. Plus généralement, des études de gain de fonction et de perte de fonction ont permis de mettre en évidence que les lncRNA sont potentiellement impliqués dans divers processus biologiques chez les mammifères [2, 15]. Ces processus incluent, par exemple, le maintien de la pluripotence des cellules souches embryonnaires, la différenciation cellulaire, la régulation du cycle cellulaire et la réponse immunitaire. Les lncRNA peuvent réguler l'expression des gènes par divers mécanismes. Les aspects moléculaires de ces mécanismes ont été détaillés dans des revues récentes [2, 4, 15]. Les lncRNA peuvent potentiellement lier l'ADN, des protéines ou d'autres ARN, formant des plateformes d'échafaudage et permettant ainsi l'interaction entre diverses molécules fonctionnelles *(Figure 4)*. Certains lncRNA peuvent modifier le contexte chromatinien près de leurs gènes cibles, en recrutant des facteurs de transcription ou des acteurs de la modification d'histones ou de la réorganisation de la chromatine, favorisant ou réprimant ainsi la transcription de gènes cibles, selon le

| Noms | Fonctions | Mécanismes | Références |
|---|---|---|---|
| *Linc-MD1* | Contrôle temporel du programme de différenciation musculaire | Séquestration des miR-133 et miR-135 | [39] |
| *HOTAIR* | Répression transcriptionnelle des gènes *HOXD* | Recrutement des complexes répresseurs PRC2 et LSD1 | [18, 19] |
| *HOTTIP* et *Mistral* | Activation transcriptionnelle de gènes *HOXA* | Recrutement du complexe activateur WDR5/MLL | [20, 21] |
| *UCHL1-as* | Contrôle de la traduction de l'ARNm du gène *UCHL1* | Ciblage de l'ARNm vers le ribosome | [22] |
| *XIST* | Répression mono-allélique du chromosome X chez les mammifères femelles | En partie *via* le recrutement des complexes répresseurs | [17] |

***Tableau I.*** ***Exemples de lncRNA impliqués dans des processus biologiques physiologiques.*** *Linc-MD1 : lincRNA muscle differentiation 1.*

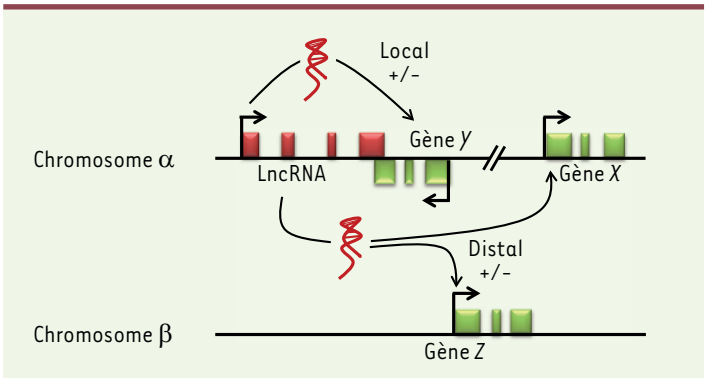



***Figure 3.*** ***Régulation de l'expression génique par les lncRNA.*** Les lncRNA peuvent contrôler l'expression des gènes voisins (action locale) ou indépendamment de la localisation de leurs gènes cibles (action distale). Ce contrôle peut avoir un impact positif ou négatif (+/-) sur l'expression des gènes cibles.

contexte. Parmi les lncRNA qui ont été fonctionnellement caractérisés se trouve *XIST* (*X-inactive specific transcript*), un transcrit de 19-kb, dont le gène est localisé sur le chromosome X [17]. *XIST* est directement impliqué dans l'inactivation du chromosome X chez les mammifères femelles. Une fois transcrit, *XIST* est retenu dans le noyau et recouvre le chromosome X inactif. De plus, il interagirait avec le complexe répresseur Polycomb 2 (PRC2), permettant le recrutement ciblé de ce complexe et contribuant ainsi au maintien de l'inactivation du chromosome X [17]. De façon intéressante, *XIST* est à son tour régulé par d'autres lncRNA, tels que *TSIX* et *XITE* (*X-inactivation intergenic transcription element*) [17]. D'autres exemples de lncRNA localisés dans des régions génomiques soumises à l'empreinte parentale, comme *AIRN* (*antisens of IGF2R RNA non-protein coding*), *H19* ou *KCNQ1OT1*, sont également impliqués dans l'inactivation de l'expression génique, *via* leurs associations avec des complexes répresseurs associés à la chromatine [6]. De même, le lncRNA *HOTAIR* (*Hox antisense intergenic RNA*), dont le gène est localisé dans le locus *HOXC*, servirait d'échafaudage pour les complexes PRC2 et LSD1 (*lysine specific demethylase 1*), deux complexes reliés à la répression transcriptionnelle [18, 19], et favoriserait leur recrutement au sein du locus *HOXD*. Au contraire, les lncRNA *Mistral* et *HOTTIP* (*HOXA transcript at the distal Tip*) faciliteraient l'expression des gènes *HOXA* en recrutant le complexe épigénétique activateur WDR5 (*WD repeat-containing protein 5*)/MLL (*mixed lineage leukemia*) [20, 21].

Les lncRNA sont aussi largement impliqués dans les processus post-transcriptionnels liés à la biogenèse des ARNm, tels que l'épissage, le transport, la traduction et la dégradation des ARNm (*Figure 4*). Par exemple, *UCHL1-as* (*ubiquitin carboxyl-terminal hydrolase isozyme L1-antisens*), un lncRNA antisens qui chevauche partiellement la partie 5' du gène *UCHL1*, contribue positivement à la traduction de l'ARNm du gène *UCHL1* [22]. Par ailleurs, des lncRNA peuvent agir comme des « éponges » pour prévenir la liaison de miARN à leurs ARNm cibles. Un cas extrême est représenté par *CDR1-as/ciRS-7* (*circular RNA sponge for miR-7*), un lncRNA circulaire exprimé chez l'humain, qui possède 70 sites de liaison pour *miR-7* [23].

Bien que le mécanisme d'action de nombreux lncRNA soit maintenant établi, il reste néanmoins possible que l'expression de certains lncRNA soit plutôt le reflet d'une structure particulière de la chromatine ou d'un sous-produit d'un mécanisme impliquant le processus transcriptionnel. Cela est le cas de l'expression des transcrits germinaux produits au cours du processus de réarrangement somatique des gènes codant les récepteurs pour l'antigène [24]. De plus, certains ARN non codants, appelés *enhancer* ARN (eARN), sont produits à partir d'éléments *cis*-régulateurs distaux [16]. À l'heure actuelle, le rôle de ces eARN sur l'activité transcriptionnelle du gène cible n'est pas encore clairement définie, puisque ceux-ci pourraient également être tout simplement des

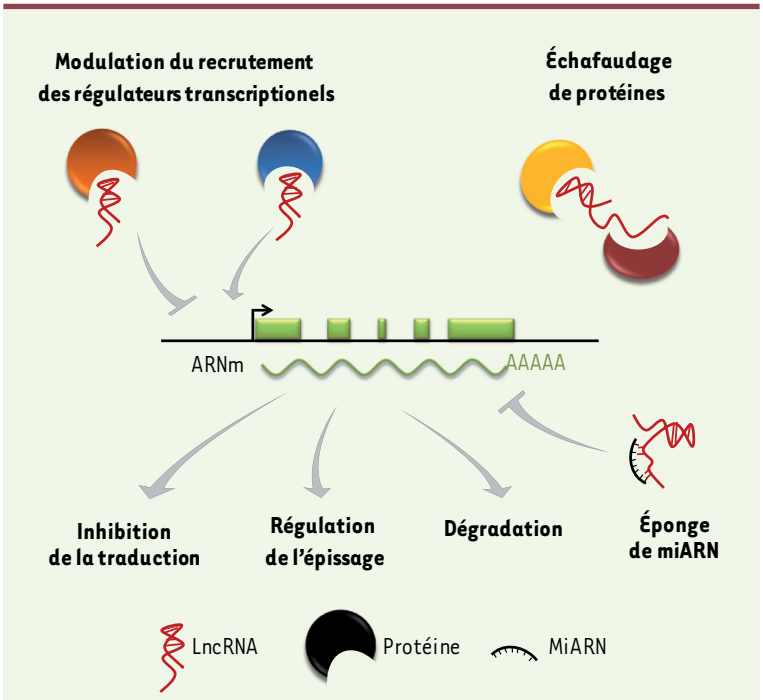



***Figure 4.*** ***Exemples de mécanismes d'action des lncRNA.*** Les lncRNA peuvent favoriser l'interaction de protéines, guider des complexes protéiques vers leurs gènes cibles, séquestrer des protéines ou des miARN (éponge de miARN) et influencer les processus post-transcriptionnels liés aux ARNm, comme la traduction, l'épissage et la dégradation de l'ARNm. Un gène codant (en vert) est représenté à titre indicatif. Les mécanismes d'action favorisant positivement ou négativement l'expression des gènes sont représentés par des flèches grises.

sous-produits d'éléments régulateurs actifs. En ce sens, nous avons démontré récemment que les couples lncRNA/ARNm transcrits de façon divergente sont le reflet d'un mécanisme spécialisé de la régulation transcriptionnelle impliquant des promoteurs bidirectionnels [7].

## Le rôle des lncRNA dans le développement des maladies et des cancers

Étant donné leurs contributions aux processus physiologiques, une altération du niveau d'expression des lncRNA pourrait mener au développement de pathologies, comme le cancer. En effet, de nombreuses études démontrant une dérégulation des lncRNA dans divers types de cellules cancéreuses suggèrent que les lncRNA peuvent agir comme des suppresseurs de tumeurs ou des oncogènes potentiels [4] *(Tableau II)*. Parmi les exemples de lncRNA associés au cancer, le rôle de *HOTAIR* a été le plus étudié. Plusieurs publications ont rapporté une surexpression de *HOTAIR* dans différents cancers, où il participerait parfois à la formation de métastases [25, 26]. *MALAT1* (*metastasis associated lung adenocarcinoma transcript 1*) est un autre lncRNA très étudié ; sa surexpression est associée à l'état métastatique de tumeurs. Son utilisation comme marqueur pronostique dans le cancer du poumon a d'ailleurs été suggérée [4]. De même, des fonctions oncogéniques ont été suggérées pour plusieurs autres lncRNA dans, entre autres, le cancer du foie (*HULC*, *highly up-regulated in liver cancer*) [27], de la prostate (*PCA3*, *prostate cancer antigen 3*) [28, 29] et du rein (*MVIH*, *microvascular invasion in HCC*) [30]. Finalement, certains lncRNA, tels que *LincRNA-p21* et *Meg3* (*maternally expressed 3*), ont été impliqués dans la modulation de la réponse p53 [31, 32].

La participation des lncRNA au développement de pathologies ne se limite cependant pas aux cancers. Le nombre d'études suggérant l'implication de lncRNA dans le développement de diverses maladies ne cesse d'augmenter [4]. Ainsi, *CDKN2B-as1 (ANRIL)*, un lncRNA impliqué dans plusieurs cancers [33], est aussi associé à l'artériosclérose [34] et à des risques cardiovasculaires accrus [35]. De la même manière, le développement de diverses pathologies, comme la maladie d'Alzheimer, le diabète néonatal transitoire et la dystrophie facio-scapulo-humérale, pourrait découler de la dérégulation de l'expression d'un lncRNA *(Tableau II)*.

Enfin, outre la dérégulation de l'expression des lncRNA, plusieurs études génétiques ont révélé la présence de mutations dans leurs séquences primaires [4]. Étant donné que la plupart des variants génétiques identifiés par des études d'association pangénomiques (GWAS) sont localisés en dehors des gènes codants, on peut s'attendre à ce qu'un grand nombre de ces mutations affectent des lncRNA [1], bien que les conséquences de ces mutations sur les fonctions des lncRNA doivent être démontrées.

| Noms | Pathologies | Réf. |
|---|---|---|
| *CDKN2B-as1* | Artériosclérose, risques cardiovasculaires, divers cancers | [33-35] |
| *BACE1-as* | Alzheimer | [40] |
| *DBE-T* | Dystrophie facio-scapulo-humérale | [41] |
| *MVIH* | Carcinome hépatique | [30] |
| *HOTAIR* | Cancer du sein, cancer colorectal | [25, 26] |
| *HULC* | Cancer du foie | [27] |
| *lncRNA HYMAI* | Diabète néonatal transitoire | [42] |
| *PCA3* | Cancer de la prostate | [28] |
| *LincRNA-p21* | Sarcome, lymphome, cancer du poumon | [31] |
| *MALAT1* | Cancer du poumon, de l'utérus, du sein, du pancréas, du rein, du côlon et de la prostate | [43-45] |

***Tableau II.*** ***Exemples de lncRNA associés à des pathologies.*** *BACE1-as* : *β-amyloid converting enzyme 1-antisense* ; *DBE-T* : *D4Z4 binding element transcript* ; CDKN2B-as1 : (ANRIL), un lncRNA impliqué dans divers cancers ; MVIH : *microvascular invasion* ; HOTAIR : *Hox antisense intergenic RNA* ; HULC : *highly up-regulated in liver cancer* ; PCA3 : *prostate cancer antigen 3* ; MALAT1 : *metastasis associated lung adenocarcinoma transcript 1*.

## Les lncRNA comme agents thérapeutiques et biomarqueurs

Étant donné leur rôle clé dans la régulation de l'expression génique, les lncRNA représentent des cibles thérapeutiques potentielles. Bien qu'ils soient encore à leur début, des essais cliniques utilisant des agents thérapeutiques à base d'ARN sont déjà à l'étude [15]. La plupart d'entre eux implique de petits ARN interférents (siRNA) ou des oligonucléotides antisens (ASO) qui mènent à la dégradation de leurs ARN cibles. Par exemple, l'utilisation *in vivo* chez la souris d'un ASO dirigé contre le lncRNA *Bdnf-as* (*brain-derived neurotrophic factor*) lève la répression de *Bdnf* et permet la prolifération neuronale [36]. Alternativement, les ASO pourraient agir comme agents bloquants en prévenant la liaison d'un lncRNA avec une protéine ou une séquence d'ADN ou d'ARN. Cependant, il est à noter que la distribution et la délivrance des ASO dans les organismes multicellulaires restent, jusqu'à ce jour, un obstacle majeur au développement de ces agents thérapeutiques [15].

Puisque plusieurs de ces lncRNA ont été associés au développement de métastases ou à la progression de tumeurs, les lncRNA peuvent servir de biomarqueurs potentiels pour le dépistage et le pronostic de cancers. Quelques-uns, comme *PCA3* et *HULC*, sont déjà utilisés pour le diagnostic du cancer de la prostate et du carcinome hépatocellulaire [27, 28]. De façon intéressante, des acides nucléiques (ARN et ADN) très stables peuvent être détectés dans les fluides corporels, comme le sang, le plasma et l'urine [37]. Il semblerait qu'il y ait une bonne corrélation entre les niveaux d'acides nucléiques circulants et les modifications génomiques, épigénétiques ou transcriptionelles associées aux tumeurs. Ces acides nucléiques seraient sécrétés ou relâchés dans le sang par des cellules cancéreuses en apoptose ou en nécrose. Notamment, des lncRNA, de par leur structure secondaire ou leur présence dans des microparticules, ont été détectés dans les fluides corporels [27, 28].

## Conclusion

Les lncRNA apparaissent comme des régulateurs clés de l'expression génique. Les quelques exemples qui ont été étudiés jusqu'à présent démontrent clairement leurs rôles dans plusieurs processus physiologiques fondamentaux. De même, la mise en évidence de leur implication dans de nombreuses pathologies, dont notamment le cancer, ne cesse d'augmenter. Leurs utilisations potentielles comme biomarqueurs et cibles thérapeutiques sont, par ailleurs, prometteuses. Le défi majeur à l'heure actuelle reste la caractérisation fonctionnelle de la grande majorité des lncRNA. Toutefois, l'identification et la cartographie de la totalité des lncRNA présents dans le génome humain devraient accélérer l'atteinte de cet objectif. Cela implique des efforts conjoints entre la génomique fonctionnelle, l'épigénomique et la bioinformatique. La mise en place de consortiums internationaux, tels que le projet ENCODE et le projet européen « A *BLUEPRINT of haematopoietic epigenome* » [38], partenaire de l'IHEC (*international human epigenome consortium*), joue un rôle déterminant dans ce contexte. On peut déjà être certain qu'avec tous les nouveaux questionnements découlant des fonctions associées aux lncRNA, l'intérêt des chercheurs à l'égard des lncRNA n'est pas près de s'estomper ! ◊


## SUMMARY

### Functions of lncRNA in development and diseases

The transcription of essentially the entire eukaryotic genome generates a myriad of non-coding RNA species that show complex overlapping patterns of expression and regulation. In the last decade, several large scale genomic analyses have shed light on the widespread existence of long non-coding RNAs (lncRNAs) in mammals. Although the function of most lncRNAs remains unknown, many of them have been suggested to play important roles in the regulation of gene expression during normal development and diseases, including cancers. Indeed, functional studies have demonstrated that lncRNAs participate in various biological processes, including reprogramming of pluripotent stem cells, oncogenic progression and cell cycle regulation. In this review, we summarize recent findings about the biology of lncRNAs and their functions in normal and pathological development in mammals. ◊



## REMERCIEMENTS

*Nous remercions le Dr Jean Imbert pour la lecture critique de cette synthèse. Le travail de recherche dans le laboratoire de S. Spicuglia est financé par l'Union européenne (FP7 programme n° 282510-BLUEPRINT), l'ARC (projet n° SFI20111203756) et A*MIDEX (projet n° ANR-11-IDEX-0001-02). L.T.M. Dao est financée par le ministère de l'Éducation et de la formation du Vietnam. Nous nous excusons auprès de ceux dont les travaux, pour des raisons de place, n'apparaissent pas dans cette synthèse.*


## LIENS D'INTÉRÊT

*Les auteurs déclarent n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.*

## RÉFÉRENCES


1. Derrien T, Guigo R, Johnson R. The long non-coding RNAs: A new (p)layer in the dark matter. *Front Genet* 2011 ; 2 : 107.
2. Hu W, Alvarez-Dominguez JR, Lodish HF. Regulation of mammalian cell differentiation by long non-coding RNAs. *EMBO Rep* 2012 ; 13 : 971-83.
3. Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. *Cell* 2013 ; 154 : 26-46.
4. Shi X, Sun M, Liu H, *et al.* Long non-coding RNAs: a new frontier in the study of human diseases. *Cancer Lett* 2013 ; 339 : 159-66.
5. Esteller M. Non-coding RNAs in human disease. *Nat Rev Genet* 2011 ; 12 : 861-74.
6. Koerner MV, Pauler FM, Huang R, Barlow DP. The function of non-coding RNAs in genomic imprinting. *Development* 2009 ; 136 : 1771-83.
7. Lepoivre C, Belhocine M, Bergon A, *et al.* Divergent transcription is associated with promoters of transcriptional regulators. *BMC Genomics* 2013 ; 14 : 914.
8. Kelley DR, Rinn JL. Transposable elements reveal a stem cell specific class of long noncoding RNAs. *Genome Biol* 2012 ; 13 : R107.
9. Kapusta A, Kronenberg Z, Lynch VJ, *et al.* Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* 2013 ; 9 : e1003470.
10. Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. *Nature* 2012 ; 482 : 339-46.
11. Guttman M, Amit I, Garber M, *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 2009 ; 458 : 223-7.
12. Spicuglia S, Maqbool MA, Puthier D, Andrau JC. An update on recent methods applied for deciphering the diversity of the noncoding RNA genome structure and function. *Methods* 2013 ; 63 : 3-17.
13. Derrien T, Guigo R. De longs ARN non codants activateurs de la transcription des gènes. *Med Sci (Paris)* 2011 ; 27 : 359-61.
14. Derrien T, Johnson R, Bussotti G, *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 2012 ; 22 : 1775-89.
15. Yang L, Froberg JE, Lee JT. Long noncoding RNAs: fresh perspectives into the RNA world. *Trends Biochem Sci* 2014 ; 39 : 35-43.
16. Orom UA, Shiekhattar R. Long noncoding RNAs usher in a new era in the biology of enhancers. *Cell* 2013 ; 154 : 1190-3.
17. Lee JT. Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control. *Nat Rev Mol Cell Biol* 2011 ; 12 : 815-26.
18. Rinn JL, Kertesz M, Wang JK, *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 2007 ; 129 : 1311-23.
19. Tsai MC, Manor O, Wan Y, *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 2010 ; 329 : 689-93.
20. Bertani S, Sauer S, Bolotin E, Sauer F. The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin. *Mol Cell* 2011 ; 43 : 1040-6.
21. Wang KC, Yang YW, Liu B, *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 2011 ; 472 : 120-4.

## RÉFÉRENCES


22. Carrieri C, Cimatti L, Biagioli M, *et al.* Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature* 2012 ; 491 : 454-7.
23. Memczak S, Jens M, Elefsinioti A, *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 2013 ; 495 : 333-8.
24. Spicuglia S, Zacarias-Cabeza J, Pekowska P, Ferrier P. Epigenetic regulation of antigen receptor gene rearrangement. *F1000 Biol Rep* 2010 ; 2 : 23.
25. Kogo R, Shimamura T, Mimori K, *et al.* Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res* 2011 ; 71 : 6320-6.
26. Gupta RA, Shah N, Wang KC, *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 2010 ; 464 : 1071-6.
27. Panzitt K, Tschernatsch MM, Guelly C, *et al.* Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA. *Gastroenterology* 2007 ; 132 : 330-42.
28. Fradet Y, Saad F, Aprikian A, *et al.* uPM3, a new molecular urine test for the detection of prostate cancer. *Urology* 2004 ; 64 : 311-6.
29. Ferreira LB, Palumbo A, de Mello KD, *et al.* PCA3 noncoding RNA is involved in the control of prostate-cancer cell survival and modulates androgen receptor signaling. *BMC Cancer* 2012 ; 12 : 507.
30. Yuan SX, Yang F, Yang Y, *et al.* Long noncoding RNA associated with microvascular invasion in hepatocellular carcinoma promotes angiogenesis and serves as a predictor for hepatocellular carcinoma patients' poor recurrence-free survival after hepatectomy. *Hepatology* 2012 ; 56 : 2231-41.
31. Huarte M, Guttman M, Feldser D, *et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* 2010 ; 142 : 409-19.
32. Zhou Y, Zhong Y, Wang Y, *et al.* Activation of p53 by MEG3 non-coding RNA. *J Biol Chem* 2007 ; 282 : 24731-42.
33. Pasmant E, Laurendeau I, Sabbagh A, *et al. ANRIL* ou l'étrange histoire d'un grand ARN non codant. *Med Sci (Paris)* 2010 ; 26 : 564-6.
34. Holdt LM, Beutner F, Scholz M, *et al.* ANRIL expression is associated with atherosclerosis risk at chromosome 9p21. *Arterioscler Thromb Vasc Biol* 2010 ; 30 : 620-7.
35. Liu Y, Sanoff HK, Cho H, *et al.* INK4/ARF transcript expression is associated with chromosome 9p21 variants linked to atherosclerosis. *PLoS One* 2009 ; 4 : e5027.
36. Modarresi F, Faghihi MA, Lopez-Toledano MA, *et al.* Inhibition of natural antisense transcripts *in vivo* results in gene-specific transcriptional upregulation. *Nat Biotechnol* 2012 ; 30 : 453-9.
37. Tong YK, Lo YM. Diagnostic developments involving cell-free (circulating) nucleic acids. *Clin Chim Acta* 2006 ; 363 : 187-96.
38. Adams D, Altucci L, Antonarakis SE, *et al.* Blueprint to decode the epigenetic signature written in blood. *Nat Biotechnol* 2012 ; 30 : 224-6.
39. Cesana M, Cacchiarelli D, Legnini I, *et al.* A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* 2011 ; 147 : 358-69.
40. Faghihi MA, Modarresi F, Khalil AM, *et al.* Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat Med* 2008 ; 14 : 723-30.
41. Cabianca DS, Casa V, Bodega B, *et al.* A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. *Cell* 2012 ; 149 : 819-31.
42. Temple IK, Shield JP. Transient neonatal diabetes, a disorder of imprinting. *J Med Genet* 2002 ; 39 : 872-5.
43. Ji P, Diederichs S, Wang W, *et al.* Malat-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 2003 ; 22 : 8031-41.
44. Yamada K, Kano J, Tsunoda H, *et al.* Phenotypic characterization of endometrial stromal sarcoma of the uterus. *Cancer Sci* 2006 ; 97 : 106-12.
45. Lin R, Maeda S, Liu C, *et al.* A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. *Oncogene* 2007 ; 26 : 851-8.
46. Abel Y, Clerget G, Bourguignon-Igel V, *et al.* Les petits ARN nucléolaires nous surprennent encore ! *Med Sci (Paris)* 2014 ; 30 : 297-302.



**TIRÉS À PART**
S. Spicuglia

# Annexe III : Autres articles

## Les articles publiés :

**Article 5: H3K4 tri-methylation provides an epigenetic signature of active enhancers.**
Contribution : Analyses des données ChIP-seq (Fig.4)

**Article 6: Rôle des longs ARN non codants dans le développement normal et pathologique.**
Contribution : Participation dans la rédaction et création des figures (Fig.1 ,3 et 4)

**Article 7: Triggering the TCR developmental checkpoint activates a therapeutically targetable tumor suppressive pathway in T-cell Leukemia.**
Contribution : Analyses des données *Microarray* – GSEA (Fig.4E et 4F)

## Les articles soumis :

**Article 8: TCRα rearrangments identify a subgroup of NKL-deregulated adult T-ALLs associated with favorable outcome.**
Contribution : Analyses des données *Microarray* et TLDA (Fig.3 et 4)

**Article 9: CBFβ-SMMHC directly regulates ribosomal gene transcription and alters ribosome biogenesis.**
Contribution : Analyses des données *Microarray* - GSEA (Fig.1E)

**Article 10: Homeobox protein TLX3 activates miR-125b expression to promote T-cell acute lymphoblastic leukemia.**
Contribution : Analyses des données *RNA-seq* (Fig.1C, 2C et 8C)