

# THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Biochimie et Biologie Moléculaire

École doctorale : CBS2 - Sciences Chimiques et Biologiques pour la Santé

Unité de recherche : Institut de Génétique Moléculaire de Montpellier (IGMM - UMR 5535)

## Étude des éléments régulateurs de l'expression des gènes chez l'humain

Présentée par Chloé BESSIÈRE  
Le 27 Novembre 2018

Sous la direction de Charles-Henri LECELLIER

Devant le jury composé de

Dr. Anthony MATHELIER, Group leader, NCMM, Oslo

Rapporteur

Dr. Cédric NOTREDAME, Group leader, CRG, Barcelona

Rapporteur

Dr. Thérèse COMMES-MAERTEN, Professeur, Université de Montpellier

Examinateur

Dr. Raphaël MOURAD, Maître de Conférence, Université Paul Sabatier, Toulouse

Examinateur

Dr. Jean-Christophe ANDRAU, Directeur de recherche, CNRS, Montpellier

Examinateur

Dr. Charles-Henri LECELLIER, Chargé de recherche, CNRS, Montpellier

Directeur de thèse



UNIVERSITÉ  
DE MONTPELLIER



*"La vie, c'est comme une bicyclette, il faut avancer pour ne pas perdre l'équilibre."*

Albert Einstein



# Remerciements

A 8 semaines de la soutenance de thèse, il est maintenant grand temps de finaliser mon mémoire et de retourner un petit peu à la "vraie science" comme dirait mon directeur de thèse Charles Lecellier. Je vais justement commencer par le remercier, pour m'avoir accueillie dans son équipe et avoir été si disponible durant ces 3 années. Il a été patient quand j'en avais besoin, je lui en suis reconnaissante. C'est un passionné, jamais à court d'idées et à coté de qui l'on apprend beaucoup. Je le remercie également pour cette opportunité qu'il me donne, je dirais même cette chance, d'aller faire un court séjour à Vancouver. Je tiens également à remercier Laurent Bréhélin et Sophie Lèbre, sans qui notre équipe de travail ne serait pas la même, et qui apportent de la bonne humeur dans nos réunions. Ils m'ont co-encadrée sur une partie de ma thèse, m'ont donné de précieux conseils, notamment en statistiques, et ont suivi de près ou de loin le reste de mes travaux. Merci à Sophie, pour la touche féminine parmi les encadrants de notre petite équipe, à Christophe et May qui ont commencé leur thèse en même temps que moi ainsi qu'aux non-permanents (stagiaires, post-doc) qui se sont joints à nous pour quelques mois ou plus, et bonne chance à Raphaël qui prend la relève en tant que doctorant !

Je remercie les membres de mon jury d'avoir accepté l'invitation d'assister à la présentation de mon travail et de le juger, mes rapporteurs, Anthony Mathelier et Cédric Notredame, et mes examinateurs, Raphaël Mourad, Thérèse Commes et Jean-Christophe Andrau. Merci à mes rapporteurs pour leurs précieux conseils et avis positifs suite à la lecture de mon mémoire. Je remercie particulièrement Cédric Notredame, Thérèse Commes et Jean-Christophe Andrau qui m'ont suivie lors des comités de thèse ou projets en collaboration et qui ont su m'aiguiller et me donner des conseils.

De façon plus générale, je souhaite remercier tous les membres des deux instituts dans lesquels j'ai travaillé, l'IGMM et les membres de l'équipe d'Edouard Bertrand et de l'équipe de Rémi Bordonné. Même si entre biologistes et bioinformaticiens

on ne parle pas toujours la même langue, ils ont su me faire des remarques très intéressantes et pour certains m'ont aidée pendant ma courte période de manips ! Merci également pour tous les moments de partage, les anniversaires et gâteaux qui vont avec. Je remercie l'équipe MAB à l'IBC, pour les pauses cafés, les discussions, scientifiques ou non, et je remercie les personnes avec qui j'ai pu travailler sur des projets annexes enrichissants.

Je remercie grandement May, sans qui ma thèse n'aurait jamais été la même. Comme dans toute période de vie, mais encore plus amplifié en étant en thèse, on a eu des hauts et des bas et j'ai toujours pu compter sur elle. C'est mon binôme de thèse, avec qui on peut discuter de tout et se plaindre quand on en a besoin. Je suis fière d'avoir été la testeuse attitrée de toutes tes nouveautés culinaires et je pense que tu es maintenant prête à ouvrir ton restaurant (tu n'oublieras pas d'afficher ton diplôme de docteur et un poster de Panda) ! Je remercie également Moana, ma partenaire de bioinformatique depuis l'écriture de nos premières lignes de codes, pour nos discussions de soutien mutuel et pour nos sessions BU, parfois manquées mais qui ont su s'intensifier au bon moment.

Je remercie tous mes amis, ceux que j'ai rencontré à Sup Agro, ou autour du premier barbeuc de demi-anniversaire, pour les bons moments passés ensemble. Même si l'on se voit moins, les parties de bang resteront emblématiques ! Je remercie mes bichettes pour nos rencontres et moments entre filles au moins une fois par an. Je remercie Charline, après quelques temps sans se voir, il a fallu venir te retrouver en Thaïlande mais depuis, on a réussi à rattraper le temps perdu ; et mes autres amis de Peip/Polytech que j'arrive encore à voir de temps en temps. Je remercie Marie et Pauline, mes plus grandes amies de collège, même en vacances sous la pluie c'est un plaisir de se retrouver ! Je remercie les 5 doigts, pas toujours facile de se rassembler au même endroit au même moment mais les nombreux GIF que je reçois égayent mes journées, et je remercie le reste des 13 du lycée avec qui je passe encore de rares mais bons moments. Je remercie Émeline, mon amie de longue date et voisine préférée, un peu compliquée de se voir ces derniers temps, mais notre table au Café des Arts nous attend pour une session rattrapage !

Je remercie toutes les personnes que j'ai pu côtoyer à Sunsud, avec qui j'ai joué au bad ou discuté, pour les sessions volley l'été sur la plage et pour les soirées jeux, en particulier Mélanie et Samir qui sont devenus de vrais amis. Je remercie également Juvibad (ou Juvibièvre, l'un ne marche pas sans l'autre) et toutes les personnes que

j'ai rencontrées, pour les moments de défoule sur les terrains et de détente autour d'une bière, les tournois dans la bonne humeur et les soirées raclette, andouillette, ou autres idées farfelues, avec une mention spéciale pour ma Jus', amie et partenaire au top !

Je remercie Clément, avec qui j'ai fait un bon bout de chemin de vie, il m'a toujours supportée, encouragée, et m'a permis d'évoluer et d'être ce que je suis aujourd'hui. Je remercie également toute sa famille ayant toujours été adorable avec moi, et Michèle pour ses bons petits plats.

Je remercie du fond du cœur toute ma famille de m'avoir soutenue et conseillée. Ma soeur, chez qui j'ai squatté un paquet de fois le midi pour de bons petits repas, jamais sans le café et le carré de chocolat, les essentiels pour garder la ligne en thèse ; mon père pour m'avoir donné goût à la recherche, pour les petites vacances partagées en famille, et avec Natasa pour nous avoir apporté à Lola et moi, une petite Oliana pleine de vie. Je remercie ma maman, sans qui je n'en serais pas là aujourd'hui, toujours présente et bienveillante dans les périodes difficiles, elle a su me chouchouter et me soutenir du début à la fin, ainsi que Jean-marie et sa bonne humeur ; merci aussi pour les longs weekends partagés en Espagne et au Portugal avec vous et tous les enfants, qui ont été de vrais moments de bonheur. Je remercie ma mamie, ma mona et mon papou pour leur joie de vivre, pour toutes leurs attentions et bon produits de Lozère ; mes cousines, oncle et tante aveyronnais ; ma tata dont j'admire la force aujourd'hui, et mes cousins, je suis fière de vous et de votre courage. En fait, je suis sacrément fière de ma famille au complet et d'avoir un petit bout de chacun d'eux en moi.

Enfin je remercie Théo, ma bulle de réconfort pendant les périodes difficiles, tu as su me re-motiver et me donner confiance quand je baissais les bras, me supporter dans la fatigue, le stress et la râlerie, et je sais que ce n'est pas une tâche facile ! Je n'ai pas partagé la période de ma thèse, et de ma vie, la plus facile avec toi, mais tu as su faire preuve de patience, de compréhension et prendre soin de moi comme un chef avec toutes tes petites attentions, je te dédie la palme d'or du meilleur compagnon de rédaction de thèse !

Pour finir, et parce que je ne les oublierai jamais, j'ai une très grande pensée pour mon papi et mon tonton Eric parti trop tôt.



# Table des matières

<b>Table des figures</b>	<b>X</b>
<b>Introduction générale</b>	<b>1</b>
<b>1 État de l'art</b>	<b>3</b>
1.1 Organisation tri-dimensionnelle de la chromatine au sein du noyau . . . . .	4
1.1.1 De la double hélice d'ADN aux chromosomes . . . . .	4
1.1.2 Hiérarchie de la chromatine et interactions . . . . .	9
1.1.3 Compartimentation spatio-temporelle du noyau . . . . .	17
1.2 Segmentation du génome . . . . .	21
1.2.1 Gènes et annotations . . . . .	21
1.2.2 Différentes classes de gènes et leur organisation . . . . .	26
1.2.3 Séquences régulatrices . . . . .	30
1.2.4 Modèles de segmentation . . . . .	35
1.2.5 Éléments répétés . . . . .	38
1.3 Régulation de l'expression génique . . . . .	45
1.3.1 Facteurs de transcription et séquences régulatrices . . . . .	45
1.3.2 Transcription par l'ARN polymérase II . . . . .	49
1.3.3 Quantification de l'expression des gènes . . . . .	52
1.4 Dérégulation des contrôles de l'expression des gènes : variants et pathologies . . . . .	57
1.4.1 Altérations du génome . . . . .	57
1.4.2 Variants et maladies . . . . .	60
<b>2 Résultats</b>	<b>63</b>
2.1 Instructions de régulation de l'expression des gènes présentes dans la séquence ADN . . . . .	63
2.1.1 Choix du modèle statistique . . . . .	63
2.1.2 Modèles de prédiction de l'expression des gènes . . . . .	66
2.1.3 Résultats et contribution . . . . .	70

2.1.4	Conclusion . . . . .	72
2.2	Caractérisation d'une nouvelle classe de longs ARNs non-codants introniques . . . . .	101
2.2.1	Signal de transcription associé à un motif poly-T . . . . .	101
2.2.2	Caractéristiques des CAGEs associés à un motif poly-T . . . . .	101
2.2.3	Expression des CAGEs associés à un motif poly-T et lien avec leurs gènes hôtes . . . . .	103
2.2.4	Conclusion . . . . .	105
2.3	Projet annexe : modélisation de la vitesse d'elongation de l'ARN polymérase II . . . . .	156
2.3.1	Méthode expérimentale . . . . .	156
2.3.2	Méthode computationnelle de détection des fronts . . . . .	157
2.3.3	Modèle de prédiction du taux d'elongation . . . . .	159
2.3.4	Résultats . . . . .	160
2.3.5	Limites de la méthode expérimentale et du modèle . . . . .	165
2.3.6	Travaux en cours : modélisation du taux d'elongation d'un point de vue biophysique . . . . .	166
<b>3</b>	<b>Discussion et perspectives</b>	<b>168</b>

# Table des figures

1.1	De la double hélice d'ADN au chromosome . . . . .	5
1.2	Structure d'un nucléosome . . . . .	6
1.3	Différentes modifications d'histones . . . . .	7
1.4	Combinaisons de marques épigénétiques associées à des gènes activés ou réprimés . . . . .	8
1.5	Principe du Hi-C . . . . .	10
1.6	Définition des TADs . . . . .	11
1.7	TADs ou domaines similaires chez plusieurs espèces . . . . .	12
1.8	Compartiments A et B . . . . .	14
1.9	TADs versus compartiments A et B . . . . .	14
1.10	Interactions inter-chromosomes et territoires chromosomiques . . . . .	16
1.11	Découverte de l'existence des territoires chromosomiques . . . . .	17
1.12	Densité de la chromatine au sein du noyau par microscopie . . . . .	18
1.13	Lamina et LADs constitutifs et facultatifs . . . . .	19
1.14	Stratégies d'annotation du génome . . . . .	23
1.15	Catégories de lncRNA établies par FANTOM . . . . .	25
1.16	Logo des motifs associés aux sites d'épissage accepteur et donneur . .	27
1.17	Organisation d'un gène . . . . .	28
1.18	Éléments régulateurs distaux . . . . .	31
1.19	Fonctionnement des insulators . . . . .	34
1.20	Principe d'une expérience de ChIP-seq . . . . .	36
1.21	Proportion des différents composants du génome humain . . . . .	38
1.22	Classification des éléments transposables . . . . .	39
1.23	Structure d'un transposon à ADN . . . . .	40
1.24	Les différentes classes de rétrotransposons . . . . .	40
1.25	Rétrotransposition des SINES et LINEs . . . . .	42
1.26	Modèle de représentation d'un motif sous forme de PWM . . . . .	48
1.27	Structure de l'ARN polymérase II . . . . .	50
1.28	Profil de densité de l'ARN Polymérase II pour un gène hypothétique .	52

1.29 Séquençage Heliscope adapté aux données de CAGE . . . . .	55
1.30 Principales technologies de séquençage pour quantifier le transcriptome	56
2.1 Comparaison de la géométrie de la pénalisation du lasso et du ridge .	65
2.2 CAGEs associés à un motif poly-T exprimés versus non-exprimés . .	103
2.3 Profils de ChIP-seq de l'ARN Polymérase II . . . . .	157
2.4 Principe de l'algorithme de programmation dynamique de détection des fronts d'ARN Pol II optimaux . . . . .	159
2.5 Représentation du taux d'elongation de l'ARN Polymérase II en fonc- tion de la distance parcourue entre le front et la fin du gène. . . . .	160
2.6 Représentation schématique des fronts de polymérases avançant sur le gène . . . . .	161
2.7 Distribution des taux d'elongation des ARNs Pol II aux différents intervalles de temps . . . . .	162
2.8 Matrice de données des taux d'elongation de l'ARN Pol II. . . . .	163
3.1 Mutations somatiques observées dans différents types de cancers . . .	173

# Introduction générale

Au sein de chaque organisme vivant, l'expression du génome contenant l'ensemble de l'information génétique est étroitement contrôlée afin d'assurer une grande diversité de types cellulaires et de fonctions. Selon le dogme central de la biologie, les informations permettant à la cellule d'assurer ses fonctions sont contenues dans les gènes. Ces gènes donnent naissance à des protéines via la transcription de l'ADN en ARN, puis la traduction de l'ARN en chaîne d'acides aminés. Cependant, ces gènes ne représentent qu'une très petite portion de notre génome et on quantifie les exons codants des gènes à 2% du génome seulement. Le reste a longtemps été considéré comme n'ayant aucun rôle fonctionnel et a été classé sous le terme de "junk DNA", terme utilisé par le chercheur Susumu Ohno en 1972 pour caractériser l'ensemble des parties non-codantes du génome.

Au fil des années et grâce au premier séquençage complet du génome humain par le "Human Genome Project" [37] entre 1990 et 2003, un rôle important des séquences non-codantes émerge. Le projet Encyclopaedia of DNA Elements (ENCODE) [35] a vu le jour en 2003 en ayant pour but la caractérisation de l'ensemble des éléments fonctionnels du génome humain. Les auteurs de ce projet ont déclaré qu'une fonction biologique pouvait être associée à 80% du génome humain et qu'au delà des gènes codants, de nombreux éléments de l'ADN auraient un rôle dans l'apparition de pathologies. L'ADN peut en fait être vu comme un gigantesque tableau de bord avec des millions d'interrupteurs génétiques qui permettent de contrôler de manière globale ou spécifique les gènes, de façon à ce qu'ils soient "allumés" et "éteints" au bon moment et au bon endroit [*Ewan Birney (LEBM-IEB)*].

A travers cette thèse, j'ai pu évaluer le potentiel de l'information contenue dans notre séquence ADN en tant que régulateur transcriptionnel et post-transcriptionnel de l'expression des gènes. J'ai également mis en évidence une catégorie d'ARNs non-codants présents dans les introns et l'impact qu'ils pouvaient avoir au niveau de l'expression de leur gène hôte. Dans un premier chapitre introductif, nous décrirons

les mécanismes de régulation de l'expression des gènes connus aujourd'hui et la place que porte l'ADN non-codant dans cette régulation. Les résultats de ma thèse seront présentés dans un second chapitre séparé en 3 parties, les deux premières présentant les deux articles (publié ou en cours) définissant ma thèse et la troisième consacrée à un projet annexe encore en cours. Enfin, la dernière partie sera dédiée à la discussion des résultats obtenus au cours de ma thèse ainsi qu'aux perspectives.

# Chapitre 1

## État de l'art

Tous les organismes vivants sont constitués de cellules, unités de base du vivant, à l'intérieur desquelles se trouvent de nombreuses molécules leurs permettant d'être autonomes et de se reproduire. Parmi ces molécules, l'ADN est constitué de deux brins complémentaires (sens et anti-sens) sous forme d'une double hélice de nucléotides : A (Adénine), C (Cytosine), G (Guanine) ou T (Thymine). Un nucléotide est donc l'unité de base de l'ADN, et ils sont reliés entre eux par des liaisons covalentes. L'ADN étant double brin, les 4 bases sont complémentaires : le A est le complémentaire du T et le G le complémentaire du C. Le lien entre les deux brins complémentaires s'effectue par la présence de liaisons hydrogènes entre chaque paire de nucléotides. De plus, chaque brin de l'ADN est orienté de son extrémité 5' phosphate (5'P) de la 1<sup>re</sup> base vers l'extrémité 3' hydroxyle (3'OH) de la dernière base et de manière antiparallèle.

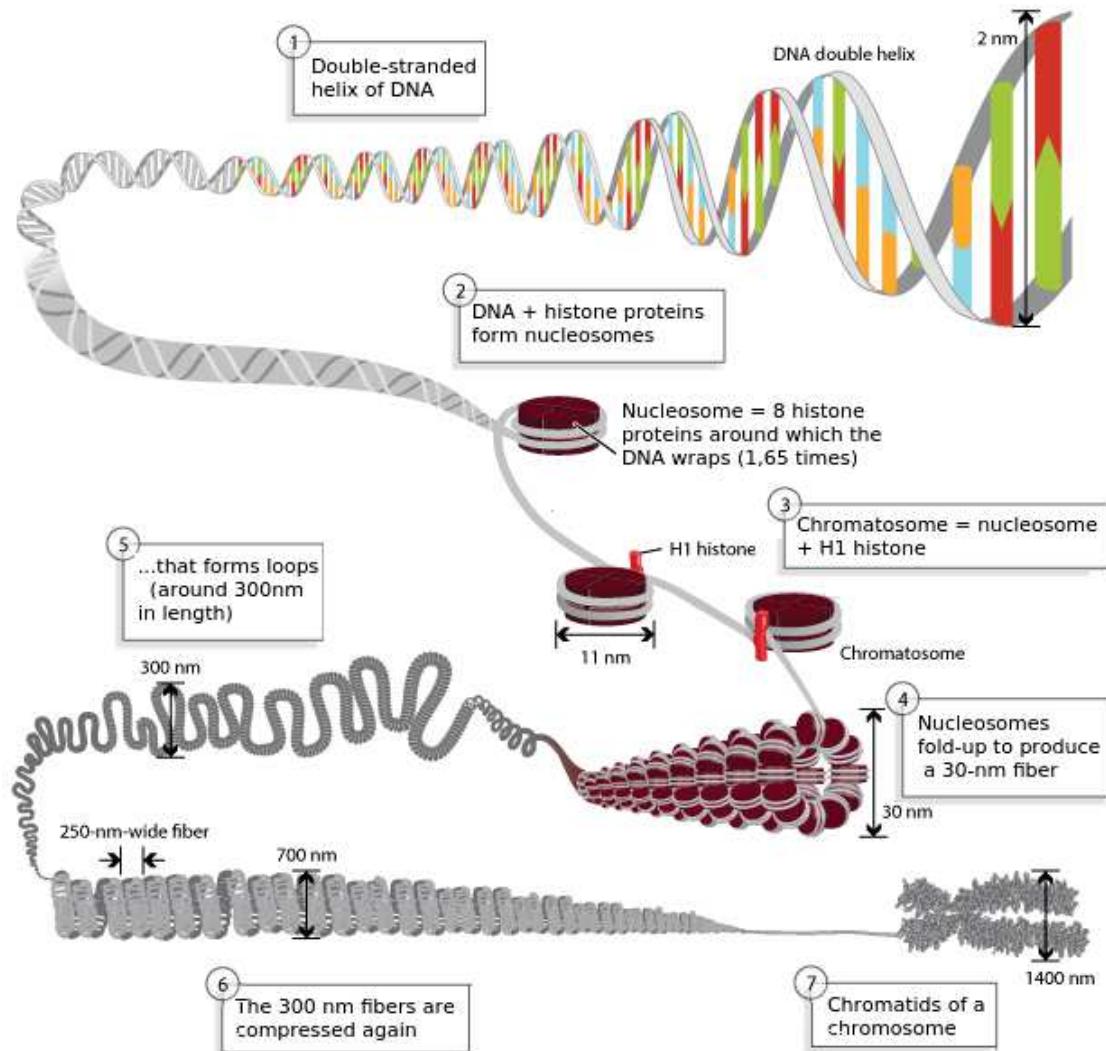
On appelle génome l'ensemble du matériel génétique encrypté dans cet ADN et qui apporte toute l'information nécessaire au développement d'une cellule et à son bon fonctionnement. Pour les organismes eucaryotes comme l'homme ou les plantes, le génome est contenu dans le noyau de la cellule. L'information portée par le génome humain est répartie sur un ensemble de 23 paires de chromosomes. Les chromosomes ont été découverts il y a plus d'un siècle par Walther Flemming [74], lors de la division cellulaire et grâce à un colorant de l'ADN qui a permis de les visualiser distinctement. Au sein du noyau, les chromosomes, et donc l'ADN, sont compactés de manière très ordonnée et l'ADN ne se retrouve jamais nu ; il est en interaction avec de nombreuses molécules dont des protéines. La structure sous laquelle se trouve l'ADN empaqueté dans le petit volume du noyau s'appelle la chromatine.

## 1.1 Organisation tri-dimensionnelle de la chromatine au sein du noyau

Au début du 20<sup>eme</sup> siècle, Heitz introduit les termes d'hétérochromatine et eu-chromatine [198]. L'hétérochromatine est la partie de la chromatine très compactée et inaccessible, qui reste condensée après la mitose. A l'inverse, l'euchromatine est décondensée. Cette observation souligne déjà l'importance de la structure chromatinnienne au sein du noyau. Au fil des années, ce principe dualiste n'est plus suffisant pour caractériser la dynamique de la chromatine et des nuances sont intégrées dans le vocabulaire : l'hétérochromatine facultative est introduite pour désigner les régions d'hétérochromatine potentiellement capables de devenir euchromatine [21], et l'hétérochromatine constitutive permet de caractériser l'hétérochromatine qui ne change pas d'état. La chromatine est une structure complexe qui peut subir des modifications chimiques réversibles permettant de moduler l'expression des gènes, sans modification de la séquence ADN. Les organismes sont ainsi déterminés par leur génome mais aussi par la modulation spécifique de l'information contenue dans la séquence primaire via des mécanismes dits épigénétiques. Le développement de techniques microscopiques et moléculaires performantes permet aujourd'hui une étude poussée de l'organisation spatiale du génome dans le noyau des cellules.

### 1.1.1 De la double hélice d'ADN aux chromosomes

Chez l'homme, l'ADN est une molécule de 3,3 milliards de paires de bases qui, mises bout à bout, atteignent une longueur totale d'environ 2 mètres. L'espace nucléaire dans lequel il est contenu ne dépasse pas quelques micromètres. Cette molécule est ainsi compactée grâce à divers repliements et interactions avec des protéines qui forment la chromatine. L'unité de base de la chromatine est le nucléosome, composé de huit protéines histones chargées positivement [141] autour desquelles l'ADN chargé négativement est enroulé (voir Figure 1.1). C'est le premier niveau de compaction de l'ADN. La succession de nucléosomes se présente sous la forme d'un "collier de perles". Les nucléosomes interagissent ensuite entre eux pour former une fibre de 30 nm de diamètre. Cette fibre est à son tour repliée sous forme de boucles s'organisant en domaines de l'ordre de la mégabase. Enfin, le dernier niveau d'enroulement est à la base des bras des chromosomes appelés chromatides. Dans cette présente sous-partie, nous allons voir que la structure de la chromatine joue un rôle important dans la régulation de l'expression des gènes.

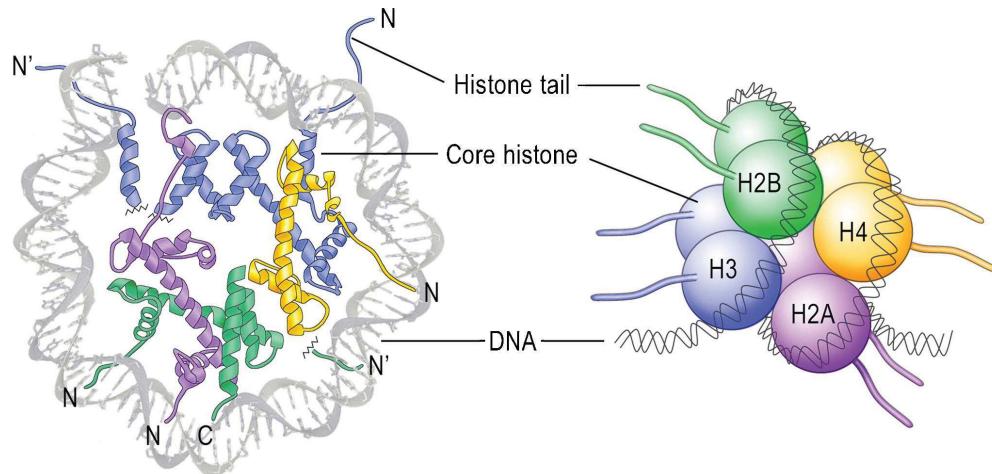


**FIGURE 1.1 – De la double hélice d'ADN au chromosome.** La double hélice d'ADN d'une largeur de 2nm est enroulée autour des coeurs d'histones appelés nucléosomes. Les nucléosomes sont à leur tour repliés pour former une fibre de chromatine de 30 nm de diamètre, elle-même enroulée sous forme de boucles étroitement repliées pour former les bras des chromosomes. [Figure adaptée de Pierce, Benjamin. Genetics : A Conceptual Approach, 2nd ed.]

## Les nucléosomes

Le cœur protéique des nucléosomes est composé de deux dimères d'histones H3-H4 et deux dimères H2A-H2B (voir Figure 1.2) [191].

Le fragment d'ADN s'enroulant autour du cœur d'histones a une longueur de 145 à 147 paires de bases (pb) correspondant à 1,65 tours. Cette structure est 6 fois plus compacte qu'un fragment d'ADN nu de même longueur. Une autre protéine, l'histone H1, est connue comme l'histone de liaison et s'associe à l'ADN se trouvant en dehors des nucléosomes, formant avec ces derniers une particule nom-



**FIGURE 1.2 – Structure d'un nucléosome.** Double hélice d'ADN autour des nucléosomes. Vue en 2D (gauche) et de façon schématique en 3D (droite) avec les différentes protéines histones (H2A, H2B, H3 et H4). [Figure extraite de [86]]

mée chromatosome et permettant de stabiliser la chromatine. L'histone H1 a un rôle important dans la structure générale de la chromatine mais son inhibition n'est pas létale, contrairement à ce qui a pu être observé pour les autres histones composant le nucléosome [131]. On trouve des nucléosomes environ toutes les 200 pb, d'où la structure caractéristique du "collier de perles". Les nucléosomes, selon leur repliement, rendent l'ADN plus ou moins accessible à la transcription et sont des modulateurs de l'expression des gènes.

### La fibre de chromatine

La succession de nucléosomes régulièrement espacés forme une fibre d'un diamètre d'environ 10 nm. Cette fibre est également repliée sur elle même pour former une autre fibre d'environ 30 nm de diamètre, correspondant à un niveau de compaction supérieur assuré par les interactions entre les nucléosomes via l'histone H1 [258]. Pour cette fibre plus large, plusieurs modèles ont été proposés pour représenter l'agencement des chromatosomes dont deux principaux. Le premier modèle est représenté par un axe imaginaire autour duquel s'enroulent les nucléosomes successifs. Ce modèle est appelé modèle "solénoïde" [72, 144] et a la forme d'une hélice simple. Le deuxième modèle propose une structure où chaque nucléosome est en contact avec ses seconds voisins et non avec les nucléosomes consécutifs. Cet agencement entraîne une structure en zigzag [279].

### Les marques épigénétiques

La modification de la structure tri-dimensionnelle de l'ADN est assurée par des

modifications chimiques de l'ADN et des protéines histones, regroupées sous le nom de marques épigénétiques.

**Modifications d'histones.** Les histones possèdent une queue N-terminale d'une 30<sup>aïne</sup> d'acides aminés exposée à des modifications post-traductionnelles (voir Figure 1.2) qui font partie des marques épigénétiques [86]. Ces modifications peuvent être de plusieurs types : mono, di et triméthylation, acétylation, ubiquitinylation, sumoylation, phosphorylation ou encore hydroxylation, et peuvent toucher plusieurs résidus particuliers : lysine (K), arginine (R), sérine (S)... Ces modifications sont nommées en fonction du type de groupement, du résidu touché et de la position de ce résidu sur la queue N-terminale (voir Figure 1.3). Par exemple, la marque épigénétique H3K4me3 est une tri-méthylation de la 4<sup>eme</sup> lysine de la queue N-terminale de l'histone H3.

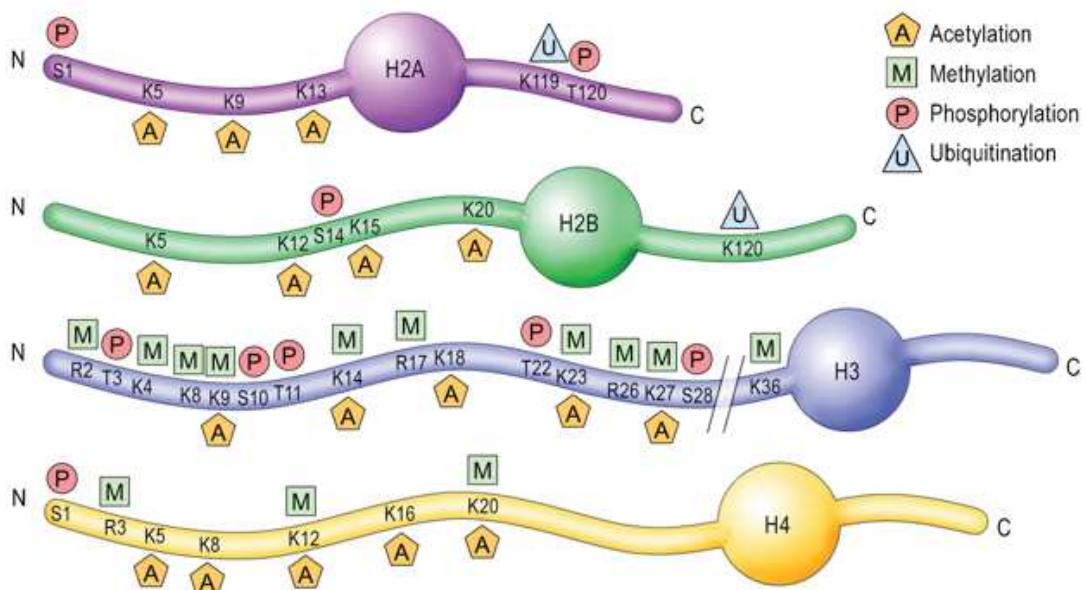
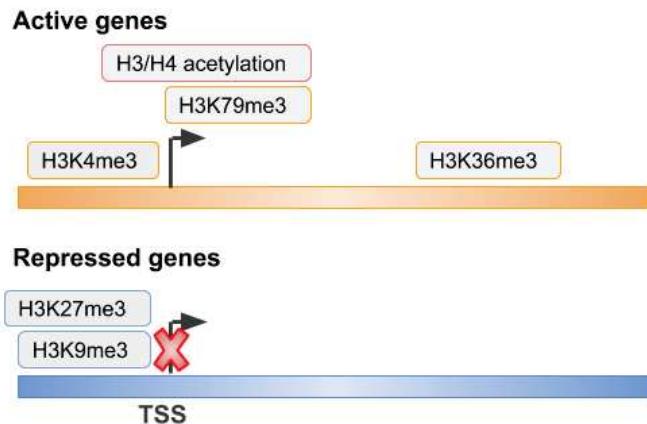


FIGURE 1.3 – **Différentes modifications d'histones.** Représentation schématique de la queue N-terminale des différentes protéines histones et de leurs modifications. [Figure extraite de [86]]

Les modifications d'histones influencent la compaction de la chromatine et son accessibilité et agissent donc sur des fonctions telles que la transcription. La partie décompactée de la chromatine, l'euchromatine, laisse place aux mécanismes de transcription. A l'opposé, la partie compactée de la chromatine, l'hétérochromatine, est caractérisée par des nucléosomes étroitement repliés sur eux mêmes. Certaines marques épigénétiques comme H3K4me3 sont utilisées comme un indice sur l'activation de la transcription. La caractérisation des différentes marques épigénétiques et leur association à un état particulier de la chromatine a conduit les chercheurs

à proposer un code histone qui associerait chaque modification à un état transcriptionnel donné [115], mais ce code a beaucoup été controversé et bien que certaines modifications soient souvent associées à un état précis, elles sont généralement en combinaison rendant le code histone beaucoup plus complexe que sa définition initiale. Ainsi, certaines combinaisons de marques épigénétiques sont souvent associées à certains types de régions : régions réprimées, hétérochromatine et régions transcrites [289, 66, 101] (voir Figure 1.4).



**FIGURE 1.4 – Combinaisons de marques épigénétiques associées à des gènes activés ou réprimés.** En orange (haut), un exemple de gène dont la transcription est active et les marques épigénétiques qui lui sont associées. En bleu (bas), un exemple de gène dont la transcription est inhibée par les marques épigénétiques annotées. [Figure adaptée de [289]]

**Méthylation de l'ADN.** Chez les mammifères, la méthylation de l'ADN correspond à l'ajout d'un groupement methyl ( $CH_3$ ) sur le carbone 5 des cytosines de l'ADN qui précèdent une guanine. On parle ainsi de dinucléotide CpG (cytosine-phosphate-guanine) et la cytosine méthylée est nommée 5-méthyl-cytosine (5mC). En 1948, R.D. Hotchkiss découvre l'existence d'une "cinquième base" de l'ADN qui possède un profil de migration en chromatographie différent de celui des 4 autres bases [103]. La découverte de l'existence de la cytosine méthylée précède même l'identification de la double hélice d'ADN par Watson et Crick. Cette modification chimique est assurée par des enzymes, les ADN méthyl-transférases (DNMT) qui ajoutent le groupement méthyl à partir d'un donneur qui est la S-adénosyl-méthionine (SAM). La présence de ces groupements méthyl est variable d'un gène à l'autre mais également d'un type cellulaire à l'autre, et est généralement associée à une répression transcriptionnelle par inhibition du recrutement des facteurs de transcription (*Transcription factors, TFs*) [118].

Les dinucléotides CpG ne sont pas très abondants dans le génome des mammifères à cause de la désamination spontanée de la 5mC en thymine [37, 108]. De plus, leur distribution le long du génome n'est pas aléatoire : ils se concentrent au niveau de régions de plus de 500 pb nommées îlots CpG et sont notamment enrichis dans la région 5' des gènes [5]. La définition d'un îlot CpG est la suivante : c'est une région d'une longueur de plus de 200 pb avec une proportion de G+C d'au moins 50% et un rapport de CpG sur GpC d'au moins 60% [81]. Dans ces régions riches en CpG, la plupart des dinucléotides CpG sont non-méthylés. Les îlots CpG se trouvent ainsi enrichis au niveau des régions promotrices des gènes et notamment des gènes ubiquitaires comme les gènes associés à la différenciation. A l'inverse, les CpG méthylys sont localisés dans les régions de chromatine compactée et sont associés à la répression des gènes [51].

### 1.1.2 Hiérarchie de la chromatine et interactions

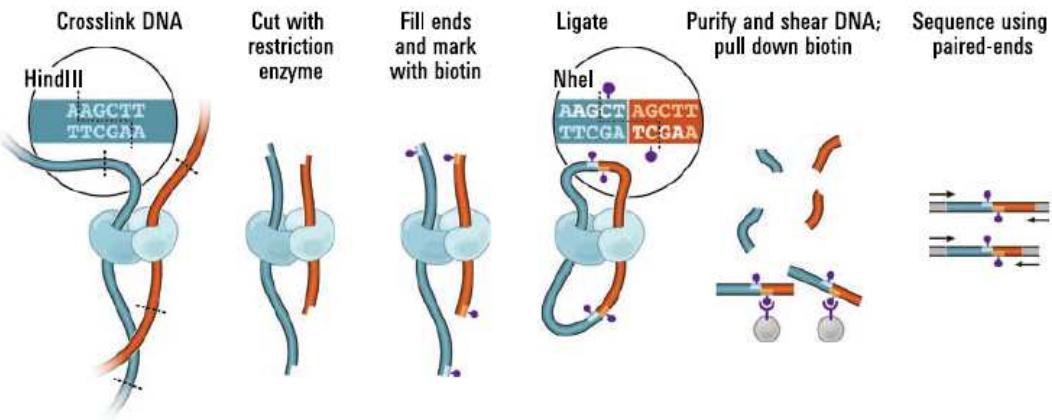
Comme nous l'avons vu ci-dessus, la fibre de chromatine est repliée de manière ordonnée et complexe. De nombreuses méthodes ont été mises en place pour étudier la structure locale de la chromatine et les portions de l'ADN qui sont en interaction. De telles études permettent l'identification des zones actives de l'ADN dans un type cellulaire donné, que ce soit les gènes ou les régions régulatrices associées.

#### Étude des interactions de la chromatine à l'échelle du génome

Un grand nombre de méthodes d'étude des interactions s'appuient sur le principe de capture de la conformation des chromosomes (*Chromosome conformation capture*, 3C) [188] et permettent la détermination de la fréquence avec laquelle un locus de l'ADN est en proximité physique avec un autre locus (entre 10 et 100 nm de distance entre les deux loci). Toutes les méthodes 3C possèdent une première étape de liaison covalente des portions de chromatine qui sont proches physiquement, ensuite la chromatine est fragmentée et les fragments obtenus sont liés pour former une molécule d'ADN hybride unique.

**Hi-C.** La méthode Hi-C (*High chromosome contact map*) est la première adaptation du 3C à l'échelle du génome. Son principe est le suivant : après immunoprécipitation de la chromatine pour figer les interactions entre protéines et ADN, il y a digestion de la chromatine en petits fragments, puis les extrémités de chaque morceau d'ADN en interaction sont marquées par un nucléotide biotinylé (voir Figure 1.5). Après ligation et sélection des fragments marqués par la biotine, ils sont séquencés.

Cette technique permet d'avoir une carte globale des interactions à l'échelle du génome [160].



**FIGURE 1.5 – Principe du Hi-C.** Les cellules sont réticulées avec du formaldéhyde ; les segments de chromatine proches (représentés en bleu et rouge) sont ainsi liés. Les protéines qui lient les fragments d'ADN sont représentées en bleu clair. La chromatine est ensuite coupée par une enzyme de restriction (exemple de HindIII ici) et les extrémités restantes de l'ADN sont marquées par la biotine (points violets). Les extrémités sont ensuite liées créant des molécules chimériques. Enfin, l'ADN est purifié, puis les jonctions biotinylées sont isolées et identifiées par séquençage. [Figure extraite de [160]]

La méthode Hi-C possède une bonne sensibilité pour déterminer les contacts à l'échelle de la mégabase et les interactions définies sont robustes. Cependant, cette méthode n'est pas très adaptée pour capturer les contacts présents à plus fine échelle ( $<40$  kb) [219] comme ceux existants entre les petits éléments régulateurs. La profondeur de séquençage, c'est à dire le nombre de reads total obtenu sur l'échantillon étudié, détermine cette résolution. En exploitant les données de Hi-C, il est possible de décrire l'organisation 3D du génome et d'établir des matrices de contact à large échelle. Ces matrices sont généralement établies par chromosome en le découplant en bins de  $x$  bases et en donnant à chaque paire de bins  $(i, j)$  dans ce chromosome une probabilité de contact  $P(i, j) = P(j, i)$ . Cette probabilité est directement dépendante du nombre de reads qui vont s'aligner sur le bin  $i$  et sur le bin  $j$ . Ces matrices sont souvent représentées sous forme de heatmap [56].

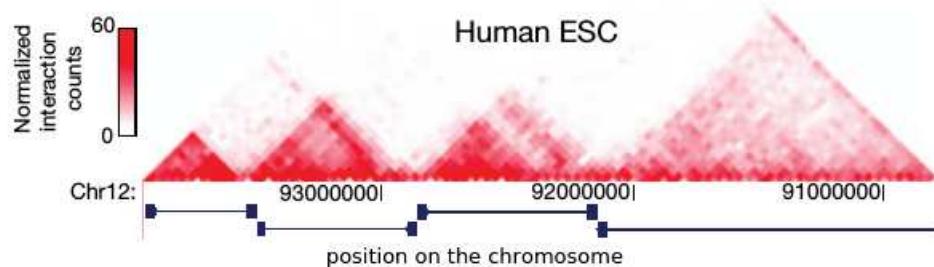
**ChIA-PET.** Une autre méthode utilisée aujourd'hui est le ChIA-PET (*Chromatin Interaction Analysis by Paired-End Tag Sequencing*). Cette technique qui s'appuie sur le principe de l'immuno-précipitation de la chromatine (*Chromatin immunoprecipitation*, ChIP) et sur le 3C a été introduite en 2009 [80, 156]. Ici, une protéine d'intérêt est ciblée, donnant accès aux régions de l'ADN en interaction l'une

avec l'autre, et avec la protéine d'intérêt.

Le ChIA-PET possède ses limites malgré une bonne résolution de l'ordre de la kilobase [80]. Cette technique nécessite de choisir une protéine à étudier, on ne peut pas avoir accès à toutes les interactions simultanément. De plus, la protéine d'intérêt peut être directement liée aux brins d'ADN séquencés, mais peut aussi faire partie d'un complexe. Ces deux situations ne peuvent pas être discriminées. Cette méthode est adaptée à la mise en évidence de réseaux d'interactions pour des facteurs de transcription, des protéines insulatrices qui servent de barrières aux interactions longue distance comme la protéine CTCF ou pour la machinerie de transcription (ARN Polymérase II).

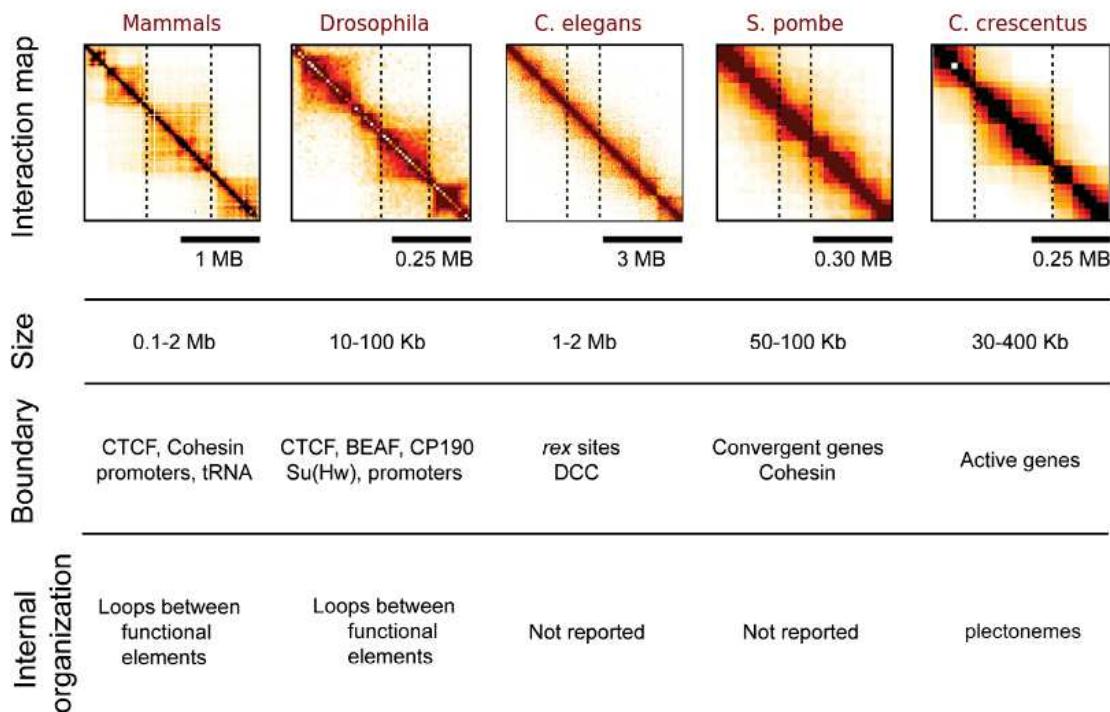
### Les domaines d'association topologique (TADs)

Sur la matrice de contact établie à partir des données de Hi-C, nous pouvons observer des domaines de fortes interactions [56, 61] (voir Figure 1.6). Ces portions correspondent à des domaines physiques occupant un espace défini dans le noyau que l'on appelle domaines d'association topologique (*Topologically associated domains*, TADs). Ainsi, un TAD est une région de l'ordre de la mégabase dans laquelle la probabilité d'interaction est forte. Ils sont définis de telle sorte que la fréquence d'interaction des loci intra-TAD soit beaucoup plus élevée que la fréquence d'interaction inter-TAD. Dixon et al. [61] ont caractérisé en 2012 les TADs chez l'homme et la souris à partir de données de Hi-C d'une résolution de 40 kilobases (kb) et en utilisant un modèle de Markov Caché (*Hidden Markov Model*, HMM) pour définir les frontières des TADs.



**FIGURE 1.6 – Définition des TADs.** Carte des probabilités d'interactions entre les différents sites du chromosome 12 chez l'homme (cellules hESC) et définition des TADs (lignes horizontales bleues). [Figure adaptée de [61]]

Chez l'homme, cette organisation spatiale est stable à travers différents types cellulaires et est également en grande partie conservée chez la souris. Les génomes de ces deux organismes sont composés d'environ 2000 TADs recouvrant 90% du génome [61]. Une organisation similaire a également été observée chez des embryons de drosophiles la même année [238, 104] et chez d'autres organismes entre 2012 et 2015 [55] (voir Figure 1.7).



**FIGURE 1.7 – TADs ou domaines similaires chez plusieurs espèces.** Les heatmaps représentent les probabilités d'interactions obtenues par Hi-C. Pour les mammifères (Mammals), c'est la région d'inactivation du chromosome X chez la souris qui est représentée. Des informations sur la taille des domaines topologiques, sur les éléments enrichis aux frontières et sur l'organisation interne de ces domaines sont présentées. [Figure adaptée de [55]]

Les TADs ont un rôle fonctionnel : leur présence facilite les interactions spatiales au sein de chaque domaine entre deux séquences et isole ces dernières des autres domaines, ce qui les empêche d'interagir avec les séquences des TADs voisins [256, 255]. Une analyse effectuée chez la souris confirme la présence particulièrement fréquente d'interactions entre promoteurs et enhancers au sein des TADs [225].

Les frontières des TADs semblent être en partie définies par des éléments génotypes. En effet, lorsque l'on supprime la frontière entre deux TADs situés au centre d'inactivation du chromosome X (*X chromosome inactivation*, XCI), une fusion partielle des deux TADs adjacents est observée [189]. Plusieurs facteurs connus caractérisent ces frontières : elles sont enrichies en transcription active et en marques

épigénétiques associées à celle-ci comme H3K4me3 et H3K36me3, en sites de liaison pour des protéines insulatrices, notamment CTCF, en gènes de ménages, en cohésine qui est un complexe protéique, mais aussi en petits éléments répétés (SINEs, que nous présenterons dans la sous-partie suivante) [61, 55, 84]. La présence de cohésine est importante pour la formation de boucles et une élimination de cette dernière entraîne une perte de tous les domaines de contacts initialement assurés par la cohésine et CTCF [218, 235, 280]. Les frontières sont également caractérisées par des motifs particuliers et chez la drosophile, le motif M1BP (*Motif 1-binding protein*) semble être celui ayant l'effet de barrière entre TADs le plus significatif. [186] L'existence de ces domaines entraîne des contraintes sur les interactions entre gènes et éléments régulateurs distaux.

### Compartiments génomiques

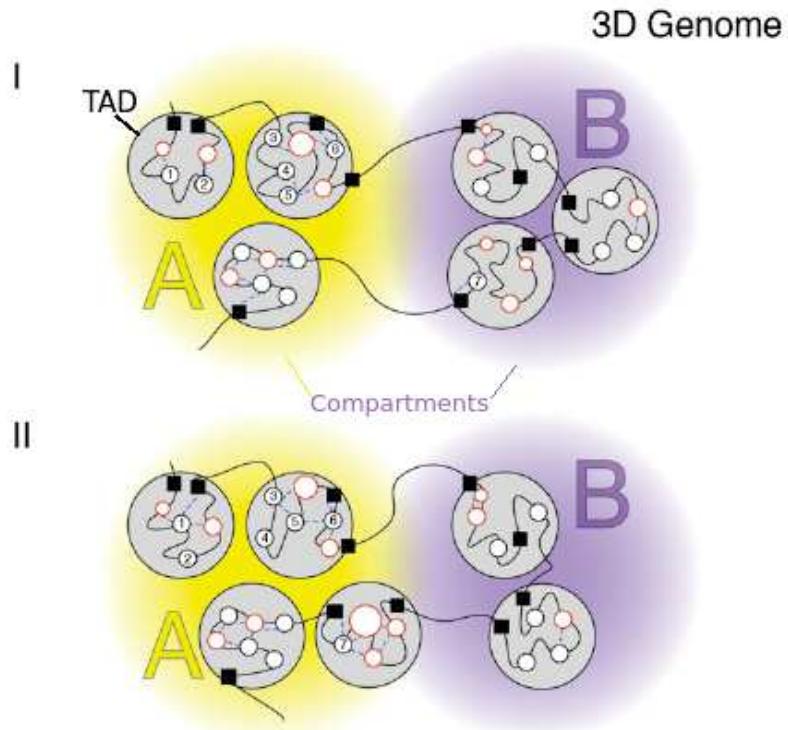
Les TADs, qui varient peu d'un type cellulaire à l'autre, appartiennent à un niveau de régulation plus large. En effet, les matrices d'interactions intra-chromosomales suggèrent une décomposition de chaque chromosome en deux ensembles de loci appelés compartiments A (en majorité transcriptionnellement actif) et B (en majorité inactif) dans lesquels les contacts sont enrichis. Ces compartiments s'alternent le long du génome et ont une taille d'environ 5 Mb chacun [84, 46] (voir Figure 1.8).

La différence entre compartiments A et B n'est pas binaire, c'est à dire que l'on n'observe pas deux états bien distincts entre les deux types de compartiments. Leur séparation corrèle avec de nombreux facteurs de l'activité transcriptionnelle comme l'accessibilité de l'ADN, la densité en gènes, le taux de GC, le temps de réplication et plusieurs marques d'histones. Ce sont ces indicateurs qui permettent une "séparation" en deux compartiments avec le compartiment A largement enrichi en euchromatine [16, 249].

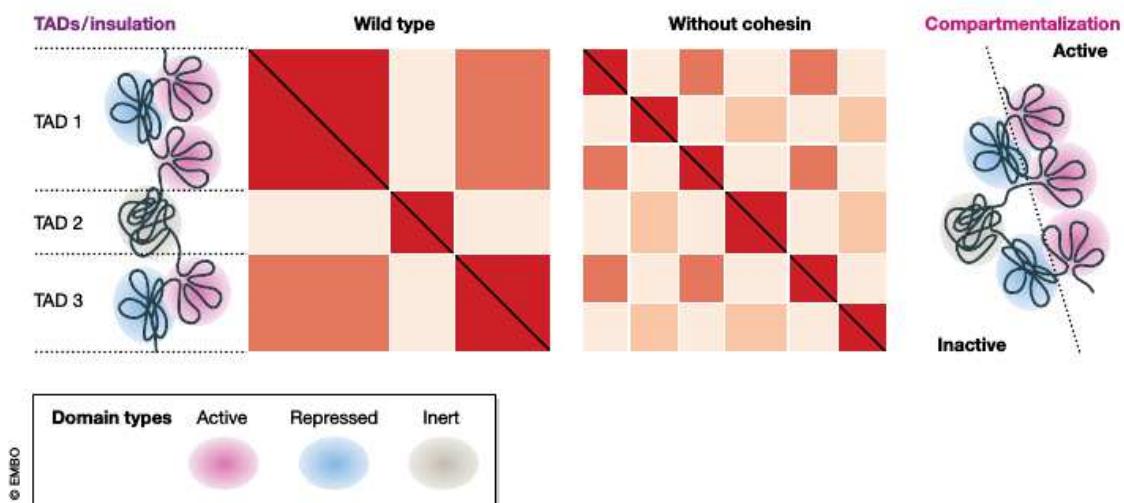
L'arrangement du génome en compartiments génomiques est bien distincte des TADs. Contrairement aux TADs majoritairement conservés à travers différents types cellulaires, les compartiments A et B sont des domaines tissu-spécifiques et dépendent de l'expression des gènes. De plus, ces compartiments s'alternant le long des chromosomes sont bien plus larges que les TADs qui eux peuvent être voisins tout en ayant un statut chromatinien similaire [215] (voir Figure 1.9).

### Les isochores

La découverte des TADs est relativement récente et rendue possible grâce au dé-



**FIGURE 1.8 – Représentation des compartiments A et B.** Les interactions entre loci ont lieu à l'intérieur des TADs (cercles gris) et les TADs sont regroupés en compartiments A et B d'activité similaire. 2 exemples de situations sont présentés : I et II. Dans la situation II, les compartiments changent par rapport à la situation I, en fonction de l'activité différente des gènes et de leurs régulateurs. Seulement les compartiments sont réarrangés mais les TADs ne changent pas. [Figure adaptée de [84]]



**FIGURE 1.9 – TADs versus Compartiments A et B.** Exemple schématisé de heatmap de données de Hi-C révélant les TADs en rouge ("wild type"). Si la cohésine est inhibée ("Without cohesin"), les interactions dominent entre les compartiments en fonction de leur activité transcriptionnelle. [Figure adaptée de [215]]

veloppement des méthodes de capture des interactions de la chromatine à l'échelle du génome. Cette découverte met en évidence une organisation du génome non aléatoire avec la présence de domaines à l'échelle de la mégabase dans lesquels les interactions sont fortes et où les gènes sembleraient être co-régulés. Cependant, l'existence d'une organisation non aléatoire du génome remonte à plusieurs dizaines d'années et il a été observé il y a 60 ans, chez un veau, que le génome possédait une composition nucléotidique très hétérogène en comparaison de la bactérie [178]. Cette composition nucléotidique particulière à l'échelle du génome a ensuite été révélée chez l'homme avec les travaux de G. Bernardi où il décrit l'existence de portions d'ADN de compositions relativement homogènes et correspondant à des contenus en GC distincts [166], nommés isochores [47]. Les isochores peuvent être séparés en deux classes majoritaires : les isochores lourds (*heavy*, H) dont le pourcentage en GC est élevé et les isochores légers (*light*, L) dont le pourcentage en GC est faible.

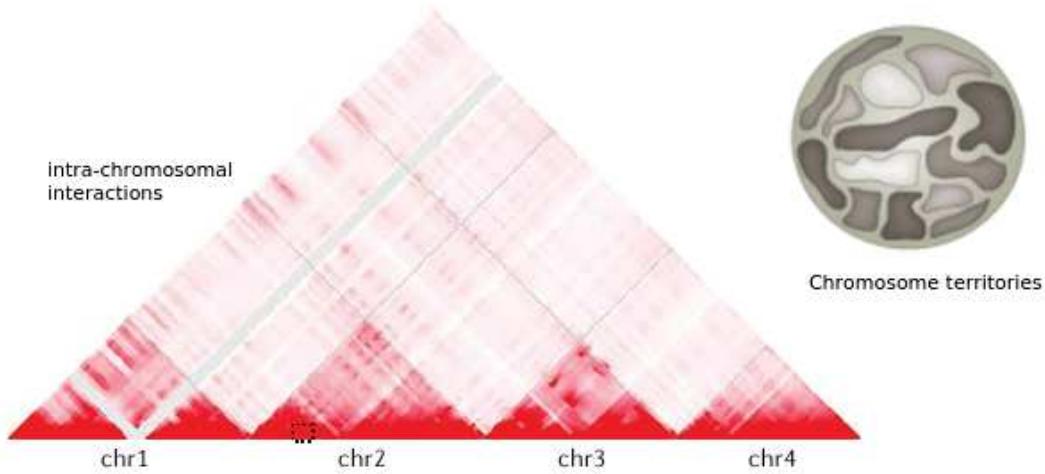
Plus récemment, Jabbari et Bernardi comparent la définition des TADs avec celle des isochores chez l'homme et la souris [111]. Cette comparaison permet de montrer que les isochores aident à définir les TADs. De plus, la conservation observée des TADs entre différents types cellulaires et entre espèces proches comme l'homme et la souris est expliquée par l'évolution conservatrice des isochores. Les domaines chromatiniens correspondant à des isochores GC-riches possèdent des interactions chromosomales plus localisées.

Bien que la découverte des isochores montre une composition génomique non homogène, le parallèle avec l'architecture particulière de la chromatine et la notion de TAD n'avait pas encore été faite jusqu'à cette dernière décennie. Cependant, un découpage du génome en domaines fonctionnels et une organisation structurée au sein du noyau avaient déjà été observés dans les années 80.

### Les territoires chromosomiques

Au niveau supérieur de l'organisation 3D de la chromatine, les interactions en *trans*, c'est à dire entre chromosomes, sont rares (voir Figure 1.10).

Les domaines d'un chromosome interagissent plutôt avec d'autres domaines du même chromosome, entraînant la formation de territoires chromosomiques (*Chromosomal territories*, CT). Les chromosomes riches en gènes sont préférentiellement localisés au cœur du noyau tandis que les chromosomes pauvres en gènes sont plutôt localisés à proximité de la membrane nucléaire [16].

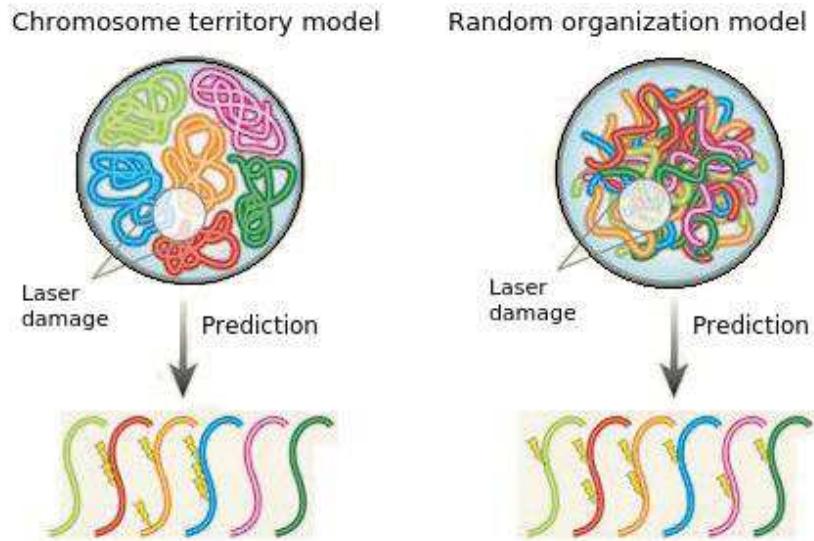


**FIGURE 1.10 – Interactions entre chromosomes et représentation schématique des territoires chromosomiques.** Les interactions inter-chromosomes sont rares et les chromosomes occupent des territoires distincts à l'intérieur du noyau (cercle gris). Ils sont schématisés par des formes irrégulières en gris plus ou moins foncé et sont appelés territoires chromosomiques. Les données de Hi-C représentées sur la heatmap en rouge proviennent de la lignée cellulaire humaine GM12878. [Figure adaptée de [16]]

**Découverte des territoires chromosomiques.** L'existence des territoires chromosomiques a été démontrée expérimentalement dans les années 80 par les frères Thomas et Christoph Cremer. Par une expérience de micro irradiation UV permettant d'endommager localement l'ADN, ils proposent un modèle qui, selon les parties du génome touchées, confirmerait l'hypothèse d'un arrangement chromosomique spécifique (petite sous-partie des chromosomes touchée) ou à l'inverse d'une répartition aléatoire du génome au sein du noyau (beaucoup de chromosomes touchés) (voir Figure 1.11) [45]. L'irradiation n'a causé des lésions que sur certains chromosomes soutenant l'hypothèse d'un arrangement spécifique au sein du noyau [293].

**Les territoires chromosomiques aujourd'hui.** Ces territoires chromosomiques ont pu être ensuite visualisés par imagerie d'hybridation *in situ* fluorescente (*Fluorescence in situ hybridization*, FISH) via des sondes et des marqueurs fluorescents [159, 205] et leur agencement les uns par rapport aux autres a été analysé. Ils ont enfin été validés par des données de Hi-C montrant que la majorité des interactions s'effectuent en *cis* et non en *trans* [160]. En effet, 2 loci d'un même chromosome, même s'ils sont séparés de plus de 200 mégabases (Mb), interagissent plus fréquemment que 2 loci proches physiquement mais sur 2 chromosomes différents [18, 184].

Les territoires chromosomiques n'existent que dans certaines cellules eucaryotes issues des organismes supérieurs comme l'homme ou la souris ; par contre chez la



**FIGURE 1.11 – Découverte de l’existence des territoires chromosomiques.** Deux hypothèses d’arrangement des chromosomes sont présentées : à gauche le modèle de territoires chromosomiques bien distincts et à droite une organisation aléatoire des chromosomes dans le noyau. L’effet que le laser aurait sur ces deux modèles est également schématisé par des petits éclairs jaunes. [Figure adaptée de [176]]

levure *Saccharomyces cerevisiae*, les chromosomes ne semblent pas être arrangés.

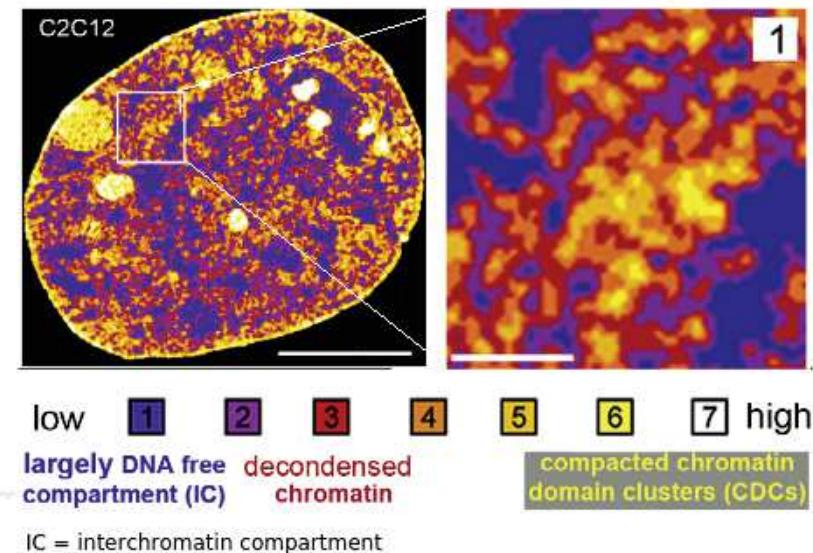
### 1.1.3 Compartimentation spatio-temporelle du noyau

Le noyau contenant la chromatine est délimité par une double membrane lipidique appelée enveloppe nucléaire. Les progrès des outils de microscopie et de séquençage de l’ADN permettent aujourd’hui d’étudier la relation entre la chromatine et les autres composés nucléaires ainsi que leur dynamique. On distingue les compartiments nucléaires inactifs, qui comprennent la chromatine compacte et inactive, des compartiments actifs formés par la chromatine transcriptionnellement active et du compartiment ”interchromatine” (IC) où l’on ne retrouve que très peu d’ADN (voir Figure 1.12).

Ce dernier compartiment est connecté aux pores nucléaires qui permettent l’entrée et la sortie des particules du noyau vers le cytoplasme et inversement [46].

#### Enveloppe nucléaire et lamina

La face interne de l’enveloppe nucléaire, en contact direct avec la chromatine et autres particules contenues dans le noyau, est recouverte d’une couche filamenteuse protéique intermédiaire appelée la lamina. Des études ont établi une corrélation



**FIGURE 1.12 – Densité de la chromatine au sein du noyau par microscopie super-résolution (3D-SIM).** Les images proviennent du noyau d'une cellule de souris C2C12. On observe en bleu foncé la région "inter-chromatine" à densité en chromatine très basse. En violet/rouge la chromatine est décondensée et active. Enfin en jaune jusqu'au blanc, la chromatine est fortement compactée. Cette coloration dépend du marquage par le DAPI qui est une molécule fluorescente capable de se lier aux bases Adénine (A) et Thymine (T) de l'ADN. [Figure adaptée de [46]]

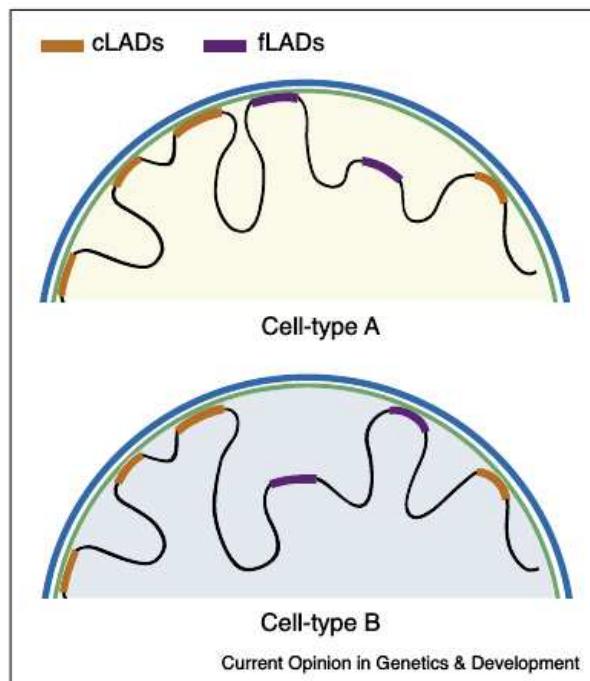
entre le positionnement des gènes à la périphérie nucléaire et leur inactivité transcriptionnelle [206, 245]. Les protéines qui composent la lamina sont des lamines qui se présentent sous la forme d'un réseau fibrillaire. On distingue les lamines B1 et B2 présentes dans le noyau de la majorité des vertébrés, des lamines A et C qui apparaissent pendant ou après la différenciation cellulaire [221]. La lamina possède un rôle important dans l'organisation de la chromatine et des pores nucléaires. En effet, les lamines peuvent se lier à la chromatine via le dimère d'histones H2A/H2B et peuvent avoir un rôle dans la régulation de l'expression des gènes [171].

### Les domaines associés à la lamina (LADs)

Il existe des régions génomiques en contact avec la lamina nucléaire que l'on appelle domaines associés à la lamina (*Lamina-associated domains*, LADs). Ils ont été initialement observés grâce à la technologie d'identification par l'ADN méthyl-transferase adénine (DamID) [268], une méthode dérivée du ChIP. L'ADN méthyl-transférase couplée à une protéine d'intérêt, ici une protéine de la lamina, entraîne la méthylation de l'adénine contenue dans les séquences GATC des régions voisines de l'ADN en contact avec ces protéines de la lamina [204]. Cette technique est couplée à une visualisation par microscopie ou à du séquençage à l'échelle du génome

(*Genome-wide*) pour pouvoir définir les LADs. Les LADs ont été définis pour de nombreux organismes comme *D. melanogaster*, *C. elegans*, et plusieurs lignées cellulaires humaines [89, 107, 204, 202]. Chez l'humain et la souris on compte plus de 1000 LADs variant de 10 kb à une 10<sup>aine</sup> de Mb chacun et distribués le long de tous les chromosomes. Ils couvrent plus d'un tiers du génome humain [267].

La majorité des gènes situés dans des LADs sont silencieux ou exprimés à un taux très faible [89, 202] et on retrouve parfois dans ces domaines de grandes portions de plus d'1 Mb sans aucune activité génique. Les LADs sont enrichis en marques histones H3K9me2 et H3K9me3 caractéristiques de l'hétérochromatine [89]. La lamina et ses protéines associées se lient ainsi à la chromatine mais aussi à des répresseurs transcriptionnels [68]. Des études des interactions de la lamina pendant la différenciation cellulaire ont permis de différencier 2 types de LADs [282, 202, 179] : les LADs constitutifs (*constitutive LADs*, cLADs) qui sont conservés entre différents types cellulaires et les LADs facultatifs (*facultative LADs*, fLADs) qui sont associés seulement à certains types cellulaires (voir Figure 1.13). La position des LADs constitutifs est conservée entre l'homme et la souris.



**FIGURE 1.13 – Interactions entre la lamina et les LADs constitutifs et facultatifs.** Les LADs constitutifs (cLADs) en marron et facultatifs (fLADs) en violet sont représentés dans deux types cellulaires exemples A et B. [Figure adaptée de [282]]

## Les corps polycomb

Les protéines du groupe Polycomb (*Polycomb group proteins*, PcG) ont été découvertes avec les protéines trithorax (*Trithorax group proteins*, TrxG) chez la drosophile et font partie d'un système de mémoire cellulaire complexe conservé dans la plupart des organismes vivants. Les PcG et TrxG ont un rôle antagoniste : les PcG peuvent moduler la chromatine pour maintenir la répression de certains gènes alors que les TrxG maintiennent leur expression. Ces protéines sont notamment connues pour leur rôle crucial dans le maintien de l'activité des gènes homéotiques qui sont les gènes contrôlant la mise en place des organes lors du développement d'un organisme [233].

## Conclusion

Les informations que nous venons de voir nous montrent une architecture hiérarchique du génome au sein du noyau de la cellule. Chaque chromosome occupe un espace délimité que l'on nomme territoire chromosomique. Au sein de ces territoires, on distingue les compartiments A (actif) et B (inactif) qui s'alternent. Ces compartiments regroupent les TADs, domaines fonctionnels dans lesquels les interactions entre éléments régulateurs sont favorisées. A l'intérieur des TADs, les interactions entre les gènes et leurs éléments régulateurs distaux s'effectuent via la formation de boucles. Ces boucles sont stabilisées par des contacts protéine/protéine comprenant les protéines architecturales (CTCF, cohésine) et des éléments régulateurs comme les facteurs de transcription. Les frontières des TADs sont aussi caractérisées par des motifs particuliers.

Dans la partie suivante nous allons voir comment s'organisent ces régions actives et inactives le long du génome et les éléments qu'elles contiennent.

## 1.2 Segmentation du génome

Depuis le séquençage complet du génome humain, plusieurs projets ont émergé pour déchiffrer cette succession de lettres et annoter les parties fonctionnelles et non fonctionnelles du génome. L'annotation systématique de tous les éléments composant le génome est rendue possible grâce aux outils bioinformatiques mais se base aussi sur les annotations connues par expérience, lors de travaux antérieurs.

### 1.2.1 Gènes et annotations

Les unités transcriptionnelles sont les gènes qui sont des portions de l'ADN recopiées sous forme de molécules d'ARN puis pouvant être traduites en protéine. Ils sont porteurs d'un caractère héréditaire précis, ce qui leur permet d'assurer une fonction spécifique dans la cellule. Nous allons voir ci-dessous comment ils sont annotés et classifiés.

#### Comparaison des annotations

L'annotation des gènes est indispensable pour pouvoir naviguer dans le génome et l'étudier. Cependant, il n'existe pas de standard universel d'annotation et de nombreux projets existent et fournissent des informations qui peuvent être similaires ou complémentaires. La première étape de l'annotation du génome humain consiste à repérer les gènes dans la séquence d'ADN brute. La seconde étape est l'annotation fonctionnelle qui permet d'associer une fonction biologique à chacun des gènes ou autres éléments du génome. L'annotation s'appuie sur les études expérimentales généralement effectuées sur des organismes modèles plus faciles à étudier que l'homme. En effet, certains gènes sont très conservés chez toutes les espèces ce qui permet de les transposer facilement chez les organismes supérieurs.

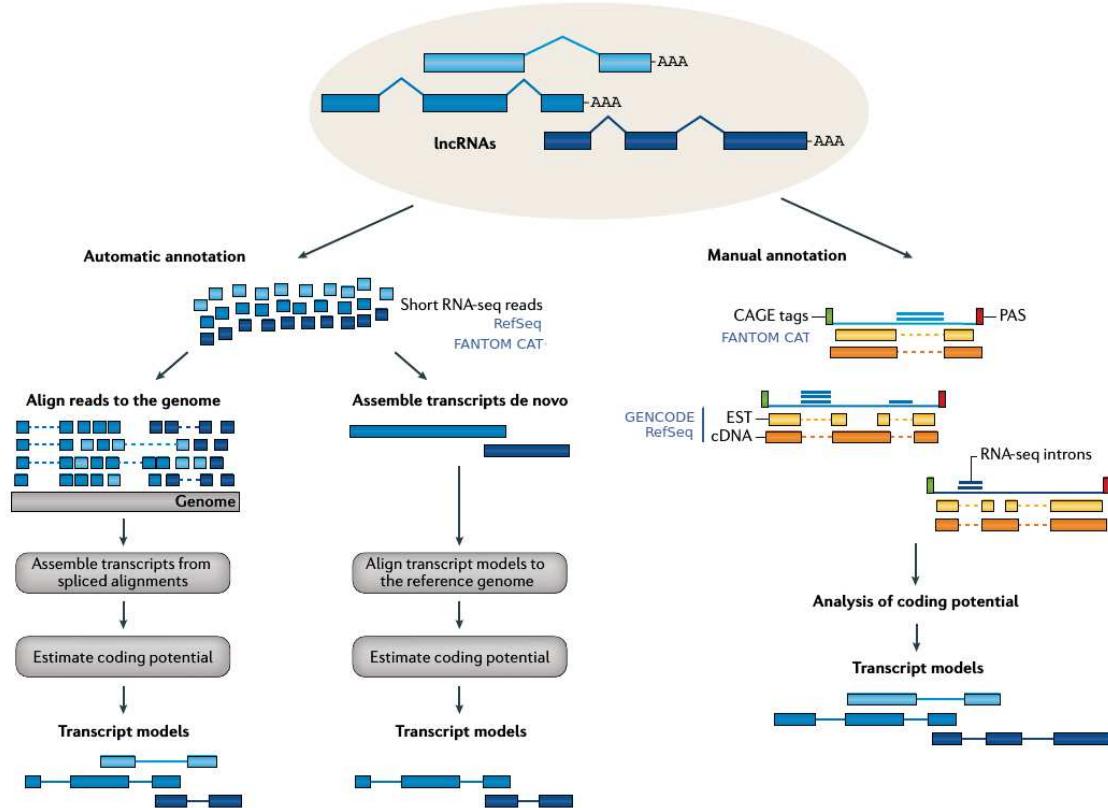
**Annotations de référence.** Ensembl [287] fait partie des nombreux projets d'annotation existants et représente une source importante d'informations sur les gènes, leur localisation dans le génome, leur fonction et les variants qu'ils peuvent contenir. Les annotations fournies par Ensembl sont essentiellement des annotations automatiques et pour chaque gène avec un identifiant unique (ENSG), un ou plusieurs transcrits sont associés (ENST).

Les transcrits eux mêmes sont aussi annotés, et leur annotation s'appuie sur des évidences expérimentales [277] et sur l'ARN messager (ARNm) ou séquences protéiques disponibles dans des databases publiques comme UniProt [38]. Si les

transcrits annotés dans Ensembl correspondent à une entrée dans une base de données protéiques, ils sont catégorisés comme connus ("known"). RefSeq (NCBI) [211] fait aussi partie des bases de données fournissant les séquences nucléotidiques et protéiques des gènes de nombreuses espèces avec leurs caractéristiques correspondantes et des annotations bibliographiques. L'annotation fonctionnelle de ces transcrits se base également sur des résultats expérimentaux et/ou sur des méthodes prédictives. De nombreux projets travaillent sur cette partie de l'annotation dont RefSeq, GENCODE [95], FANTOM [77] et UCSC [180]. L'annotation fournie par GENCODE est adoptée par la plupart des projets d'étude de la dynamique du génome comme ENCODE (ENCylopedia Of DNA Elements) [35] pour qui elle a originellement été créée, le projet GTEx (Genotype-Tissue Expression) [36] ou encore Epigenome Roadmap [224]. Aujourd'hui un total de 58,381 gènes sont annotés dans GENCODE (v28) dont 19,901 gènes codants pour des protéines correspondant à 82,335 transcrits différents.

**Annotations de nouvelles classes de gènes.** Actuellement, les projets d'annotation s'intéressent beaucoup à la partie non-codante du génome, c'est à dire ne codant pas de protéine, où de nombreuses unités fonctionnelles sont découvertes. La classe la plus représentée est celle des longs ARNs non codants (*long non coding RNA*, lncRNA), mais en comparaison aux gènes codants pour des protéines, ils ne sont pas faciles à annoter pour plusieurs raisons : ils sont peu exprimés, leur fonction n'est pas bien caractérisée et ils sont faiblement conservés au cours de l'évolution. Ainsi, une grande diversité de méthodes sont exploitées pour les annoter [264]. Comme nous pouvons le voir sur la Figure 1.14, différentes stratégies sont adoptées pour annoter les lncRNA et on distingue l'annotation automatique en exploitant les petits reads issus du séquençage RNA-seq de l'annotation manuelle.

L'annotation établie par FANTOM pour les lncRNA intègre des données de GENCODE [95], Human BodyMap, miTranscriptome [110], ENCODE [35] ainsi que leurs propres données de RNA-seq et de CAGE (*Cap Analysis of Gene Expression*, CAGE) [102], ce qui la rend robuste. Les signaux de CAGEs marquent les sites de départ de la transcription (*Transcription start sites*, TSSs) des transcrits et leur utilisation pour annoter les lncRNA leur permet d'obtenir des transcrits plus complets au niveau de l'extrémité 5' en comparaison à d'autres annotations. Avec le projet FANTOM5, le consortium a annoté 27,919 lncRNA, ce qui est largement supérieur au nombre de loci déterminés par GENCODE v28 (15,779). Les analyses effectuées sur les lncRNA lors de ma thèse s'appuieront sur l'annotation établie par



**FIGURE 1.14 – Stratégies d’annotation du génome.** Les annotations automatiques sont souvent établies à partir de données de RNA-seq et peuvent suivre 2 stratégies : soit les reads obtenus sont directement alignés sur le génome de référence pour trouver les éventuels sites d’épissage alternatif et caractériser chaque transcrit (*mapping*), soit les reads sont assemblés *de novo*, c’est à dire qu’ils sont regroupés en fragments de taille plus importante par chevauchement pour retrouver les transcrits (assemblage), puis les transcrits sont alignés sur le génome de référence. Les annotations manuelles sont basées sur plusieurs sources de données : l’ADN complémentaire (cDNA) et les marqueurs de séquences exprimées (EST) couplés à des données de RNaseq/CAGE. [Figure adaptée de [264]]

FANTOM, consortium auquel le laboratoire appartient.

### Biotype des gènes

Pour les différencier, les gènes sont classés en fonction de leur biotype. Les classes majeures utilisées par ENCODE et communes à la plupart des projets d’annotations sont listées ci-dessous :

- gènes codants pour des protéines qui contiennent un cadre ouvert de lecture (*Open reading frame*, ORF).
- longs ARNs non-codants qui sont divisés en plusieurs sous-classes en fonction de leur localisation dans le génome par rapport aux gènes codants pour des protéines connues. On distingue par exemple les lncRNA intergéniques des

lncRNA introniques.

- petits ARNs non codants divisés en plusieurs sous-classes dont les plus connues sont les ARNs ribosomiques qui sont les principaux constituants des ribosomes, les ARNs de transfert qui véhiculent les acides aminés lors de la traduction et les micro ARNs qui sont de petits ARNs non codants d'une 20<sup>aine</sup> de nucléotides pouvant contrôler la stabilité et la traduction des ARNs.
- pseudogènes qui sont similaires aux gènes codants au niveau de leur structure mais ils contiennent un décalage dans le cadre de lecture ou un codon stop prématué qui ne leur permet pas de produire de protéine.

Ces grandes classes de gènes sont séparées en sous-classes pouvant varier d'une annotation à l'autre. La majorité des petits ARNs non-codants sont annotés par alignement de séquences contre la base de données RFAM [88] qui recense un nombre important de séquences et d'informations sur les différentes familles d'ARNs.

**Gènes de petits ARNs non-codants.** Les premiers ARNs participant à la régulation de l'expression des gènes ont été découverts dans les années 50 et sont les ARNs ribosomiques (ARNr) et les ARNs de transfert (ARNt). L'ARNr est un composant essentiel des ribosomes et il est indispensable pour la synthèse de protéines [73]. L'ARNt est un ARN d'une petite centaine de nucléotides et indispensable pour la traduction d'un ARNm en protéine [241]. Ils servent d'intermédiaire entre les codons et les acides aminés qui composent la protéine finale. Pour chaque codon, un ARNt transportant l'acide aminé correspondant est associé. Une autre classe d'ARNs non codants importante est celle des petits ARNs nucléolaires (*small nucleolar RNAs*, snoRNA). Situés dans le nucléole, il jouent un rôle dans la maturation des précurseurs d'ARNr en ARNr matures et, associés à d'autres protéines, ils forment des petites ribonucléoprotéines nucléolaires (*small nucleolar Ribonucleoprotein*, snoRNPs) [6]. Ils sont souvent localisés au sein des introns des gènes codants ou non-codants et pourraient être impliqués dans certains cancers [251].

Au début des années 80, l'idée d'une implication de petits ARNs nucléaires (*small nuclear RNA*, snRNA) dans l'épissage, qui est un processus de maturation de l'ARNm, émerge. Ces ARNs sont toujours associés à des protéines ; on les trouve sous forme de complexes nommés petites ribonucléoprotéines nucléaires (*small nuclear Ribonucleoprotein*, snRNP). Les classes majoritaires de ces petits ARNs sont impliquées dans la reconnaissance des introns lors de l'épissage. D'autres snRNA sont impliqués dans des processus comme la maturation des ARNm codants pour

les histones ou comme l’assemblage des ribosomes [169].

Les microARNs sont de petits ARNs d’une 20<sup>aïne</sup> de nucléotides et ils sont présents dans la plupart des organismes vivants. Le premier microARN a été découvert chez le nématode *C. elegans* en 1993 [150]. Ces microARNs peuvent être partiellement ou parfaitement complémentaires à une partie de la région 3’UTR (3’ *Untranslated transcribed region*) de l’ARNm d’un gène cible. Après reconnaissance de l’ARNm cible, une inhibition de la traduction ou une dégradation de l’ARNm peut survenir. Leur expression est associée à des dérégulations importantes dans les cancers [161]. L’annotation des précurseurs de microARNs provient de la base de données miRBase [142].

**Gènes de longs ARNs non-codants.** Les lncRNA sont des ARNs d’une longueur supérieure à 200 nucléotides (nt) transcrits par l’ARN Polymérase II (ARN Pol II). A priori, ils ne possèdent pas de capacité codante, mais comme pour les ARNm, ils sont généralement coiffés en 5’, poly-adénylés en 3’ et possèdent une structure exons/introns que nous décrirons dans la partie suivante. Les lncRNA sont souvent classés en fonction de leur localisation dans le génome, par rapport aux gènes connus codants pour des protéines [253, 217]. FANTOM5 sépare ainsi les lncRNA en 4 classes : les lncRNA intergéniques, les lncRNA divergents, les lncRNA sens-introniques et les lncRNA antisens. La classification de FANTOM5 se base également sur la sensibilité à la DNase I, qui reflète l’ouverture de la chromatine, de la région autour des TSSs des lncRNA pour les catégoriser, permettant d’associer une partie des lncRNA à des sites de transcription active comme les promoteurs et enhancers qui seront présentés dans la partie suivante (voir Figure 1.15).

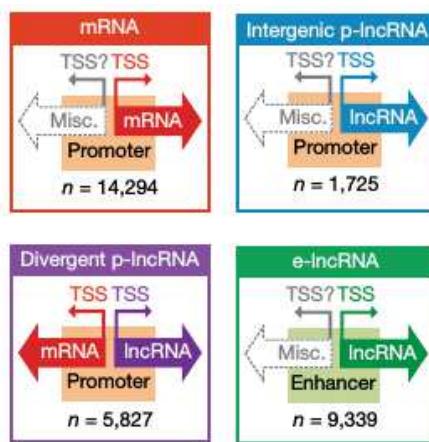


FIGURE 1.15 – Catégories de lncRNA établies par fANTOM. Le nom et le nombre de lncRNA associés sont indiqués pour chacune des 4 classes. [Figure extraite de [102]]

## 1.2.2 Différentes classes de gènes et leur organisation

La plupart des gènes du génome humain possèdent une structure exons/introns. Les exons (initialement étant la contraction d’"*expressed region*") sont généralement conservés dans le transcrit mature et contiennent, dans le cas des gènes codants, les séquences codantes pour des protéines. Les introns ("*intragenic region*") sont des séquences non-codantes qui séparent les exons les uns des autres et qui sont généralement retirés des transcrits lors de leur maturation, par un processus d'épissage.

### Structure exons/introns

**Exons.** Les ARNs immatures appelés pré-ARNs peuvent contenir à la fois introns et exons. Ils sont soumis, après transcription, à un procédé d'épissage pendant lequel les introns sont excisés. Lors de l'épissage, les exons peuvent être conservés dans l'ARN mais ils peuvent aussi être éliminés et, à partir d'un même transcrit immature, de nombreux ARNs différents sont produits. C'est ce qu'on appelle l'épissage alternatif. Une grande majorité d'exons semblent être alternativement épissés [195, 272].

**Introns.** Les introns ont été découverts en 1977 par Phillip Allen Sharp et Richard Roberts qui montrent que le gène codant pour l'ovalbumine, protéine du blanc d'oeuf de la poule, contient des séquences dont on ne retrouve pas les correspondances en acides aminés dans la protéine. Cette découverte leur a valu un prix Nobel en 1993. Les introns sont définis comme des segments de gènes transcrits en ARN mais retirés par un processus d'épissage lors de la maturation de ce dernier. Ils permettent de contribuer à la diversité des transcrits produits à partir d'un même gène. Les introns représentent la moitié du génome humain (UCSC, GCRh38) et leur taille varie de quelques dizaines de nucléotides à la mégabase. Le plus petit intron dont l'épissage a été montré expérimentalement possède une longueur de 43 nt [230]. Les introns des gènes codants ont une longueur médiane d'1,5 kb et les introns des gènes non-codants de 1,7 kb.

Les introns ont été séparés en 4 groupes selon leur mécanisme d'épissage. Les plus répandus sont épissés par le spliceosome [223]. Ces introns présents dans le pré-ARN sont caractérisés par des séquences spécifiques : premièrement le site donneur en 5' et le site accepteur en 3' localisés aux frontières entre introns et exons et qui sont reconnus lors de l'épissage, deuxièmement une boîte de branchement sur laquelle l'extrémité 5' est reliée lors de l'épissage pour former une boucle ayant la

forme d'un lasso [288]. La séquence intronique retirée lors de l'épissage du pré-ARN commence en 5' par le dinucléotide GU et fini en 3' par le dinucléotide AG (voir Figure 1.16). Les séquences consensus de ces sites sont critiques et changer une base des nucléotides conservés entraîne une inhibition de l'épissage [226].

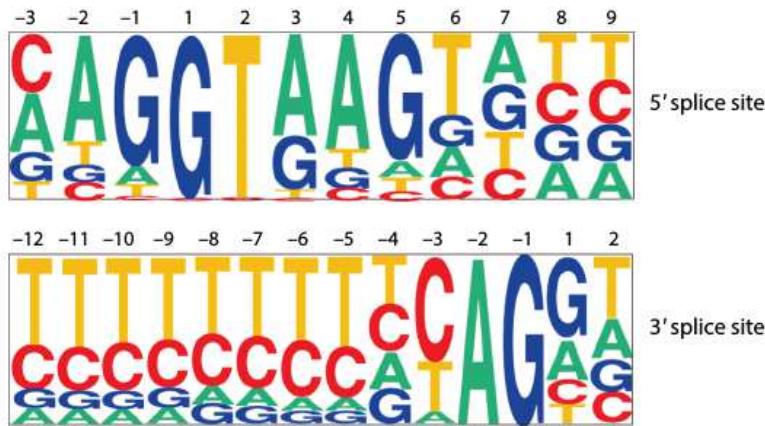


FIGURE 1.16 – Logo des sites accepteur en 5' et donneur en 3' de l'épissage des introns. Pour chacun des deux motifs, deux nucléotides sont très conservés. [Figure adaptée de [152]]

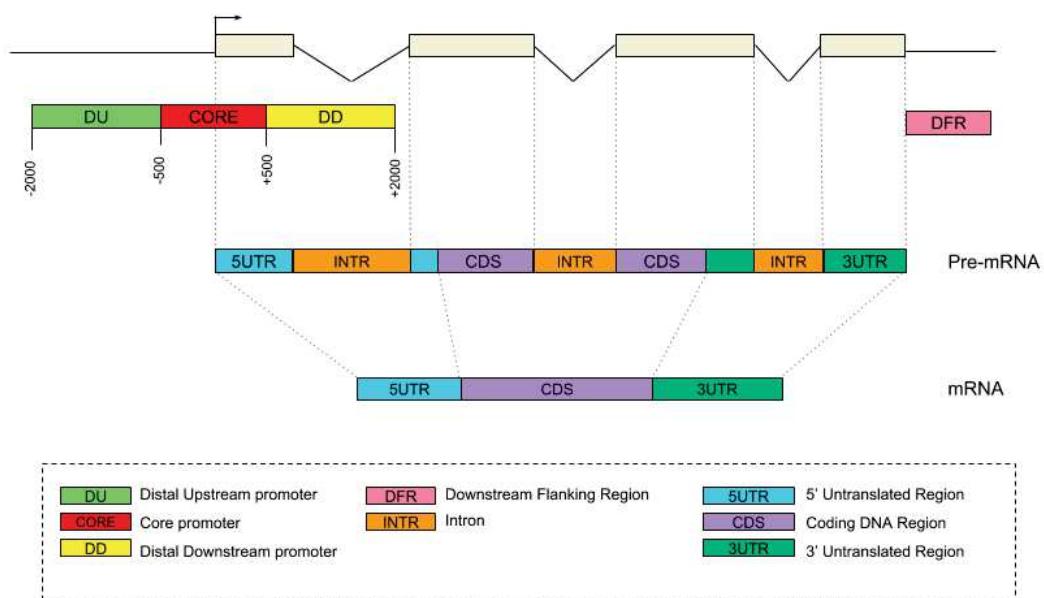
L'extrémité 3' des introns peut contenir, en amont du dinucléotide AG, une région riche en pyrimidine et particulièrement en T appelée *polypyrimidine tract*. Cette région favorise l'assemblage du splicéosome. Des protéines comme la *polypyrimidine tract-binding protein* (PTB) reconnaissent ce motif. La deuxième classe d'introns regroupe les introns des ARNs de transfert qui sont épissés via des protéines nucléaires spécifiques. Les deux dernières classes sont les groupes I et II où les introns catalysent leur propre épissage, sans l'aide de protéines [27].

Les introns ont longtemps été considérés comme des éléments génétiques "égoïstes" [165] qui ont envahi les gènes des génomes eucaryotes. Un débat sur leur origine persiste encore aujourd'hui et deux hypothèses s'opposent : d'un côté la théorie du "intron-early" qui fait l'hypothèse d'une existence très ancienne des introns, ce qui faciliterait la recombinaison des gènes et la diversité des transcrits produits ; d'un autre côté la théorie du "intron-late", soutenant l'idée d'une apparition des introns chez les eucaryotes uniquement et d'une accumulation au cours de l'évolution [140]. En effet, les introns sont présents seulement chez les eucaryotes, avec une proportion plus importante chez les primates que chez la levure, la drosophile ou encore *C. elegans*. Chez les espèces proches, il existe une grande conservation de séquence entre les introns homologues suggérant des contraintes fonctionnelles au cours de l'évolution [94]. Aujourd'hui, de nombreuses fonctions ont été associées aux introns

et ils sont impliqués dans différentes étapes allant de la transcription à la traduction pour les gènes codants. Les séquences introniques influencent ainsi l'épissage alternatif [247], elles peuvent amplifier l'expression des gènes [242], contrôler le transport de l'ARN [265] et les introns des extrémités 5' et 3' des gènes affectent la dégradation des ARNs non-sens [153]. Les introns sont également impliqués dans diverses fonctions de façon indirecte : la longueur des introns est importante dans l'évolution des génomes et ils peuvent contenir de nombreux gènes non-codants [116].

### Organisation spécifique des gènes codants pour des protéines

Chaque gène codant pour une protéine peut donner naissance à une ou plusieurs protéines selon sa complexité. Les gènes sont ainsi découpés en plusieurs parties (voir Figure 1.17) ayant chacune leur rôle dans le processus de création d'une protéine fonctionnelle à partir de la matrice ADN.



**FIGURE 1.17 – Organisation d'un gène exemple.** Un gène est découpé en plusieurs régions. Le promoteur peut être défini de différentes façons. Quand on parle de "core promoter" (rouge), on définit généralement une région réduite autour du TSS qui contient la majorité des sites de fixation pour le recrutement de l'ARN polymérase. On peut également trouver des motifs spécifiques autour du TSS mais dans des régions un peu plus éloignées (promoteurs distaux en vert clair et jaune). Le cœur du gène est séparé en introns et exons (rectangles beiges). Les exons contiennent la partie codante de la séquence (violet) et les UTRs (bleu et vert foncé).

**Exons et partie codante.** Les exons des gènes codants contiennent les parties codantes de l'ADN et des transcrits produits à partir de cet ADN qui seront traduites en protéine. La composition nucléotidique de la partie codante du gène

(*Coding sequence*, CDS) est très contrainte. En effet, une protéine est une succession d'acide-aminés (aa) encodés par les codons qui, au niveau de l'ADN, sont représentés par des triplets de nucléotides. La partie codante contient donc des triplets de nucléotides non aléatoires dont au moins un codon initiateur marquant le début de la traduction (ATG) et un codon stop (TAA, TAG, TGA) [25]. De plus, il existe seulement 20 aa pour 64 triplets de nucléotides différents, plusieurs codons peuvent donc correspondre à un même aa. On dit que le code génétique est dégénéré.

**UTRs.** Les régions non traduites situées aux extrémités 3' et 5' d'un gène sont les UTRs (*Untranslated transcribed region*, UTR). On distingue les 5'UTRs à l'extrémité 5' des 3'UTRs à l'extrémité 3'. Le 5'UTR est défini comme la région située entre le TSS et le codon d'initiation de la traduction (ATG au niveau de l'ADN, AUG au niveau de l'ARN transcrit). Il contient des éléments régulateurs et joue un rôle important dans le contrôle de l'expression des gènes. Lors de la transcription, l'extrémité 5' des ARNm est coiffée ce qui permet de le protéger et de le stabiliser. La proximité de la coiffe au codon initiateur, la composition nucléotidique et la structure secondaire du 5'UTR influencent l'initiation de la traduction [203]. L'étude de l'impact de la composition de la séquence des 5'UTRs sur l'expression des gènes chez la levure montre une influence forte sur le niveau de protéines produites, avec une contribution majeure des quelques bases précédant le codon initiateur [63]. Les 5'UTRs peuvent également contenir des cadres de lecture amont ou uORF (*Upstream open reading frame*, uORF). Les uORFs possèdent leur propre codon initiateur (*Upstream AUG*, uAUG) en amont du site principal associé au gène et leur propre codon stop [24]. Chez l'homme, ils sont présents dans environ 50% des 5'UTRs et peuvent être traduits. Leur présence corrèle avec une diminution de l'expression à la fois au niveau de la quantité d'ARNs produite et de protéines.

La région 3'UTR est localisée en aval du site de terminaison de la traduction. Cette région est impliquée dans de nombreux processus de régulation comme le clivage de l'extrémité 3' de l'ARN transcrit, la stabilité de l'ARNm et sa polyadénylation, ou encore la traduction et la localisation de l'ARNm [172]. Chez l'homme, une grande portion des gènes codants pour des protéines utilisent des sites de clivage et de polyadénylation alternatifs pour générer des 3'UTRs alternatifs ayant un impact sur le devenir de l'ARN [259, 65]. Le nombre de gènes produisant des 3'UTRs alternatifs a considérablement augmenté au cours de l'évolution, ainsi que leur longueur qui est de 140 nt en médiane pour le nématode et de 1,200 pour l'homme [112]. Malgré cette grande variété, les 3'UTRs contiennent des éléments très conservés et

ils servent de site de fixation pour de nombreuses protéines régulatrices ainsi que pour les microARNs [243].

### Pseudogènes

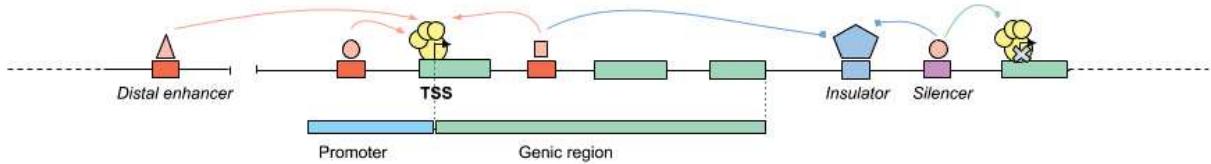
Les pseudogènes sont définis comme des séquences ayant une structure similaire aux gènes codants mais qui ne peuvent pas produire de protéine fonctionnelle. Ils ont perdu leur fonction suite à l'accumulation de mutations délétères, comme l'insertion ou la suppression de nucléotides, entraînant une perturbation du cadre de lecture qui empêche le passage de l'ARN à une protéine fonctionnelle [57]. Les pseudogènes peuvent toutefois être transcrits et peuvent même être impliqués dans la régulation de leurs gènes codants homologues [117]. Leur quantité est comparable à celle des gènes codants pour des protéines [263] et 14723 pseudogènes sont annotés dans GENCODE v28. Leur étude peut donner des indications sur l'histoire évolutive du génome.

### 1.2.3 Séquences régulatrices

La régulation transcriptionnelle de l'expression des gènes, c'est à dire lors du passage d'un locus de l'ADN à une copie ARN, est une étape critique dans l'adaptation de la cellule à son environnement, dans le temps et l'espace. Pour contrôler la transcription d'un gène particulier, une combinaison complexe d'interactions entre régions régulatrices et protéines nucléaires se met en place. Ainsi, les promoteurs situés à proximité du gène établissent des interactions avec d'autres régions régulatrices plus éloignées : les *enhancers*, les *silencers* et les *insulators*. On parle d'*enhancer* quand ces séquences permettent d'activer ou d'amplifier l'expression de leurs gènes cibles, de *silencer* quand elles répriment l'expression de leurs gènes cibles et d'*insulator* quand elles empêchent certaines interférences et interactions à longue distance. Toutes trois sont des séquences régulatrices pouvant être à plusieurs kilobases de leurs gènes cibles et agissant en *cis*, c'est à dire régulant préférentiellement les gènes voisins. Leur action est possible via le repliement de l'ADN et la formation de boucles qui permettent de rapprocher ces régions de leurs gènes cibles (voir partie 1.1.2 et Figure 1.18) [168].

#### Promoteur

Pour qu'un gène puisse être transcrit, la machinerie de transcription est recrutée au niveau d'une séquence régulatrice généralement située en amont du TSS et que l'on appelle région promotrice. Les promoteurs sont localisés sur le même brin



**FIGURE 1.18 – Éléments régulateurs distaux des gènes.** Les éléments régulateurs des gènes peuvent être localisés à leur proximité, dans la région génique ou être très éloignés. Ils recrutent des protéines spécifiques (triangle, rond et carré orange clair) pour exercer leur action régulatrice positive ou négative sur la machinerie transcriptionnelle (ronds jaunes). Les enhancers (rectangle rouge) et les silenciers (rectangle violet) respectivement activent et répriment l'expression de leurs gènes cibles. Les insulateurs recrutent des protéines (rectangle bleu + pentagone bleu) qui permettent de bloquer ou limiter l'action des éléments enhancer/silencer sur certains gènes cibles.

d'ADN que le gène à transcrire et ils sont caractérisés par des séquences nucléotidiques particulières, les motifs, sur lesquels les facteurs de transcription impliqués dans l'initiation et la régulation de la transcription se lient [39, 34] et interagissent. Le motif le plus connu présent dans 10 à 20% des gènes codants pour des protéines est la TATA-box, permettant le recrutement de la "TATA-binding protein" (TBP). Un autre motif, l'élément initiateur (Inr), est présent dans 40 à 60% des promoteurs [283]. Les promoteurs sont caractérisés par un ensemble d'éléments régulateurs variant entre les espèces mais aussi selon le type de gène [261].

Un gène peut contenir plusieurs sites de départ de transcription et donc plusieurs promoteurs. Le recrutement de la machinerie de transcription peut ainsi avoir lieu au niveau de différents promoteurs appelés promoteurs alternatifs engendrant la production d'une grande variété d'ARNs. D'après une étude de Cramer et al. en 1997 [44], la composition du promoteur qui conditionne son architecture joue un rôle sur la modulation de l'épissage alternatif. D'autres facteurs relatifs au promoteur peuvent influencer la transcription comme la méthylation de l'ADN qui varie considérablement selon les gènes et selon le type cellulaire. Une forte méthylation du promoteur corrèle avec une transcription faible ou inexiste [254].

## Enhancer

Le premier *enhancer* a été décrit lors d'une expérience de clonage avec une séquence virale (simian virus SV40) d'une longueur de 72 pb, que l'on peut retrouver chez l'homme. Sa présence dans l'environnement du gène alors étudié ( $\beta$ -globine) entraîne une augmentation considérable de son expression et ceci indépendamment de l'orientation de la séquence virale ou de sa distance au promoteur du gène cible

[7]. Aujourd’hui, de nombreux *enhancers* ont été découverts *in vivo* et dans différents types cellulaires. Les caractéristiques communes aux *enhancers* découverts expérimentalement ont permis de développer des outils pour une annotation automatique. Ainsi, les *enhancers* sont décrits comme de courts segments d’ADN ayant la capacité d’amplifier l’expression d’un ou de plusieurs gènes cibles. Ils peuvent être localisés à proximité du gène cible, au sein du gène cible ou à une distance importante (voir Figure 1.18). Quand l’*enhancer* est situé sur un chromosome différent de celui du gène cible, on parle d’action régulatrice en *trans* [163]. Leur action n’est pas dépendante de l’orientation vis à vis du gène cible. De plus, ils sont hypersensibles au traitement à la DNase I qui est une enzyme coupant l’ADN libre, et sont caractérisés par des motifs spécifiques permettant la liaison de facteurs de transcription et de co-activateurs comme p300. Enfin, ils présentent un enrichissement de certaines marques épigénétiques dont une acétylation d’histone (H3K27ac) et une monométhylation (H3K4me1) [157, 99].

Les méthodes de prédiction pour l’annotation des *enhancers* intègrent de nombreux types de données pour les repérer dans le génome : données de DNase-seq permettant d’identifier les régions décondensées et enrichies en éléments régulateurs ; détection de marques épigénétiques particulières (H3K4me1, H3K27ac) avec un ratio H3K4me1/H3K4me3 élevé ; liaison de co-activateurs et d’acetyl-transférase ainsi que de nombreux facteurs de transcription. Malgré ces indicateurs, les *enhancers* ne sont pas facile à détecter. En 2010, deux études mettent en évidence la transcription bidirectionnelle par l’ARN polymérase II des *enhancers* en petits ARNs non-codants, appelés les eRNAs (*enhancer RNA*, eRNA) [134, 50]. Ces eRNAs sont décrits comme des marques de l’activité des *enhancers* et leur transcription corrèle avec l’activité des *enhancers* actifs. Aujourd’hui, un des atlas d’*enhancers* les plus importants est celui établi par FANTOM5. En effet, les données de CAGEs utilisées pour détecter les *enhancers* actifs et transcrits couvrent plus de 800 échantillons différents provenant de cellules primaires, de tissus et de lignées cellulaires [3]. Dans un contexte de compréhension des mécanismes de régulation par les *enhancers*, les *super-enhancers* ont été introduits et décrits comme une classe de régions régulatrices enrichies en co-activateurs et plus particulièrement en médiateurs qui sont des co-activateurs transcriptionnels multi-protéiques [209]. Les *super-enhancers* peuvent être vus comme des clusters d’*enhancers*.

Pour exercer leur activité régulatrice, les *enhancers* interagissent directement ou indirectement avec leurs gènes cibles et différents modèles ont été envisagés [139] :

dans le modèle du "tracking", les protéines régulatrices sont chargées au niveau de l'*enhancer* et se déplacent le long de la chromatine pour atteindre le promoteur ; dans le second modèle dit de "linking", les protéines régulatrices viennent se polymériser en direction du promoteur ; enfin, pour le dernier modèle dit du "looping" qui est le plus courant, promoteur et *enhancer* interagissent directement par formation d'une boucle.

Malgré un support croissant de l'hypothèse d'un rôle biologique pour les eRNAs, des études suggèrent qu'ils ne sont pas nécessaires à la définition d'un *enhancer* fonctionnel [92], et leur fonction reste à éclaircir. Cependant, il est clair que les *enhancers* jouent un rôle essentiel dans le contrôle spatio-temporel de l'expression des gènes et ils sont nécessaires pour des processus comme la différentiation ou le développement [17]. La différence d'expression des *enhancers* à travers les tissus permet de mieux comprendre les mécanismes de contrôle de la diversité des types cellulaires partageant le même génome.

**ePromoteur.** Les promoteurs et *enhancers* partagent de nombreuses caractéristiques chez les mammifères. Ces deux types de séquences régulatrices présentent un signal de transcription bidirectionnel et elles sont fixées par des facteurs de transcription et cofacteurs de manière similaire et dépendante de leur composition nucléotidique [41]. Ces ressemblances suggèrent des mécanismes d'action communs et montrent que la dichotomie établie entre les deux classes n'est pas exclusive. De plus, il a été montré que 2 à 3% des promoteurs de gènes codants dans un type cellulaire donné possédaient une activité d'*enhancer* [49]. Ces ePromoteurs ont des caractéristiques génomiques et épigénomiques qui leurs sont propres et semblent contenir des éléments régulateurs ayant un double rôle pour assurer à la fois la fonction de promoteur et d'*enhancer*.

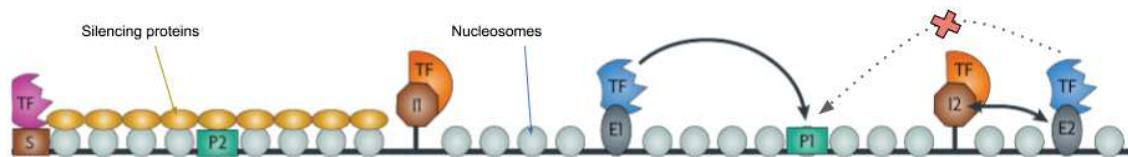
### Silencer

On parle souvent des facteurs de transcription se fixant au niveau de séquences régulatrices pour activer ou amplifier la transcription du gène cible, mais il existe aussi des TFs inhibiteurs qui peuvent inactiver l'expression du gène cible via des séquences régulatrices appelées *silencers* (voir Figure 1.18)) [190]. Leur découverte remonte aux années 80 chez la levure, puis peu de temps après chez le rat [19, 148]. Les *silencers* ont des propriétés communes avec les *enhancers* comme leur fonctionnement indépendant de l'orientation et de la distance au promoteur cible. Le recrutement des TFs répresseurs au niveau des *silencers* peut aussi être accompagné

du recrutement de co-facteurs négatifs (co-répresseurs). La fonction répressive de ces éléments peut être établie selon plusieurs modèles : par blocage de la fixation des activateurs proches ou directement par compétition pour le même site de fixation [168]. Les *silencers* et répresseurs associés peuvent également empêcher les activateurs et facteurs de transcription généraux (*General transcription factors*, GTFs) d'accéder aux promoteurs, via la création d'une structure répressive de la chromatine. Enfin, il semblerait qu'ils puissent dans certains cas bloquer l'assemblage du complexe d'initiation de la transcription (*Preinitiation complex*, PIC) [28].

### Insulator

Les *insulators* sont des éléments régulateurs pouvant empêcher l'action des *enhancers* ou *silencers* sur leurs promoteurs cibles (voir Figure 1.19).



**FIGURE 1.19 – Fonctionnement des *insulators*.** Les nucléosomes sont représentés par des cercles bleus clair et les protéines induisant la condensation de la chromatine sont représentées par des ovales jaunes. Le premier *insulator* "I1" permet de bloquer l'effet répresseur du *silencer* "S" en bloquant la propagation d'une structure condensée de la chromatine. L'*enhancer* "E1" entouré par deux *insulators* "I1" et "I2" est capable d'amplifier la transcription de son gène cible en communiquant avec le promoteur "P1". L'*enhancer* "E2" ne peut pas communiquer avec "P1" car il est bloqué par l'*insulator* "I2". [Figure adaptée de [214]]

Plusieurs études ont montré l'importance des *insulators* dans la régulation de l'expression des gènes et des mutations de ces derniers peuvent entraîner des défauts de développement [69]. Les *insulators* sont caractérisés comme bloqueur d'*enhancers* quand ils sont situés entre un promoteur et un *enhancer* et de barrière quand ils sont entre un promoteur et un *silencer* [266] (voir Figure 1.19). Ils peuvent posséder une seule de ces deux caractéristiques ou les deux. Ils ont une taille généralement comprise entre 0.5 et 3 kb et fonctionnent de manière indépendante à l'orientation par rapport à leurs cibles.

Les *insulators* sont impliqués dans le contrôle de plusieurs gènes soumis à empreinte qui sont des gènes dont un seul des deux allèles hérités du père et de la mère est exprimé. Ainsi, chez l'homme, la protéine CTCF, qui est un répresseur transcriptionnel et une des protéines les plus associées aux *insulators*, joue un rôle pour

les gènes H19/Igf2 soumis à empreinte où elle bloque l'activation de l'allèle maternel Igf2 par un *enhancer* [78].

Toutes ces régions régulatrices sont caractérisées par des marques épigénétiques et un environnement particulier. Ainsi, des modèles ont été développés pour définir de manière automatisée les frontières des différentes portions du génome.

### 1.2.4 Modèles de segmentation

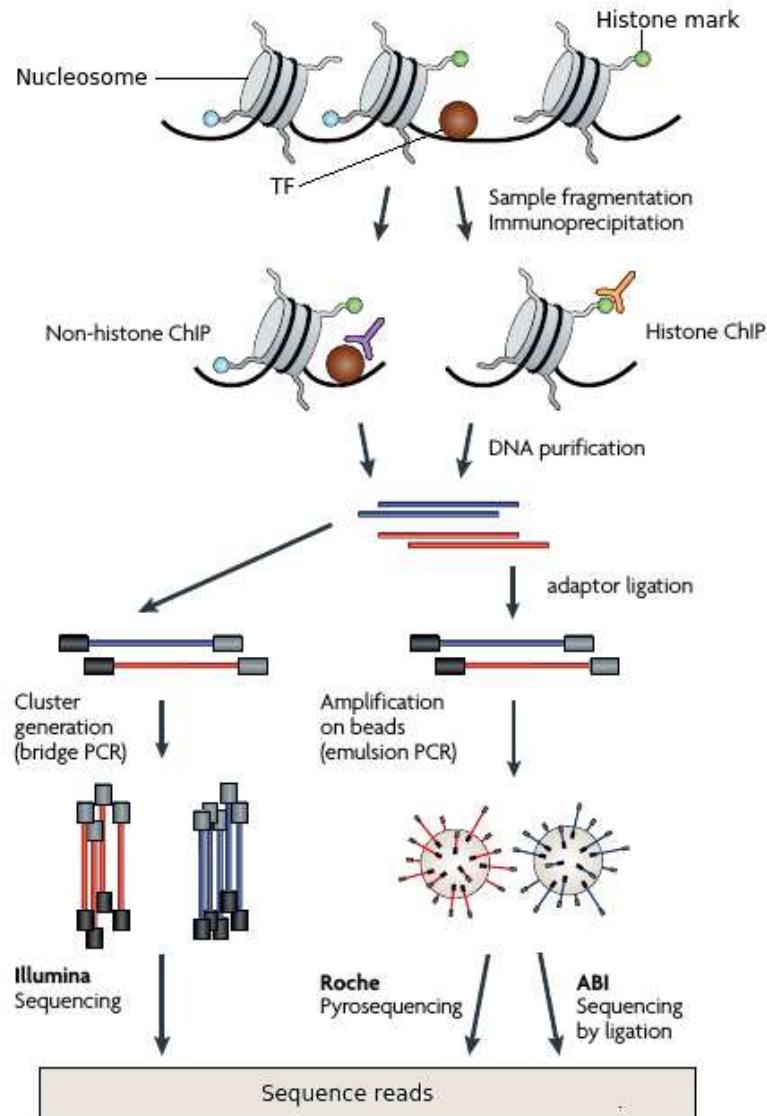
#### ChIP-seq

L'immunoprecipitation de la chromatine combinée à du séquençage haut débit (*Chromatin immunoprecipitation followed by sequencing*, ChIP-seq) permet d'établir les profils de distribution, à l'échelle du génome, de nombreuses marques épigénétiques, nucléosomes et facteurs de transcriptions et ce, dans différents types cellulaires. Dans toute expérience de ChIP-seq, les fragments d'ADN associés à une protéine d'intérêt sont enrichis. Les protéines sont ensuite réticulées à l'ADN par un traitement des cellules au formaldéhyde, puis l'ADN est fragmenté par sonication. Enfin, grâce à un anticorps spécifique de la protéine d'intérêt et/ou de la marque épigénétique, les complexes protéine/ADN sont immunoprecipités et les fragments d'ADN sont relâchés, purifiés, marqués par des adaptateurs, amplifiés et séquencés (voir Figure 1.20).

Jusqu'à la purification de l'ADN et le marquage par des adaptateurs, le protocole est similaire pour les différentes plateformes de séquençage existantes. L'amplification des fragments marqués peut ensuite être effectuée de différentes façons selon la technologie de séquençage choisie. En 2007, les profils de 20 marques d'histones, de H2A.Z, de la RNA polymérase II et de la protéine CTCF sont établis par ChIP-seq [9]. Aujourd'hui, grâce à la quantité de données générées [35], il est possible d'étudier la combinaison de ces différentes marques et leur corrélation avec l'expression des gènes.

#### Projets internationaux recensant des données épigénomiques

L'étude de l'épigénome fourni des informations sur le profil des marques épigénétiques et des protéines gouvernant l'expression des gènes dans une cellule. Une grande diversité de techniques existe pour les étudier : ChIP, digestion de l'ADN par la DNase I (DNase), traitement au bisulfite pour étudier la méthylation de l'ADN et bien d'autres. Le profilage des ARNs pour quantifier les transcrits permet de



**FIGURE 1.20 – Principe d'une expérience de ChIP-seq.** Le concept de base de cette technique, jusqu'à la purification de l'ADN, est commun à toutes les plateformes de séquençage. Pour une description plus détaillée, voir texte. [Figure adaptée de [196]]

comprendre l'impact de la combinaison des marques épigénétiques sur l'expression des gènes. Les données relatives à l'épigénome sont générées pour de nombreuses marques épigénétiques et de nombreux types cellulaires. Elles sont regroupées et rendues accessibles par des consortium internationaux dont les plus importants sont ENCODE [35] et Roadmap Epigenomics [147]. ENCODE possède plus de 9000 jeux de données couvrant un grand nombre de lignées cellulaires, cellules primaires et tissus. Le projet Roadmap Epigenomics a également généré une collection importante de données pour des lignées cellulaires primaires et tissus sains ou malades, chez l'enfant et l'adulte, recouvrant 111 épigénomes différents, avec un supplément de 16 épigénomes du projet ENCODE.

Les données de ChIP-seq disponibles dans les bases de données publiques s’accumulent et intégrer les informations provenant de plusieurs projets est intéressant pour l’identification et l’étude des régulateurs transcriptionnels. Cependant, les protocoles et logiciels utilisés pour analyser ces données sont souvent hétérogènes. ReMap intègre les données de ChIP-seq de facteurs de transcription du projet ENCODE et d’autres ressources publiques (Gene Expression Omnibus, GEO [8], ArrayExpress [138]) pour proposer un atlas standardisé de régions régulatrices appelées aussi modules cis-régulateurs (*Cis-regulatory modules*, CRMs) [29].

### Modèles ChromHMM et Segway

Pour mieux comprendre l’organisation particulière du génome et distinguer les régions régulatrices des régions transcrives, réprimées, etc, des modèles de segmentation du génome ont été développés. Ces différentes régions, que l’on appelle états chromatiniens, sont définies sur la base de combinaisons particulières de modifications d’histones et correspondent à des régions fonctionnelles différentes. Le but de ces méthodes *de novo* est d’apprendre, à partir de données expérimentales, des combinaisons particulières de marques épigénétiques significatives et caractéristiques des différents états.

ChromHMM (*Chromatin multivariate Hidden Markov Model*) [66] est un modèle d’apprentissage des états chromatiniens développé par un groupe du consortium ENCODE qui permet la caractérisation des fonctions biologiques des différents états. ChromHMM utilise un modèle de Markov caché multivarié permettant de capturer des combinaisons complexes de modifications de la chromatine. Ainsi, pour chaque segment du génome d’une taille définie par défaut à 200 pb et dans un type cellulaire particulier, ChromHMM renvoie l’état chromatinien le plus probable, basé sur la présence/absence de nombreuses marques épigénétiques. Il permet de capturer des classes connues d’éléments génomiques comme promoteurs, enhancers ou régions répétées mais capture aussi de nouvelles classes d’éléments et aide à l’annotation du génome non-codant. Indépendamment et la même année, un autre modèle de segmentation du génome appelé Segway est développé [101]. Segway utilise un réseau Bayésien dynamique (*Dynamic Bayesian network*, DBN) qui peut être vu comme une généralisation du HMM, pour segmenter et regrouper les informations extraites des données génomiques de ChIP-seq et d’ouverture de la chromatine du projet ENCODE, avec une résolution d’1 pb. Le nombre d’états chromatiniens est fixé à 25 pour pouvoir interpréter facilement les résultats.

Les modèles ChromHMM et Segway ont été implémentés séparément et diffèrent sur plusieurs points : méthode utilisée, résolution génomique, ensemble d'apprentissage... Cependant, elles partagent de nombreuses variables et ont un but commun : segmenter le génome en plusieurs états chromatiniens selon le profil des marques épigénétiques. Ainsi, une segmentation du génome combinant les résultats de ces deux modèles a été établie [75]. Quand les états définis par ChromHMM et Segway séparément sont concordants, ils sont associés à une des 7 classes définissant le modèle combiné. Dans le cas contraire, ils ne sont pas assignés. Les modèles de segmentation aident à la caractérisation et l'annotation du génome non codant et permettent également de trouver des caractéristiques communes aux CRMs comme les enhancers décrits précédemment.

### 1.2.5 Éléments répétés

Comme nous l'avons déjà mentionné, les exons des gènes codants représentent 2% seulement du génome. Les 98% restants correspondent à l'ADN non codant et environ 50% du génome est représenté par les éléments répétés qui sont des séquences que l'on retrouve en plusieurs copies (voir Figure 1.21).

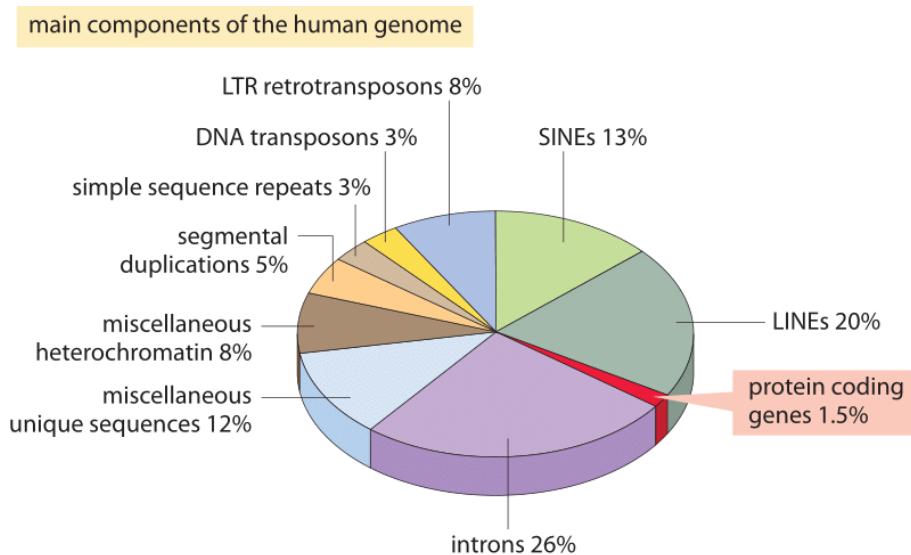


FIGURE 1.21 – Proportion des différents composants majeurs du génome humain.  
[Diagramme extrait de l'ouvrage *Biology by the numbers*]

Parmi les éléments répétés, la famille des éléments transposables est la plus représentée. Ces éléments que l'on appelle aussi "gènes sauteurs" sont des séquences de l'ADN capables de se répliquer et de se déplacer au sein du génome. Ils ont été

mis en évidence pour la première fois il y a presque 70 ans par Barbara McClintock [173]. En étudiant le maïs et la diversité qu'il peut y avoir dans la coloration de ses grains, elle montre pour la première fois que des éléments du génome peuvent se dupliquer et se déplacer le long des chromosomes, et qu'ils peuvent influencer l'expression des gènes et par conséquent, modifier le phénotype qui en découle.

### Classification des éléments transposables

Les éléments transposables peuvent se déplacer dans le génome par deux mécanismes distincts : soit par transposition (mécanisme de "couper-coller"), soit par rétrotransposition (mécanisme de "copier-coller") [128]. Les transposons se déplacent via une matrice ADN et ne sont pas dupliqués alors que les rétrotransposons sont d'abord copiés en ARN et peuvent se dupliquer en de multiples copies [64]. Les différentes classes d'éléments transposables sont présentées sur la figure 1.22.

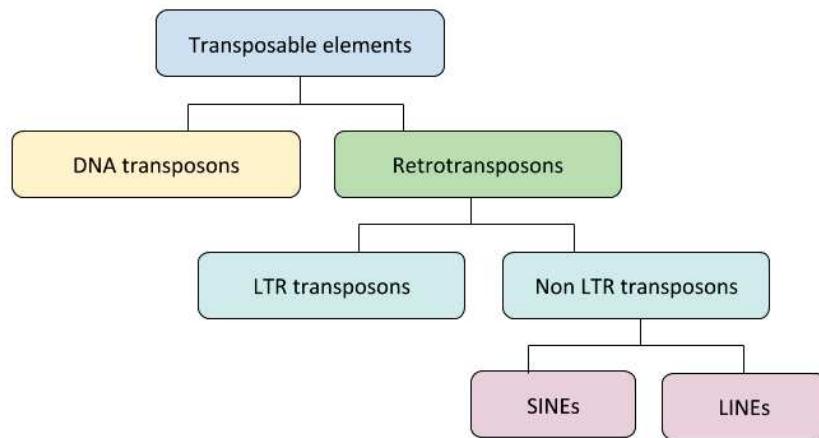
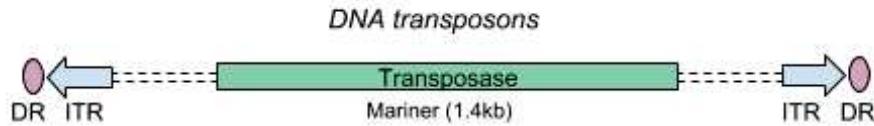


FIGURE 1.22 – Classification des éléments transposables.

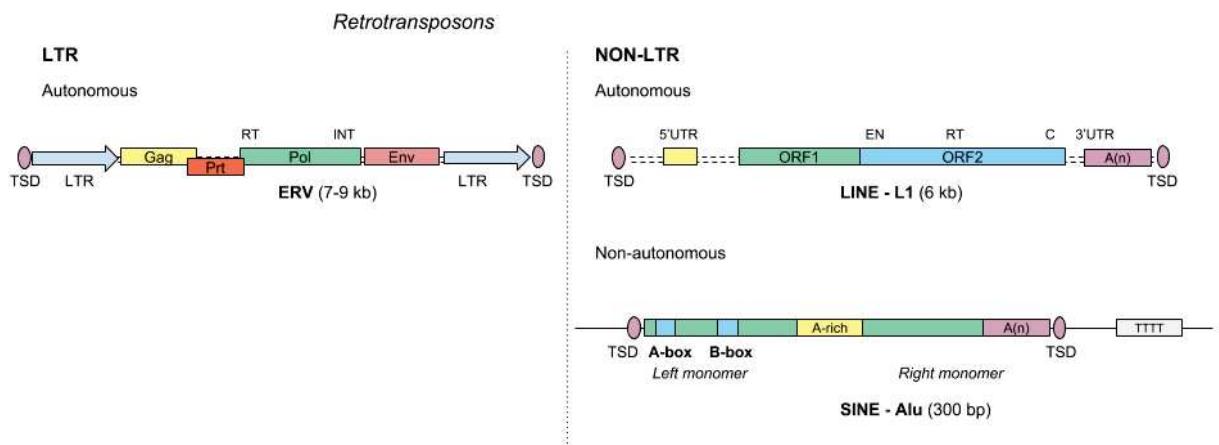
Les différentes familles d'éléments transposables et leurs séquences consensus sont répertoriées dans une base de donnée commune appelée Repbase [121] (<http://www.girinst.org/repbase/>).

**Les transposons à ADN.** Les transposons à ADN utilisent un intermédiaire ADN simple ou double brin pour se déplacer [70]. Les transposons actifs possèdent une séquence codant pour une transposase entourée de deux bornes : les répétitions terminales inversées (*Inverted terminal repeats*, ITR) (voir Figure 1.23). La transposase va reconnaître le transposon au niveau de ses ITRs et l'exciser avant de l'insérer ailleurs dans le génome. Ces éléments sont généralement d'une longueur comprise entre 1.3 et 2.4kb [183]. Ils peuvent altérer l'expression des gènes codants en s'insérant dans les introns, exons ou dans les régions régulatrices.



**FIGURE 1.23 – Structure type d'un transposon à ADN.** La classe de transposons *Mariner* est retrouvée chez de nombreux organismes, dont l'homme. La partie centrale du transposon code pour une transposase. Aux extrémités on trouve les répétitions terminales inversées (ITR) et de petites répétitions directes (DRs). [Figure adaptée de [85]]

**Les rétrotransposons.** Les rétrotransposons représentent la majeure partie des éléments répétés (voir Figure 1.21). Ils utilisent une molécule ARN comme intermédiaire de transposition. Ils sont ainsi copiés en ARN puis rétro-transcrits à l'aide d'une transcriptase inverse (*Reverse Transcriptase*, RT) en une molécule d'ADN complémentaire qui est ensuite intégrée dans le génome. Contrairement à la majorité des transposons à ADN, l'élément est ici dupliqué. Les rétrotransposons sont séparés en deux grandes classes selon qu'ils possèdent de longues répétitions terminales (*Long Terminal Repeat*, LTR) à leurs extrémités 5' et 3' (rétrotransposons à LTRs) ou qu'ils n'en possèdent pas (rétrotransposons sans LTR) [85].



**FIGURE 1.24 – Les différentes classes de rétrotransposons.** Sur la gauche les rétrotransposons à LTRs avec les sites de duplication (*target site duplication*, TSD), les longues répétitions terminales (*long terminal repeat*, LTR) et les différents gènes (Gag, Pol, Env). Sur la droite les rétrotransposons sans LTR. [Figure adaptée de [85]]

Les rétrotransposons à LTRs (voir Figure 1.24) ont une structure proche de celle des rétovirus. Ils possèdent un promoteur interne localisé dans le LTR en 5' et sont transcrits en ARN par l'ARN Pol II. Entre les deux LTRs situés aux extrémités 5' et 3', des gènes codants pour des protéines de structures et des protéines ayant une

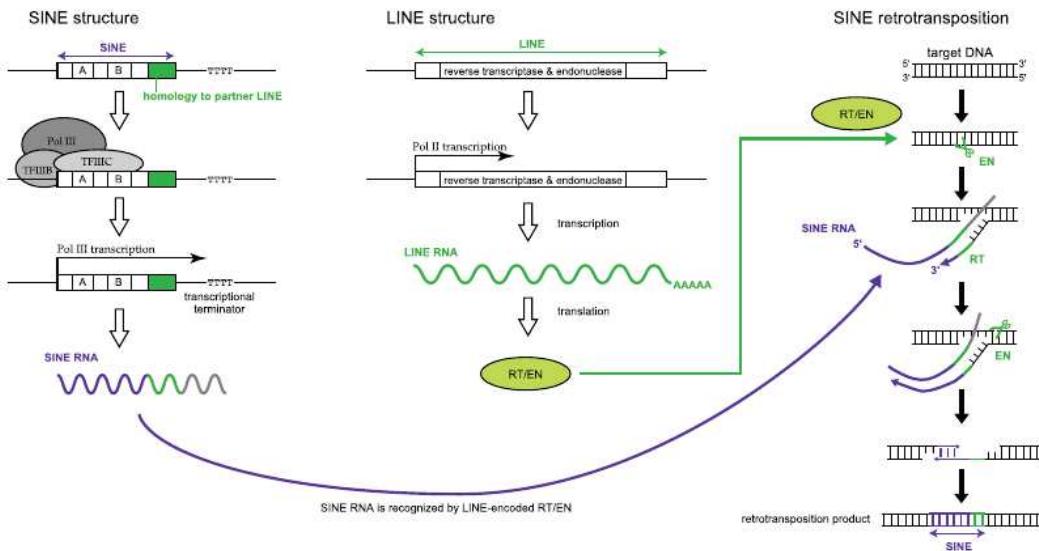
activité enzymatique sont présents. Le gène Gag code pour une pseudo-particule virale (*Virus-like particles*, VLP) dans laquelle la transcription inverse a lieu. Le gène Pol code pour plusieurs fonctions enzymatiques : une protéase permettant de couper les polyprotéines en plusieurs segments, une reverse transcriptase qui permet de transcrire l'ARN en ADN complémentaire et enfin une intégrase qui permet l'intégration du segment d'ADN produit dans le génome hôte [97]. Le gène Env code quant à lui pour des protéines d'enveloppe.

Les rétrotransposons sans LTR se séparent en deux groupes majeurs : les longs éléments nucléaires intercalés (*Long interspersed nuclear elements*, LINEs) et les petits éléments nucléaires intercalés (*Short interspersed nuclear elements*, SINEs) (voir Figure 1.24). Les LINEs, d'une longueur moyenne de 6kb, possèdent tout le matériel nécessaire pour assurer leur rétrotransposition : ils sont autonomes. La famille des LINE-1 est la seule classe active de rétrotransposons autonomes [10, 192]. Ces éléments possèdent un promoteur interne et sont transcrits en ARN par l'ARN polymérase II. Les transcrits produits sont ainsi coiffés en 5' et poly-adénylés en 3'. Ils contiennent également deux cadres de lecture : l'ORF1 est similaire à Gag et code pour une protéine de liaison à l'ARN (*RNA binding protein*, RBP) et l'ORF2, similaire à Pol, code pour une RT et une endonucléase [143]. Ces protéines vont permettre la copie de l'ARN en ADN complémentaire et son insertion dans le génome hôte. Les protéines résultant des LINEs servent également à la rétrotransposition de la deuxième grande classe de rétrotransposons que forment les SINEs (voir Figure 1.25). En effet, ces derniers ne sont pas autonomes et utilisent la machinerie des LINEs pour leur rétrotransposition [59].

Malgré la faible proportion de LINEs actifs dans le génome humain, ils ont un rôle dans le processus d'inactivation du chromosome X. En effet, il sont impliqués dans la formation d'un compartiment nucléaire réprimé dans lequel les gènes sont recrutés pour être inactivés [31]. De plus, une réactivation de l'expression des LINE-1 a été observée dans de nombreux cancers [222, 12], notamment le cancer de la prostate où l'ORF1 est exprimée dans 40-50% des cas [20].

## SINEs

Les SINEs représentent environ 13% du génome humain (voir Figure 1.21). Ils possèdent un promoteur interne et sont transcrits par l'ARN polymérase III, comme les gènes codants pour les ARNs de transfert. Comme nous l'avons vu dans la sous-partie précédente, ils ne sont pas autonomes et ne codent pour aucune protéine.



**FIGURE 1.25 – Rétrotransposition des SINES et LINEs.** Comme vu précédemment, les SINES contiennent généralement deux séquences appelées A et B-box, qui sont reconnues par des facteurs de transcription et la polymérase III, et une séquence homologue aux LINEs (rectangle vert). La transcription du SINE débute quand la polymérase est recrutée au niveau du promoteur et termine au premier élément de terminaison TTTT rencontré. Les LINEs codent pour une protéine à activité reverse transcriptase (RT) et endonucléase (EN). Enfin, la rétrotransposition des SINES et LINEs est assurée par les protéines codées par les LINEs qui viennent couper l'ADN et initier la transcription inverse (synthèse de l'ADN complémentaire). La reconnaissance du transcrit ARN du SINE est effectuée via la région 3' homologue à celle des LINEs. [Figure extraite de [105]]

Ils ne possèdent pas non plus de séquence spécifique pour la terminaison de la transcription, la polymérase s'arrête donc au niveau de la répétition de 4 T ou plus la plus proche, en aval du SINE (voir Figure 1.25) [52]. Les protéines codées par les LINEs permettent le clivage de l'ADN et la rétrotransposition des SINES transcrits en ADN complémentaire (ADNc), par reconnaissance au niveau de leur extrémité 3' (queue poly-A), homologue à celle des LINEs [59].

**Les Alus.** Les SINES les plus représentés chez l'homme sont les Alus avec plus d'1 million de copies et une longueur d'environ 300 pb. Leur structure générale est présentée sur la Figure 1.24. Ils sont composés de deux monomères dérivant du gène codant pour l'ARN 7SL et séparés par une région riche en A (séquence consensus : A<sub>5</sub>TAC A<sub>6</sub>) [40, 54].

La présence de la queue poly-A au niveau de l'extrémité 3' des Alus est nécessaire à une rétrotransposition efficace. Comme nous l'avons vu ci-dessus, les Alus n'étant pas autonomes, la queue poly-A sert de site d'initiation à la reverse transcriptase. Il a été montré que la longueur de la queue poly-A avait un impact sur la rétro-

transposition des Alus et qu'entre une longueur de 10 et de 50 nt la fréquence de transposition est accrue d'un facteur 10 [60, 227]. Les Alus sont divisés en plusieurs familles selon leur similarité de séquence. Les copies de ces rétrotransposons ont accumulé des mutations au cours de l'évolution, leur séquence reflète donc leur âge. On distingue ainsi 3 sous-familles d'Alus : les AluJ, AluS, et AluY, qui peuvent être encore divisées en sous-sous-familles [53]. La longueur de la queue poly-A varie au sein de ces différentes classes : les vieux Alus, soit les AluS et AluJ, ont une longueur moyenne de 21 pb et les jeunes Alus, les AluY, ont une longueur moyenne de 26 pb. Les Alus actifs les plus récents ont tous une longueur supérieure à 40 pb [227].

**Activité des SINEs.** La majorité des SINEs ont accumulé des mutations au cours de l'évolution et sont aujourd'hui inactifs. Les transcrits dérivant des SINEs sont peu abondants, il est donc difficile de les détecter dans les cellules somatiques [106] malgré leur grand nombre dans le génome humain. Il a toutefois été montré que les SINEs étaient plus transcrits dans les cellules germinales primordiales, qui sont les précurseurs des cellules spermatogéniques et ovocytes, que dans les cellules somatiques [105]. L'expression tissu-spécifique est régulée par des mécanismes épigénétiques, et notamment par la méthylation de l'ADN au niveau des dinucléotides CpG.

Les régions riches en SINEs sont aussi riches en gènes alors que l'on retrouve plutôt les LINEs dans les régions intergéniques [37]. Leur localisation au niveau de la chromatine n'est pas non plus aléatoire : les SINEs sont enrichis au centre de la chromatine et les LINEs s'accumulent plutôt à la périphérie [15, 89, 106] suggérant leur implication dans l'évolution de la structure de la chromatine. La méthylation importante de ces éléments peut aussi influencer l'expression des gènes voisins. Leur activité peut également être à l'origine d'altérations génomiques et de maladies [133]. Bien que les fonctions des SINEs ne soient pas bien comprises, il a été montré à travers une étude réalisée à l'échelle du génome qu'ils partageaient de nombreuses caractéristiques avec les enhancers [252]. Ils présentent un profil de marques épigénétiques proche de ces derniers et sont observés en interaction avec les promoteurs voisins.

## Conclusion

L'avancée des technologies de séquençage a permis de mettre en évidence la transcription d'une grande partie du génome non codant, considéré autrefois comme ADN poubelle. Ces dernières années, le nombre de publications sur les petits et longs ARNs non-codants a augmenté de façon exponentielle. Toutefois, les fonctions de ces ARNs ne sont pas bien caractérisées et malgré des évidences sur leur implication dans les mécanismes de régulation de l'expression des gènes et leur association à certaines maladies, leur mode de fonctionnement reste à déterminer.

## 1.3 Régulation de l'expression génique

Comme nous l'avons vu précédemment, selon le type cellulaire, les gènes ne sont pas exprimés de la même façon et les cellules savent s'adapter en réponse à des stimuli qu'elles reçoivent et en fonction de leur environnement. L'expression des gènes, que l'on peut mesurer par le taux d'ARNs transcrits, est régulée au niveau transcriptionnel, en partie via les promoteurs des gènes et leurs interactions avec des éléments *cis*-régulateurs, et au niveau post-transcriptionnel lors de la maturation de l'ARN produit et de la traduction en protéine fonctionnelle.

### 1.3.1 Facteurs de transcription et séquences régulatrices

Les facteurs de transcription sont des protéines nécessaires à l'initiation et à la régulation de la transcription des gènes. On distingue les facteurs généraux, indispensables au recrutement de la polymérase, et les facteurs spécifiques qui permettent chacun de moduler l'expression d'une quantité réduite de gènes en réponse à un signal biologique.

#### Facteurs de transcription

Pour que les facteurs de transcription reconnaissant des motifs de l'ADN puissent accéder à leurs sites de fixation, l'ADN doit être décondensé. Il existe une classe particulière de facteurs de transcription appelés les facteurs pionniers qui sont capables de se lier directement à la chromatine condensée et rendent ainsi l'ADN accessible au recrutement des autres facteurs de transcription et des enzymes de modification des histones [286].

Les facteurs de transcription, autre que les facteurs généraux, permettent de réguler spécifiquement l'expression des gènes dans les cellules. Leur action s'effectue par reconnaissance de sites particuliers de l'ADN sur lesquels ils vont venir se fixer. Ces sites de fixation (*Transcription factor binding sites*, TFBSS) sont de longueur comprise entre 6 et 30 nucléotides. Ils peuvent être aussi bien situés à proximité du gène cible, généralement au niveau du promoteur où ils sont enrichis, qu'à des centaines de paires de bases du promoteur (au niveau des enhancers et silencers). On distingue deux types de TFs : les facteurs activateurs et les facteurs répresseurs de la transcription. Les facteurs activateurs possèdent au moins deux domaines : un domaine de fixation à l'ADN qui va reconnaître la séquence spécifique de liaison et un domaine d'activation qui peut agir en attirant les GTFs, l'ARN Pol II ou par action indirecte en modifiant la structure de la chromatine. Les facteurs répresseurs,

comme leur nom l'indique, vont empêcher ou limiter la transcription en masquant les surfaces d'activation, par compétition de fixation avec un activateur, par interaction directe avec les GTFs ou encore par modification de la structure de la chromatine. Cependant, ces fonctions ne sont pas exclusives et un même co-facteur peut être activateur ou répresseur selon le contexte. Les TFBSSs ne sont pas stricts, c'est à dire qu'un même TF peut reconnaître plusieurs sites de l'ADN pouvant différer de quelques nucléotides. On dit qu'ils sont dégénérés [13]. Cette observation entraîne le développement de modèles synthétiques pour la description des sites de fixation. Plusieurs types de représentation de ces motifs et de leur spécificité existent, afin de conserver au mieux l'information sur les propriétés de fixation de chacun des TFs connus. Ils vont aussi permettre de prédire la présence de sites de fixation potentiels dans les séquences inconnues.

### Modèles de représentation des motifs

Le premier modèle de représentation le plus courant est la séquence consensus (chaîne de caractères) [250]. Cette séquence représente tous les sites de fixation possibles par une lettre de l'alphabet dégénéré déterminé par l'IUPAC (*International Union of Pure and Applied Chemistry*). Ce code prend en compte toutes les combinaisons possibles d'un ou plusieurs des 4 nucléotides et suivant une suite de règles précises déterminées par Cavener [26]. Par exemple, R correspond à A ou G, Y à C ou T et N à n'importe quelle base. Cette représentation est simple mais elle a comme inconvénient de ne donner aucune information sur la probabilité d'avoir chacune des bases possibles à chaque position, sauf quand il y a exclusivité d'un nucléotide.

Une autre façon plus conservatrice de représenter un motif est de construire une matrice, très souvent sous la forme d'une matrice de poids appelée *Position weight matrix* (PWM) (voir Figure 1.26). La construction de cette matrice passe par le comptage des occurrences de chaque nucléotide à chaque position du motif (*Position frequency matrix*, PFM, voir Figure 1.26). En divisant ces occurrences par le nombre de séquences, on obtient la matrice de probabilité (*Position probability matrix*, PPM) qui donne pour chaque position du motif, les différentes probabilités d'avoir chacun des 4 nucléotides [274]. Ces probabilités sont souvent corrigées pour les petits jeux de données, pour qu'il n'y ait pas de valeur nulle, et sont calculées de la façon suivante :

$$p(b, i) = \frac{fb, i + s(b)}{N + \sum_{b' \in [A, C, G, T]} s(b')} \quad (1.1)$$

où  $p(b, i)$  est la probabilité corrigée d'avoir la base  $b$  en position  $i$  du motif, avec

$f_{b,i}$  le nombre de bases  $b$  en position  $i$ ,  $s(b)$  la fonction *pseudocount* qui permet d'ajouter une valeur faible au nombre  $f_{b,i}$  pour ne pas avoir de fréquence nulle et  $N$  le nombre de séquences alignées (exemple sur la Figure 1.26). La matrice PWM est ensuite construite en divisant les probabilités de chaque nucléotide à chaque position obtenues en équation 1.1 par les probabilités attendues (*background*), puis en convertissant à une échelle log :

$$W_{b,i} = \log_2 \frac{p(b, i)}{p(b)} \quad (1.2)$$

avec  $p(b)$  la probabilité *background* de la base  $b$ .

Cette notation permet de prendre en compte le *background* et donc les probabilités hétérogènes d'obtenir chacune des bases.

Une alternative aux PWMs pour représenter les motifs de fixation des TFs se fait par l'utilisation de modèles de Markov cachés qui permettent de prendre en compte plus de subtilités qu'avec les PWM [177]. Il est par exemple possible de prendre en compte la dépendance entre les positions des motifs ou d'effectuer des insertions ou des délétions. Les *Transcription factor flexible models* (TFFMs) en sont un exemple [170]. Avec le nombre important de données de ChIP-seq disponibles, ces modèles permettent d'améliorer les performances de prédiction pour les TFs ayant des caractéristiques de fixation flexibles et permettent de mieux discriminer le vrai signal de ChIP-seq du *background*. Ces modèles permettent également de calculer un score d'occupation des TFs le long d'une séquence.

Pour capturer les préférences de fixation des TFs, il existe des méthodes *in vivo* et des méthodes *in vitro* [82]. *In vivo*, la méthode la plus répandue est le ChIP-seq qui permet de cibler directement un TF d'intérêt et de séquencer les portions d'ADN sur lesquelles ils sont fixés. Cependant, cette technique dépend de l'abondance en protéine, de l'efficacité de la réticulation et de la spécificité de l'anticorps choisi. De plus, il est possible de capturer des complexes et on ne sait pas si le TF que l'on regarde est réellement fixé sur l'ADN. *In vitro*, la technique généralement utilisée est la méthode SELEX (*Systematic Evolution of Ligands by EXponential enrichment*) [207] qui se base sur le principe suivant : les protéines d'intérêt sont incubées en présence d'un grand nombre d'oligonucléotides générés aléatoirement et viennent se fixer au niveau des motifs qu'elles reconnaissent. Les complexes sont ensuite extraits par immunoprecipitation et les séquences reconnues sont amplifiées. Une autre technique *in vitro* répandue est celle du *Protein binding microarray* (PBM) [82] où une protéine d'intérêt est mise en contact avec une puce sur laquelle de nombreux oli-

## Data collection and alignment

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	T	A	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	T	A	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	G	G	G	A
Site 5	T	G	C	C	A	A	A	G	T	G	G	T	C	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	C	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	C	C
Site 8	C	T	C	C	T	A	C	A	T	G	G	G	C	C

Source binding sites

## Number of observed nucleotides at each position

Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

## Position Weight Matrix: normalized frequency matrix converted to a log scale

Position weight matrix (PWM)

<b>A</b>	-1.93	0.79	0.79	-1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	-1.93	-1.93	0.79
<b>C</b>	0.45	-1.93	0.79	1.68	-1.93	-1.93	-1.93	0.45	-1.93	-1.93	-1.93	-1.93	0.00	0.79
<b>G</b>	0.00	0.45	-1.93	-1.93	-1.93	-1.93	-1.93	-1.93	0.66	-1.93	1.30	1.68	1.07	-1.93
<b>T</b>	0.15	0.66	-1.93	-1.93	1.07	0.66	0.79	0.00	0.00	0.79	-1.93	-1.93	-0.66	-1.93

## Visualization with a sequence logo

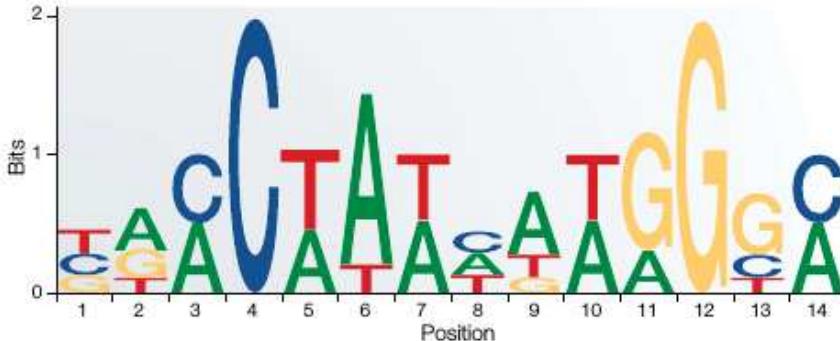


FIGURE 1.26 – Modèle de représentation d'un motif sous forme de matrice de probabilité (PWM). Un ensemble de séquences, obtenues par séquençage des fragments d'ADN sur lesquels le facteur de transcription d'intérêt est fixé, sont collectées et alignées. Une matrice contenant le nombre d'observations de chaque nucléotide à chaque position du motif est créée, appelée matrice de fréquence. Enfin, cette matrice est normalisée et passée au log. [Figure adaptée de [274]]

gonucléotides double brins sont fixés. Les événements de fixation sont identifiés par immuno-détection en utilisant un anticorps spécifique couplé à un fluorophore.

Plusieurs databases recensent directement les PWMs associées aux différents TFs chez les vertébrés comme Transfac [278], JASPAR [229], HOCOMOCO (Homo sa-

piens COmprehensive MOdel Collection ) [146] et cisBP [275]. Pour JASPAR, les profils des sites de fixation des TFs sont vérifiés manuellement. Ils se présentent sous la forme de matrices de fréquences (*Position frequency matrices*, PFM) et de TFFM [170], dans 6 groupes taxonomiques différents dont les vertébrés. Depuis sa création en 2004, la base de données a été mise à jour plusieurs fois, la dernière version est de 2018 et comporte 579 PFM pour les vertébrés [130]. La version v11 de HOCOMOCO intègre les modèles de TFBSS pour 680 TFs chez l'homme et 453 chez la souris sous forme de PWMs établies à partir de données expérimentales de ChIP-seq et de HT-SELEX. En plus des PWMs classiques, HOCOMOCO fourni également des PWMs considérant les fréquences des dinucléotides au lieu de simples nucléotides.

Pour faciliter l'utilisation des motifs de fixation des TFs (connus et prédicts) la base de données CIS-BP (Catalog of inferred sequence binding preferences, CIS-BP) a été créée en 2014 (<http://cisbp.ccbr.utoronto.ca>) [275]. Ce catalogue intègre des données provenant d'environ 300 espèces et couvrant plus de 250 familles de TFs. Les données de CIS-BP proviennent de plusieurs sources dont Transfac, JASPAR et HOCOMOCO et contient des scores de fixation de TFs pour des 8-mers, des PWMs et des motifs consensus.

### 1.3.2 Transcription par l'ARN polymérase II

La transcription de la matrice ADN en ARN est réalisée par des ARN polymérases. On en compte 3 principales chez les eucaryotes : ARN polymérases I, II et III. L'ARN polymérase II synthétise l'ARN messager, précurseur de la protéine mais elle transcrit aussi de nombreux autres ARNs non codants. Ici nous nous intéressons à la transcription par l'ARN polymérase II.

#### Structure de l'ARN polymérase II

L'ARN polymérase II est constituée d'une douzaine de sous unités numérotées de 1 à 12 dont deux sous-unités plus larges que les autres : la RPB1 (encodée par le gène POLR2A) et la RPB2 (encodée par le gène POLR2B). La RPB1 possède un domaine carboxy-terminal (CTD) formant une queue et permet entre autre à l'ARN Pol II de se fixer sur sa matrice ADN en interagissant avec de nombreux facteurs. Les différentes phases de phosphorylation du CTD au cours des étapes clés de la transcription permettent le recrutement des facteurs nécessaires.

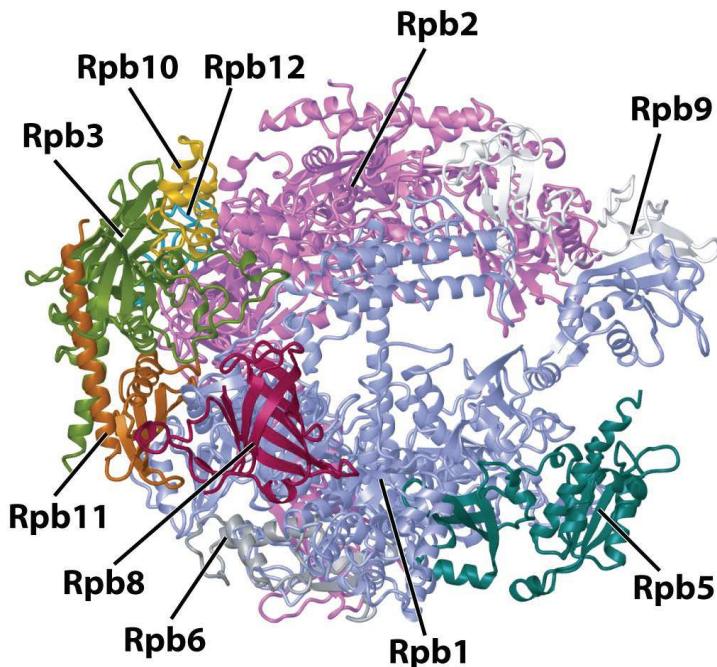


Figure 21-10 Principles of Biochemistry, 4/e  
© 2006 Pearson Prentice Hall, Inc.

**FIGURE 1.27 – Structure du complexe protéique de l’ARN polymérase II.** Représentation 3D de la polymérase II et ses différentes sous unités. [Figure extraite de l’ouvrage ”Principles of Biochemistry”, 2006]

### Différentes étapes de la transcription

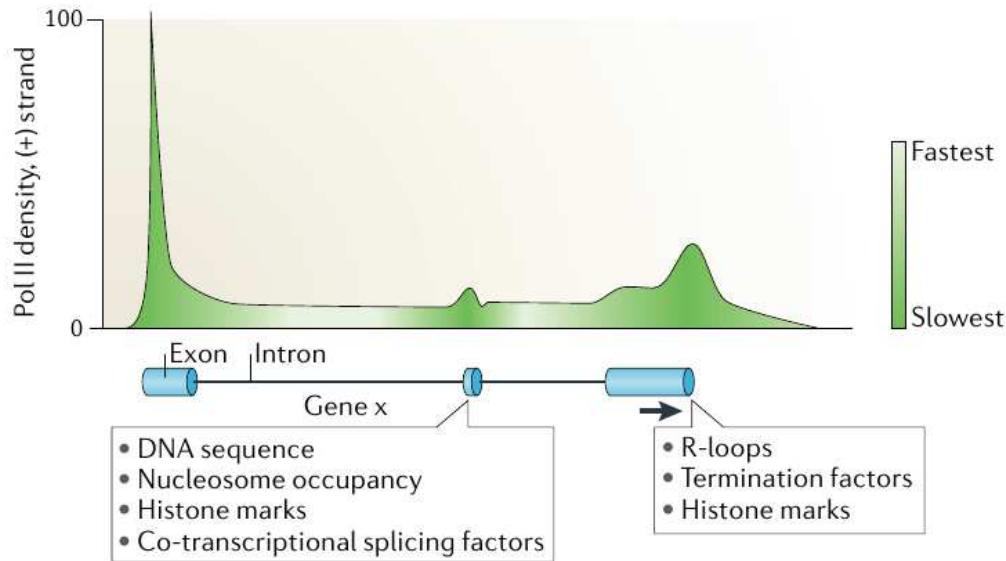
**Initiation.** La transcription par l’ARN Pol II est contrôlée par des facteurs de transcription généraux. Elle débute par une phase que l’on appelle initiation, pendant laquelle le complexe d’initiation de la transcription est formé par assemblage de l’ARN Pol II avec les GTFs (TFIIA, TFIIB...). En effet, l’ARN Pol II des eucaryotes ne reconnaît pas seule la séquence sur laquelle elle doit se fixer pour transcrire. La formation du PIC suit une série d’événements. TFIID est impliqué dans la liaison du PIC à la région promotrice. TFIIA stabilise ensuite la liaison du complexe à l’ADN et interagit avec les activateurs/inhibiteurs de la transcription [42]. Le facteur général TFIIH, via une activité hélicase, enclenche l’ouverture de la double hélice d’ADN. Il possède également une activité kinase responsable de la phosphorylation post-traductionnelle des séries 5 et 7, présentes dans la queue C-terminale de l’ARN Pol II composée d’une 50<sup>aïne</sup> de répétitions de la séquence *Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7*. Cette phosphorylation provoque la dissociation du complexe d’initiation et l’entrée dans la phase d’elongation [58].

**Élongation.** Au cours de l’elongation, c’est la sérine 2 qui est à son tour phosphorylée. Durant cette phase, l’ARN Pol II copie le brin complémentaire au brin

codant (5' vers 3') contenant la séquence d'intérêt, de manière à obtenir une copie conforme de la région à transcrire. L'incorporation des nucléotides se fait par complémentarité des bases. Cette étape joue un rôle crucial dans la régulation de l'expression des gènes puisque la quantité d'ARN à un instant précis dépend directement de la productivité de la polymérase. Pour que la polymérase puisse progresser sur sa matrice ADN, les nucléosomes doivent être temporairement retirés, même si dans certains cas un désassemblage partiel suffit. Il existe une série de facteurs facilitant le passage de la polymérase comme les facteurs de remodelage de la chromatine et les chaperons d'histones [237].

Des méthodes pour suivre la cinétique de la polymérase dans un type cellulaire donné ont été développées ces dernières années afin de mieux comprendre les facteurs qui influencent sa progression. Dans un premier temps, la polymérase est généralement bloquée temporairement au niveau des TSSs par l'action d'une drogue (exemple de drogue largement utilisée : 5,6-dichlorobenzimidazole 1- $\beta$ -D-ribofuranoside, DRB), ce qui permet de vider le génome des polymérases en cours d'elongation et de synchroniser l'initiation de la transcription. Une fois l'action de la drogue relâchée, le front de polymérase se déplaçant est suivi sur différents temps. Cette étape peut par exemple être réalisée en utilisant le GRO-seq (*Global run-on sequencing*) [43] qui permet de cibler les sites où l'ARN Pol II est transcriptionnellement active [119, 48], ou encore par ChIP-seq contre l'ARN Pol II [271]. Toutes ces techniques et études de l'elongation permettent de faire le lien entre l'elongation de l'ARN Pol II et de nombreux facteurs pouvant l'influencer comme les marques épigénétiques, la proportion introns/exons ou la quantité de CG. Ainsi, il semblerait que les taux d'elongation soient plus faibles à proximité du promoteur et dans les exons que dans le reste du gène (voir Figure 1.28), et également en présence de séquences terminales longues répétées et d'une forte méthylation de l'ADN. Les marques épigénétiques H3K79me2 et H4K20me1 peuvent influencer positivement l'elongation. Enfin, la polymérase semble ralentir au niveau des jonctions introns/exons, ce qui pourrait avoir un impact sur l'épissage co-transcriptionnel et favoriser la diversité des transcrits produits [76].

**Terminaison.** La terminaison de la transcription est généralement couplée à la polyadénylation qui est un processus post-transcriptionnel où une queue poly-A est ajoutée en aval de l'ARNm transcrit. Ce processus est associé à un ralentissement de l'ARN Pol II suivi du clivage du transcrit au niveau du motif de signal poly-A qui est une séquence AAUAAA. Le fragment d'ARN restant est dégradé par une



**FIGURE 1.28 – Profil de densité de l'ARN Polymérase II pour un gène hypothétique.**  
Densité de l'ARN polymérase II et taux d'élongation varient le long du gène. Sur l'axe des y, la densité en ARN Pol II est indiquée et l'échelle de couleur de vert à blanc représente le taux d'élongation. L'ARN Pol II est plus lente au niveau du promoteur (pause) et accélère une fois rentrée en phase d'élongation. Il semblerait également que l'ARN Pol II ralentisse à la rencontre des exons et soit en pause au site de terminaison. Le taux de GC, l'occupation par les nucléosomes et les marques épigénétiques semblent également être corrélés à la vitesse d'élongation. [Figure extraite de [120]]

exonucléase recrutée au niveau du site de polyadénylation, dans le sens 5' vers 3' [210]. Comme nous l'avons vu dans la partie 1.2.2, il existe des sites de polyadénylation alternatifs entraînant une terminaison de la transcription pouvant être effectuée à différentes positions. Le mécanisme selon lequel la transcription se termine et le choix du site de polyadénylation ne sont pas bien connus. Une technique de séquençage du transcriptome temporaire (*Transient transcriptome sequencing*, TT-seq) est développée en 2016 par Schwalb et al.[234] Cette méthode collecte et séquence tous les ARNs synthétisés sur une courte durée de 5 minutes, ce qui lui permet de détecter une grande partie des ARNs peu stables et/ou peu exprimés. Elle a permis de mettre en évidence la transcription de la polymérase en aval du site de polyadénylation sur une fenêtre de plusieurs kb, dans laquelle on trouve en moyenne 4 sites de terminaison.

### 1.3.3 Quantification de l'expression des gènes

Dans le cadre de ma thèse, l'expression des gènes désigne le nombre de transcrits produits à partir des gènes et qui sont présents dans la cellules à un instant t. A partir d'une méthode choisie qui cible tous les ARNs ou un sous ensemble d'ARNs ayant

une propriété commune, il est possible de séquencer ces ARNs, retrouver les gènes dont ils proviennent et quantifier pour chaque gène le nombre de transcrits associés. Cette expression dépend des régulations transcriptionnelles et des régulations post-transcriptionnelles, mais selon le gène, n'est pas forcément corrélée à la quantité de protéines. Il existe de nombreuses méthodes pour quantifier l'expression des gènes à l'échelle du génome.

### RNA-seq

Le RNA-seq est la technique la plus répandue pour étudier le transcriptome qui correspond à l'ensemble des ARNs présents dans une cellule à un instant donné. Le transcriptome varie au cours du temps, selon le type cellulaire et sous l'action de l'environnement. Le RNA-seq est une technique permettant de caractériser et quantifier les transcrits sur toute leur longueur.

**Protocole du RNA-seq.** Après avoir récolté les ARNs d'un ensemble de cellules dans une condition donnée, ils sont fragmentés de manière aléatoire, avec une taille pouvant aller jusqu'à quelques centaines de paires de bases. Aux extrémités des fragments d'ARN obtenus, des adaptateurs sont ajoutés [185]. On obtient ainsi une librairie que l'on dépose sur un support solide : la *flowcell*. Cette étape permet d'amplifier chaque fragment par hybridation sur les oligonucléotides de la *flowcell* que l'on peut ensuite fournir au séquenceur. La lecture par le séquenceur se fait base par base, par extension des fragments via des nucléotides couplés à des fluorochromes qui, lors de leur fixation, libèrent un signal de fluorescence. Les fragments sont ainsi numérisés et renvoyés sous forme de lectures que l'on appelle *reads*, d'une taille limitée par la vitesse du séquenceur et par son taux d'erreur. Ces lectures sont toutes encodées dans un fichier FastQ qui contient pour chacune, son identifiant, la séquence elle-même qui est une suite de A, T, G et C ou N quand la base lue est indéterminée et le code qualité. Avec les *reads* obtenus et pour un organisme étudié relativement bien annoté comme l'homme, il est possible de les aligner sur le génome de référence. De nombreux outils d'alignement existent. L'application intéressante dans le contexte de ma thèse est la quantification de l'expression des gènes.

**Quantification de l'expression des gènes et des transcrits.** La quantification de l'expression des gènes s'effectue par comptage du nombre de *reads* qui s'alignent sur chacun des gènes dont les bornes sont définies par les annotations, sous l'hypothèse que le nombre de *reads* associé à un certain gène soit proportionnel à l'abondance de son ARN correspondant dans la cellule. Parfois, les *reads* obtenus

proviennent d'une jonction d'épissage, c'est à dire qu'il sont à cheval sur deux exons, et on ne les retrouve pas directement sur le génome de référence. Des outils existent pour les aligner et les prendre en compte dans le comptage. L'expression des gènes représentée par le comptage des *reads* peut être exprimée en plusieurs unités. La plus populaire est le RPKM (*Reads per kilobase per million*). Exprimer l'expression en RPKM permet de normaliser les résultats par la profondeur du séquençage (nombre total de reads dans l'expérience) et par la longueur des gènes. Il existe des mesures alternatives similaires au RPKM comme le FPKM (*Fragments per kilobase per million*) qui est adapté au séquençage *paired-end* où deux reads peuvent correspondre à un même fragment sans qu'il soit compté 2 fois et le TPM (*Transcripts per kilobase per million*). Il est également possible d'estimer l'expression de chaque transcript. Afin d'optimiser la distribution des *reads* qui s'alignent sur plusieurs transcrits, des algorithmes sont développés comme la méthode des moindres carrés ou l'algorithme espérance-maximisation (*Expectation Maximization*, EM) qui comprend une fonction de vraisemblance. Ce dernier est utilisé dans l'outil RSEM [154]. Malgré les algorithmes existants, il n'est pas facile de quantifier les *reads* provenant de chaque transcrit, les différents transcrits étant en grande partie chevauchants.

## CAGE

La technologie de CAGE permet de caractériser l'extrémité 5' des transcrits possédant une coiffe [137]. Toutes les petites séquences obtenues sont généralement d'une longueur de 20 nucléotides et sont appelées tags. Le principe du CAGE est le suivant : l'ARN est tout d'abord soumis à une étape de transcription inverse des ARNs en ADNc. La coiffe en 5' est biotinylée et l'ARN est digéré en 3' par une RNase. Les molécules ARN/ADNc biotynilées sont isolées via la protéine streptavidine, libérées, puis un adaptateur est ajouté. Pour le séquenceur Heliscope utilisé par le consortium FANTOM, c'est une séquence poly-A qui est ajoutée en 3' de l'ADNc en tant qu'adaptateur. Enfin, ces fragments sont chargés sur la *flowcell* qui contient des oligonucléotides poly-T permettant l'hybridation des queues poly-A des ADNc. Comme le génome humain contient de nombreuses séquences poly-T, cette technique capture les ADNc poly-adénylés mais peut aussi capturer les séquences poly-A internes (voir Figure 1.29).

## Capturer l'ARN naissant

De nombreuses méthodes sont aujourd'hui utilisées pour quantifier l'expression des gènes. Les deux techniques présentées ci-dessus permettent de capturer l'ARN généralement stable et décroché de la chromatine. Même si le CAGE permet de

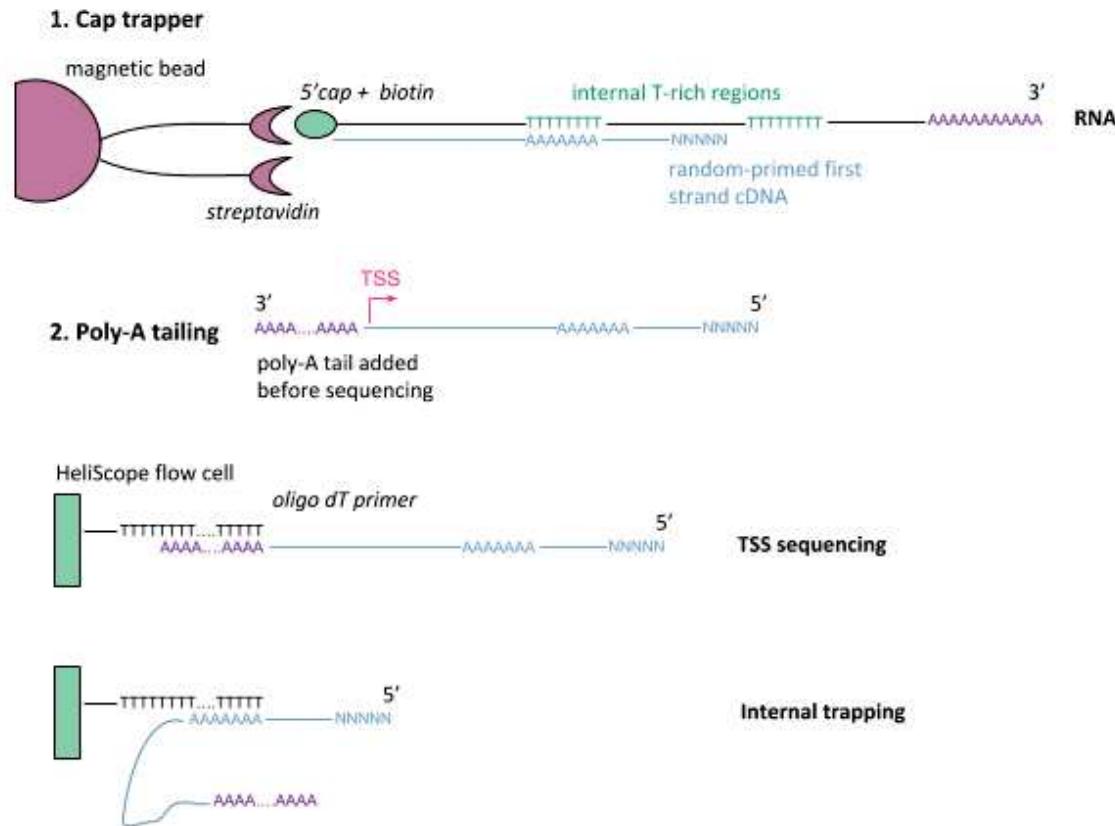


FIGURE 1.29 – Protocole du CAGE avec la technologie Heliscope. Après synthèse de l'ADNc, les hybrides ARN/ADN sont biotinyrés. Ces molécules sont ensuite capturées via des billes magnétiques associées à la streptavidine. Les ADNc simple brin sont relâchés et une queue poly-A est ajoutée en amont du TSS. C'est cette queue poly-A qui va permettre l'hybridation sur la *flowcell* HeliScope sur laquelle des oligonucléotides poly-T sont présents. Ces oligonucléotides peuvent aussi capturer des séquences poly-T internes (poly-A sur l'ADNc).

mieux caractériser l'extrémité 5' des ARNs et donc d'améliorer la détection des TSSs en comparaison au RNA-seq, ces deux techniques ne sont pas adaptées pour capturer les ARNs en cours de transcription ou les ARNs vite dégradés. En effet, elles mesurent la quantité globale, dans un état d'équilibre, de tous les ARNs, y compris mûris. Il existe des méthodes pour capturer les ARNs naissants, comme le GRO-seq [43] mentionné dans la partie 1.3.2 et le NET-seq (*Native elongating transcript sequencing*). Pour le GRO-seq, les ARNs Pol II transcrivent en présence de 5-bromouridine 5'-triphosphate (Br-UTP) pour produire des ARNs marqués à la BrdU (5-bromo-2-deoxyuridine). Ces derniers sont ensuite isolés par des billes magnétiques couplées à un anticorps de la BrdU, purifiés et séquencés. Le NET-seq permet également de capturer les ARNs naissants par immunoprecipitation des ARNs Pol II en cours d'elongation suivi du séquençage des extrémités 3' des fragments d'ARN liés à ce complexe d'elongation. Cette méthode permet d'obtenir un alignement des transcrits naissants à une résolution d'un nucléotide [32].

## Mesure de la quantité d'ARNs traduits

Comme nous venons de le voir, la régulation de la transcription des gènes est étroitement régulée par l'environnement épigénétique et les facteurs de transcription se fixant au niveau d'éléments cis-régulateurs particuliers. Pour les gènes codants pour des protéines, en plus des contrôles transcriptionnels, l'expression des protéines est contrôlée au niveau de la traduction. La technique du Ribo-seq (*Ribosome sequencing*) a été développée en 2009 [109] pour mesurer la quantité de ribosomes liés à l'ARN en cours de traduction. Après blocage des ribosomes sur leur matrice, les fragments d'ARNs occupés et protégés sont clivés et séquencés (voir Figure 1.30).

L'aperçu des méthodes principales pour la détection des ARNs matures ou naissants est présenté sur la figure ci-dessous (Figure 1.30).

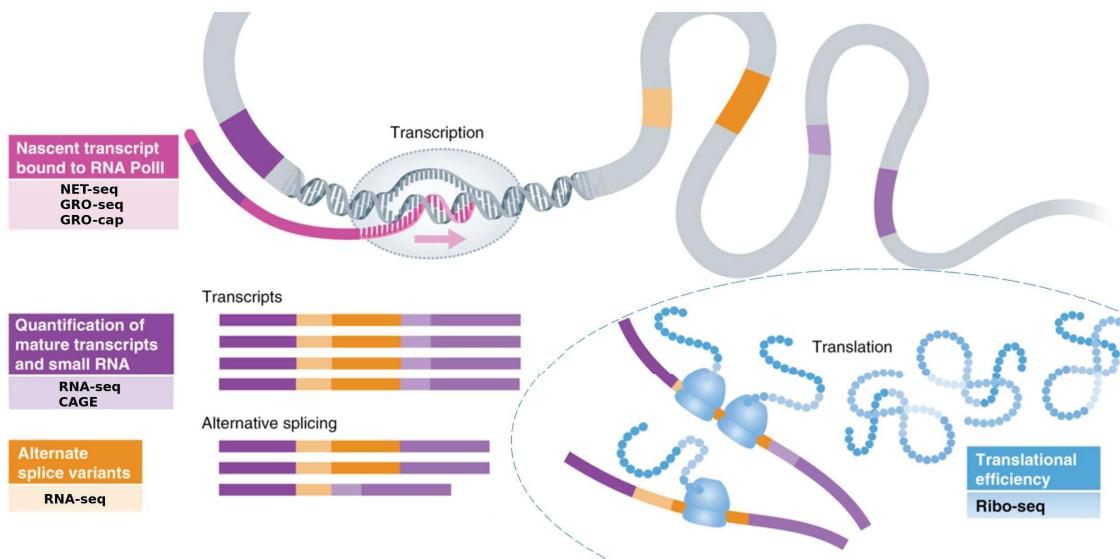


FIGURE 1.30 – Principales technologies de séquençage permettant la quantification du transcriptome. [Figure adaptée de [246]]

Dans la partie suivante, nous aborderons les variations existantes entre les génomes de différents individus et l'impact qu'elles peuvent avoir au niveau phénotypique.

## 1.4 Dérégulation des contrôles de l'expression des gènes : variants et pathologies

Chaque individu possède son propre génome et contribue à la diversité de notre espèce. Mais, si on compare deux génomes humains, on observe 99,9% de séquences identiques [The Human Genome Project]. La spécificité de chacun se trouve donc dans les 0,1% restants. Bien que l'idée d'un individu un génome soit encore très présente, il existe également des variations au sein d'un même individu. Le séquençage en cellule unique (*Single-cell*) permet de les détecter et d'étudier leurs effets intra-individu [145]. Comme nous l'avons vu dans la partie précédente, la régulation de l'expression des gènes joue un rôle fondamental pour le bon fonctionnement des cellules et pour leur adaptation aux différents stimuli extérieurs. Cependant, dans le cas de maladies, le contrôle de l'expression des gènes est dérégulé, et de tels dysfonctionnements peuvent survenir suite à la présence ou l'apparition de mutations. Les progrès des techniques d'analyses de ces dernières années permettent d'accéder au patrimoine génétique de chaque individu et de le croiser avec des informations cliniques comme les symptômes d'une maladie. Il est ainsi possible d'établir plus facilement un diagnostic sur des maladies génétiques ou d'estimer les risques d'une personne à développer un cancer par exemple. Les altérations possibles du génome sont de plusieurs types que nous allons voir dans la présente partie.

### 1.4.1 Altérations du génome

#### Variations génomiques

Le terme de variants génomiques décrit toutes les différences qu'il peut y avoir entre deux individus, mais qui peut aussi, avec l'émergence des variations intra-individuelles, faire référence aux variants observés d'une cellule à l'autre. On distingue les variations ponctuelles qui n'affectent qu'une seule base, les variations structurales qui sont des altérations génomiques sur des segments de plus de 1 kb [71] et les aneuploïdies qui sont caractérisées par une variation du nombre d'un ou de plusieurs chromosomes.

**Substitution nucléotidique.** Parmi les variations ponctuelles, on retrouve le *Single Nucleotide Variant* (SNV) qui correspond au changement d'un seul nucléotide à une position particulière. On parle de *Single Nucleotide Polymorphism* (SNP) quand on retrouve le variant chez plus d'1% de la population. Un SNP apparaît en moyenne tous les 300 nucléotides [239]. Quand ces variations se trouvent dans la par-

tie codante des gènes, elles peuvent avoir un impact sur la protéine produite. Elles sont ainsi classées en fonction de leur effet : silencieux quand il n'y a aucun impact, faux-sens quand cela modifie l'acide aminé codé ou non-sens quand cela génère un codon stop prématué. Dans la partie non-codante du génome, évaluer l'impact des variations ponctuelles est un challenge actuel. En effet, les mécanismes d'action des éléments régulateurs du génome non codant ne sont pas encore bien connus. Quand un variant est présent dans le génome des cellules tumorales, en comparaison au génome de cellules germinales, on parle de mutation [125]. Une mutation peut être rare et il est difficile de différentier les mutations directives (*driver mutations*) qui sont causales des passagères (*passenger mutations*), surtout dans la partie non-codante du génome.

**Insertions/délétions.** On peut également observer des insertions ou délétions de plusieurs bases consécutives regroupées sous le nom d'*indels*. Ces variations ne sont pas forcément délétères pour la cellule et leur impact va dépendre de plusieurs facteurs comme la nature du variant ou sa localisation [71]. Par contre, elles ont un impact direct sur la longueur de la portion d'ADN dans laquelle on les trouve.

**Variation du nombre de copies.** Les variants structuraux sont définis comme des altérations génomiques sur des segments de plus de 1 kb [71]. La variabilité du nombre de copies (*Copy number variation*, CNV) qui fait partie des variants structuraux correspond à la répétition de portions plus ou moins grandes du génome. Comme pour les variations d'une base, on parle de polymorphisme quand le CNV est présent chez plus d'1% de la population. Ces variations sont retrouvées chez de nombreux mammifères et les CNVs les mieux caractérisés sont les répétitions de tri-nucléotides (*trinucleotide repeats*, TNRs) [200]. Il existe d'autres types de variants structuraux touchant des portions de chromosomes que nous ne détaillerons pas ici [136].

**Short tandem repeats.** Les microsatellites aussi appelés *short tandem repeats* (STRs) sont de courtes séquences d'ADN (1 à 6 pb) répétées consécutivement un certain nombre de fois. Il a été estimé qu'environ 3% du génome humain est représenté par les STRs [33]. Les STRs sont fréquemment mutés, avec des taux de mutation plus élevés que les taux moyens à travers l'ensemble du génome [83]. Les mutations des STRs ont été associées à plus de 30 pathologies comme les troubles neurologiques ou troubles du développement [182]. Leur contribution dans la régulation de l'expression des gènes chez l'homme a également été étudiée et a permis

de mettre en évidence plus de 2000 *expression STR* (eSTR) [91].

### Données de variants génomiques

**dbSNP.** dbSNP est une base de données de SNPs humains hébergée par le NCBI. Cette collection de polymorphismes comprend des SNPs, des petites insertions et délétions, des éléments rétroposables qui sont des fragments répétés de l'ADN provenant de molécules d'ARN inversement transcrrites (voir partie 1.2.5) et enfin des microsatellites aussi appelés *short tandem repeats* [135]. Chaque entrée de cette base de données est associée à plusieurs informations, comme le contexte nucléotidique de la séquence entourant le polymorphisme, l'occurrence du polymorphisme et les conditions expérimentales.

**Consortium GTEx.** Le consortium GTEx pour "Genotype-Tissue Expression" a été lancé en 2010 pour étudier l'expression et la régulation des gènes chez l'homme dans de nombreux tissus post-mortem [164]. En recueillant un grand nombre d'échantillons, il est ainsi possible d'analyser les effets des variations génomiques propres à chaque tissu et leur lien avec l'expression des gènes. Aujourd'hui, l'impact des modifications présentes dans notre code ADN est mieux compris. Une des découvertes du consortium est qu'un même variant présent dans de multiples tissus peut avoir un effet différent selon le tissu concerné.

Dans le cadre de mon projet de thèse, nous nous intéressons tout particulièrement aux associations entre variants génomiques et expression des gènes. Le terme d'association se réfère à la co-occurrence d'un variant génomique et d'un caractère phénotypique spécifique plus fréquemment que ce que l'on pourrait observer par hasard. Les approches de QTL (*quantitative trait loci*) peuvent être appliquées à n'importe quel caractère qui peut être quantifié et qui possède un locus défini dans le génome [22]. Il existe ainsi des analyses de QTL d'épissage (*splicing QTL*), du niveau d'expression protéique (*protein QTL*), de la méthylation (*methylation QTL*) ou encore des modifications d'histones (*histone QTL*). Pour évaluer l'influence d'un variant sur le niveau d'expression d'un gène, il existe l'analyse eQTL (*expression quantitative trait loci*). Les eQTLs sont mis en évidence en collectant simultanément des données de variations génétiques et d'expression pour de nombreux individus et en regardant pour chacun les potentielles associations entre génotype et niveau d'expression du gène d'intérêt. Généralement, ces associations sont calculées pour tous les variants situés à proximité du gène, pour trouver de potentiels variants *cis*-régulateurs. Les associations en *trans*, c'est à dire éloignées des gènes cibles sont

plus difficiles à mettre en évidence. Comme de nombreuses études ont mis en avant l'importance du génome non-codant dans la régulation des gènes, il est nécessaire d'avoir une vision globale des variants pouvant affecter l'expression des gènes dans une condition particulière. Les analyses à l'échelle du génome (*Genome-wide association studies*, GWAS) ont été développées pour comprendre et identifier les gènes dont l'expression est affectée par les variations génomiques [23]. Elles permettent de tester des associations statistiques entre des polymorphismes et la variabilité d'un caractère quantitatif (comme l'expression d'un gène) sur un ensemble d'individus à priori non apparentés [167].

dbGaP, appartenant à GTEx, est une base de dépôt publique comprenant des phénotypes, génotypes, caractéristiques (comme fumeur ou non), l'âge et des données de séquences. Elle représente une source de données importante dont l'accès public est limité au résumé des informations. En accès contrôlé, l'expression des gènes et les génotypes associés sont disponibles.

### 1.4.2 Variants et maladies

#### Variants dans les cancers

Le cancer est un groupe de maladies qui se caractérise par une prolifération de cellules anormales au sein d'un tissu. Les cellules cancéreuses sont caractérisées par une division cellulaire invasive. Contrairement aux cellules normales qui meurent automatiquement après un certain nombre de divisions, les cellules cancéreuses se divisent sans fin : elles sont immortelles. Lors de l'évolution de la maladie, elles peuvent endommager les organes ou tissus dans lesquels elles se trouvent mais peuvent aussi se détacher de la tumeur et migrer vers d'autres tissus et organes. Ces caractéristiques propres aux cellules tumorales sont liées à l'activation/inactivation de gènes particuliers : sur-expression des oncogènes régulant positivement la prolifération cellulaire, sous-expression/inactivation des gènes suppresseurs de tumeurs ou encore des gènes du système de réparation de l'ADN. Ainsi, chaque cancer est caractérisé par des erreurs présentes dans le génome du patient atteint comme les mutations somatiques, la variation du nombre de copies, etc. Un exemple aujourd'hui bien connu est celui de la mutation des gènes BRCA1 et BRCA2, gènes du système de réparation, qui augmentent considérablement le risque de contracter un cancer du sein [181]. Les mutations associées à chaque cancer varient et des études ont pu associer des signatures de mutations à chaque type et sous-type de cancer [149, 2]. Aujourd'hui, le profil d'expression d'un grand nombre de tumeurs a été

établi chez différents patients et des bases de données regroupent ces informations.

**The Cancer Genome Atlas (TCGA).** The Cancer Genome Atlas (TCGA) (<http://cancergenome.nih.gov/>) est le fruit d'une collaboration entre le National Cancer Institute (NCI) et l'Institut National de Recherche sur le Génome Humain (National Human Genome Research Institute, NHGRI), les deux parties de la National Institutes of Health. Avec l'International Cancer Genome Consortium (ICGC), ce sont les deux principaux projets fournissant des informations cliniques et une quantité considérable de données issues d'analyses par séquençage haut débit de génomes tumoraux. TCGA analyse un très grand nombre d'échantillons pour chaque cancer étudié pour avoir des informations sur le profil génomique de chaque cancer tout en ayant une signification au niveau statistique. De plus, pour chaque patient étudié, un échantillon de tissu affecté et un échantillon normal (en général sang) sont analysés, ce qui permet d'identifier les changements au niveau du génome qui joueraient un rôle dans le développement du cancer. Le but de ce projet est ainsi de comprendre le lien entre les variations génomiques et les différents types et sous-types de cancers [262].

### Variants et autres maladies

Le cancer est la première cause de décès en France pour les moins de 65 ans. C'est une maladie complexe et plus de 200 formes ont été décrites avec pour chacun de nombreux sous-types [93]. Cependant, d'autres maladies peuvent être associées à des variations du génome. Ainsi, des maladies comme l'autisme chez les enfants ont pu être corrélées à des mutations particulières [113]. Un autre exemple est celui de la maladie d'Alzheimer qui est une maladie neurodégénérative caractérisée par un déclin global et progressif des facultés cognitives (apprentissage, langage, compréhension...) et de la mémoire. Parmi les facteurs à risque d'apparition de la maladie d'Alzheimer, on retrouve deux protéines : la Translocase of Outer Mitochondrial Membrane 40 Homolog et l'Apolipoprotéine E (TOMM40-APOE), dont les loci ont été associés au déclin cognitif non pathologique et à la maladie d'Alzheimer [199]. Il semblerait que la longueur d'un variant poly-T (rs10524523), qui est un STR, soit corrélée à la maladie avec des longueurs importantes du poly-T associées à un risque accru de développer la maladie d'Alzheimer.

## Conclusion

Au fil de ce chapitre, nous avons eu un aperçu des différents niveaux de régulation de l'expression des gènes. La régulation de leur expression est ainsi assurée par une combinaison complexe d'éléments. Une grande portion du génome reste encore largement inexplorée, et bien qu'elle soit en grande partie transcrrite, on ne comprend pas encore le rôle des ARNs non-codants qui sont produits. Les résultats de ma thèse sur l'étude des éléments régulateurs de l'expression des gènes sont présentés dans le chapitre suivant.

# Chapitre 2

## Résultats

### 2.1 Instructions de régulation de l'expression des gènes présentes dans la séquence ADN

Comme nous l'avons vu précédemment, la régulation de l'expression des gènes est orchestrée par de nombreuses régions régulatrices qui interagissent entre elles via des protéines et qui permettent d'assurer la spécificité des différentes cellules de notre organisme. Pour mieux comprendre les mécanismes à l'origine de cette régulation, je me suis intéressée, dans la première partie de ma thèse, à la prédiction de l'expression des gènes uniquement à partir de l'information contenue dans notre séquence ADN pour différents types de cancers. Des approches ont déjà été développées pour prédire l'expression des gènes mais reposent généralement sur des données expérimentales. Elles seront présentées dans une deuxième sous-partie, après avoir défini le modèle de régression linéaire, suivies de ma contribution au projet et résultats.

#### 2.1.1 Choix du modèle statistique

Les outils statistiques nous aident à décrire et interpréter des données quantitatives, mais aussi à modéliser et prédire certains phénomènes. Il existe de nombreux outils statistiques applicables à des données d'expression de gènes, que l'on choisi en fonction du type de données à analyser et des hypothèses faites sur celles-ci [273]. L'approche que nous choisissons d'utiliser est la régression linéaire. En effet, il s'agit d'un modèle simple qui nous permet d'effectuer une approche exploratoire sur l'ensemble des données disponibles (composition nucléotidique, motifs) malgré l'hypothèse de départ très forte d'une relation linéaire entre expression des gènes et information contenue dans la séquence ADN.

## Modèle de régression linéaire multiple

Dans tous les problèmes rencontrés en statistiques, on a des variables non aléatoires que l'on note ici sous forme de vecteur :  $X = (X_1 \dots X_p)$  dites explicatives. Chaque variable  $X_i$  contient  $N$  observations soit  $X_i = (x_{1i} \dots x_{Ni})^T$  et  $Y$  est une variable aléatoire dite expliquée. Le but est de trouver une relation entre  $Y$  et  $X$ , soit  $Y = f(X_1 \dots X_p)$ . On dira que le modèle est de régression linéaire si :

$$f(X_1 \dots X_p) = \beta_0 + \beta_1 * X_1 + \dots + \beta_p * X_p + \epsilon = \beta_0 + \sum_{j=1}^p (\beta_j * X_j) + \epsilon$$

avec  $\beta_0 \dots \beta_p$  des paramètres inconnus supposés constants à estimer. On ajoute un terme d'erreur  $\epsilon$  qui représente l'erreur du modèle (erreur liée à la technique, erreur par omission de certaines variables...). L'écriture matricielle du modèle est la suivante :  $Y = X\beta + \epsilon$ . Les paramètres  $\beta_0 \dots \beta_p$  sont généralement estimés par la méthode des moindres carrés qui consiste à minimiser la somme des carrés des résidus (*Residual sum of squares*, RSS), c'est à dire, minimiser l'erreur entre le modèle et ce que l'on observe réellement :

$$RSS(\beta) = \sum_{i=1}^N \epsilon_i^2 = \sum_{i=1}^N (y_i - (\beta_0 + \sum_{j=1}^p \beta_j x_{ij}))^2$$

Contrairement à l'estimation par le principe du maximum de vraisemblance, cette méthode d'estimation ne nécessite pas que l'on pose l'hypothèse de normalité des résidus mais les résultats obtenus sont les mêmes car le problème de minimisation de départ est identique dans les deux cas.

**Limites de la régression linéaire.** Pour des problèmes simples où l'on ne possède que peu de variables explicatives, la régression linéaire multiple est bien adaptée. Cependant, nous proposons ici de modéliser l'expression des gènes en fonction des compositions nucléotidiques et motifs dans différentes régions. Le nombre de variables peut ainsi vite être très important. Le but étant de capturer une combinaison de variables importantes permettant d'expliquer au mieux l'expression des gènes observée pour chacun des échantillons considérés, nous entrons dans un problème de sélection de variables en grande dimension.

## Modèle Lasso

Le LASSO (Least Absolute Shrinkage and Selection Operator) est une méthode de régression linéaire avec sélection de variables qui consiste à estimer les paramètres

$\beta_i$  pour  $i = 1 \dots p$  et d'éliminer les variables  $X_i$  non pertinentes en ramenant leurs coefficients à 0 [260]. On cherche ainsi à résoudre le problème d'optimisation suivant :

$$\hat{\beta} = \operatorname{argmin}_{\beta} \left( \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p \beta_j * x_{ij})^2 + \lambda \sum_{j=1}^p |\beta_j| \right)$$

avec  $\lambda$  un paramètre strictement positif. Par rapport à la régression linéaire classique, une pénalité est ajoutée, aussi appelée terme de régularisation, qui correspond à la norme  $l_1$  de  $\beta$ , soit sa valeur absolue pondérée par un paramètre. La géométrie  $l_1$  est représentée sur la figure 2.1. Cela permet de mettre certains coefficients à 0. Avec  $\lambda = 0$ , cela revient à un estimateur des moindres carrés. A l'inverse, avec un  $\lambda$  très grand, tous les coefficients ou presque sont nuls.

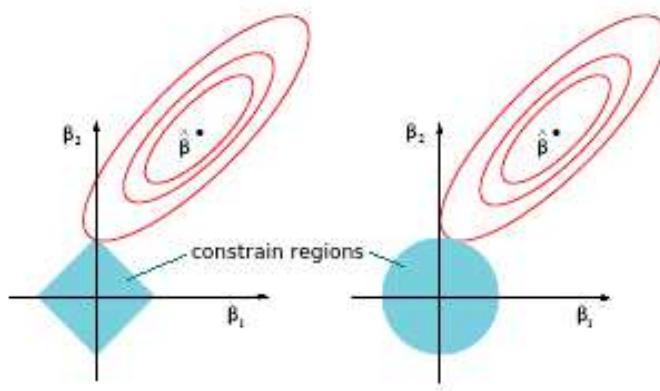


FIGURE 2.1 – Représentation schématique de la comparaison entre la géométrie du lasso (gauche) et du ridge (droite). Les aires bleues représentent les régions de contraintes  $|\beta_1| + |\beta_2| \leq t$  pour le lasso et  $\beta_1^2 + \beta_2^2 \leq t$  pour le ridge et les ellipses rouges correspondent à l'erreur. [Figure adaptée de Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning, 2nd ed. 2009.]

**Validation croisée et évaluation du modèle.** Les données d'apprentissage sont importantes pour construire un modèle robuste. Lorsque l'on veut évaluer les performances d'un modèle, on regarde ses prédictions en comparaison à des valeurs réelles. Cependant, il est important de ne pas apprendre et évaluer un modèle sur un seul et même ensemble de données car cela engendre des problèmes de surapprentissage, c'est à dire que le modèle résultant est trop spécifique au jeu de données sur lequel il a été entraîné. Pour éviter cela, la validation croisée peut être adoptée. Cette méthode consiste à découper le jeu de données dont on dispose en  $K$  groupes de même taille tirés aléatoirement, qui vont chacun servir à leur tour d'échantillon test. les  $K - 1$  groupes restant sont dédiés à l'apprentissage du modèle. Une erreur de test peut alors être calculée pour chacun des  $K$  groupes, et l'estimation de l'erreur du modèle par validation croisée correspond à la moyenne

des erreurs. Une valeur de prédiction peut aussi être calculée pour chaque gène ce qui nous permet ensuite de calculer une corrélation de Spearman entre le vecteur des valeurs d'expression prédictes et le vecteur des valeurs d'expression observées. Cette corrélation, en plus de l'erreur quadratique moyenne, nous permet d'évaluer notre modèle.

## Deep learning

Les algorithmes de *deep learning* font partie des outils de *machine learning* les plus utilisés ces dernières années dans le domaine de la bioinformatique. Ils sont adaptés à l'identification de motifs ou règles complexes à partir de larges échantillons de données riches en paramètres. Ils ont notamment fait leurs preuves dans le domaine de l'analyse d'images, de la médecine à la détection d'éléments cellulaires [281]. Ils sont aujourd'hui adaptés à la compréhension du code de régulation de l'expression des gènes et des effets transcriptionnels des variations génomiques, difficiles à détecter avec la complexité de la partie non codante du génome humain. Des outils de *deep learning* ont ainsi été développés pour prédire l'effet des variations de la séquence sur les marques épigénétiques (DeepSEA [291], DanQ [212]). De manière concomitante à mon travail, des modèles de prédiction de l'expression des gènes à partir de la séquence ADN basés sur des réseaux de neurones convolutifs (convolution neural network, CNN) ont été publiés en 2018 [290, 1, 129] et seront présentés dans la partie suivante.

### 2.1.2 Modèles de prédiction de l'expression des gènes

Dans le premier chapitre de ce présent mémoire, nous avons donné un aperçu des différents facteurs pouvant influencer la transcription des gènes codants pour des protéines par l'ARN Pol II. L'ouverture de la chromatine et sa structure sont en partie conditionnées par les marques épigénétiques et influencent la transcription par un contrôle de l'accès aux gènes à transcrire. Si la chromatine est ouverte, les gènes et leurs promoteurs sont accessibles par les TFs autre que les facteurs pionniers, qui peuvent alors reconnaître des motifs spécifiques et activer ou inhiber la transcription. La régulation de l'expression des gènes est un champ de la biologie très étudié et nous allons voir dans cette sous partie quelques études qui ont orienté nos travaux sur la prédiction de l'expression des gènes à partir de la séquence.

## Prédire l'expression des gènes à partir données expérimentales de marques épigénétiques et facteurs de transcription

La quantité de données expérimentales générées sur les interactions entre protéines et ADN ou ARN étant en pleine croissance depuis le début des années 2000, de nombreux modèles ont été développés pour établir le lien entre l'expression des gènes et les marques épigénétiques ou facteurs de transcription. En 2009, Ouyang et al. proposent un modèle de régression linéaire pour étudier la relation entre l'activité d'une 12<sup>aine</sup> de facteurs de transcription quantifiée par ChIP-seq *in vivo* et l'expression des gènes dans des cellules embryonnaires de souris [193]. Dans la même optique, Cheng et Gerstein [30] et Park et Nakai en 2011-2012 [197] intègrent dans leurs modèles de régression linéaire, à la fois des facteurs de transcription et des modifications d'histones ou autres marques épigénétiques comme la méthylation de l'ADN et les îlots CpG, toujours chez la souris. Cheng et Gerstein montrent que la fixation des TFs possède de meilleures performances prédictives quand on se restreint à une petite région autour du TSS des gènes alors que les modifications d'histones agissent plutôt sur des régions plus larges autour des gènes. Leur modèle suggère également une redondance entre la fixation des TFs et les modifications d'histones en tant que variables prédictives.

En 2010, Karlić et al. développent un modèle prédictif de l'expression des gènes chez l'homme dans des cellules CD4 T+, établi à partir de modifications d'histones seulement [126]. Un peu plus tard, McLeay et al. construisent un modèle de régression intégrant cette fois ci, non pas la fixation des TFs *in vivo* mais leur prédiction *in silico* dans 2 tissus différents, chez l'homme et la souris [174]. Les performances sont équivalentes à celles obtenues avec les données *in vivo*. Ils testent également la contribution de données sur l'accessibilité de la chromatine et les modifications d'histones et voient les performances de leur modèle augmenter. La même année, Dong et al. généralisent les approches des précédentes études à 7 types cellulaires [62]. Pour chaque type cellulaire les résultats sont comparés en fonction du type d'ARN (polyA+ ou polyA-), du compartiment cellulaire (noyau, cytoplasme...) et de la technique utilisée pour mesurer la quantité d'ARN (CAGE, RNA-seq, RNA-PET). Les résultats confirment la relation entre les marques épigénétiques et l'expression des gènes dans différentes conditions mais suggèrent une influence de la technique utilisée avec des performances un peu plus élevées sur des données de CAGE et une différence dans les mécanismes de régulation selon le compartiment cellulaire et la polyadénylation de l'ARN.

Récemment, deux modèles linéaires ont été développés pour prédire l'expression des gènes à partir de données expérimentales. Le premier, RACER [158], est un modèle de régression Lasso pour prédire l'expression des gènes dans des conditions de leucémie myéloïde en intégrant des données de ChIP-seq de facteurs de transcription ainsi que des données de méthylation issues de la lignée cellulaire K562. Le second modèle, TEPIC [231], intègre des variables basées sur la séquences qui sont les scores de motifs (PWM), couplés à des données expérimentales d'ouverture de la chromatine. Nous montrons dans le présent article que ces modèles donnent de bonnes performances pour la prédiction de l'expression des gènes mais ils sont fortement influencés par l'ouverture de la chromatine.

### Prédire les marques épigénétiques à partir de la séquence

Les modèles que nous venons de voir se concentrent seulement sur la prédiction de l'expression des gènes à partir des TFs et des marques épigénétiques. Mais qu'en est-il de ces facteurs eux-mêmes ? Peut-on prédire les marques épigénétiques à partir d'un niveau supérieur de régulation qui serait basé sur la séquence ADN ? Dans les années 2000, de nombreuses études ont identifié les différences épigénétiques existant entre les différents types cellulaires ou encore entre cellules de tissus normaux et tumoraux [9, 98, 240]. Il a été également montré que les marques épigénétiques ne sont pas distribuées aléatoirement à travers le génome [14] et la spécificité de ces marques dépend en grande partie de la séquence ADN. En 2013, 3 études identifient chez l'homme des variants influençant la fixation des TFs, les modifications d'histones ou l'occupation par l'ARN Pol II [175, 132, 127]. Ces variants génomiques peuvent aussi affecter des sites de régulation distaux comme les enhancers. Des modèles ont été développés dans le but de prédire les profils de marques épigénétiques dans différentes conditions, à partir de la séquence ADN.

En 2006, Segal et al. [236] ont modélisé les motifs de fixation des nucléosomes par une matrice contenant la probabilité d'observer les différents dinucléotides à des positions spécifiques. En 2007, Lee et al. utilisent un modèle de régression Lasso incluant des scores de motifs de TFs et des données de structure de l'ADN pour prédire le positionnement des nucléosomes chez la levure [151]. D'autres études se sont intéressées aux modifications d'histones. Même si les enzymes impliquées dans ces modifications n'interagissent pas directement avec la séquence ADN, elles peuvent être recrutées à des loci précis par l'intermédiaire de TFs ou d'ARNs non codants. Les premières études se concentrent sur le recrutement des PcG qui inhibent l'expression des gènes via la marque H3K27me3 [79], et il a été montré que certains

motifs et paires de motifs permettent de discriminer les régions avec et sans PcG chez la drosophile [96]. En 2009, Yuan et al. [284] publient un modèle de prédiction des modifications d'histones à partir de la séquence chez l'homme avec des performances variables selon les marques prédictives. Pour certaines marques comme H3K4me3 et H3K4me1, le modèle performe plutôt bien et l'enrichissement local en CpG n'est pas le seul dinucléotide à expliquer les variations observées. Plus récemment en 2015, Whitaker et al. développent un modèle d'analyse (Epigram) pour prédire les modifications d'histones et la méthylation de l'ADN à partir de motifs, par type cellulaire [276]. Les éléments *cis*-régulateurs identifiés interagissent avec les facteurs qui établissent et maintiennent les modifications épigénétiques. Le modèle développé a permis d'identifier des motifs spécifiques à chaque marque et à chaque cellule. En 2016, Quante et Bird [213] discutent l'effet de petits motifs fréquents et répétés de 2 à 5 pb sur l'épigénome. Ils suggèrent l'idée d'un modèle où ces petits motifs permettent de recruter des protéines pour modifier la structure de la chromatine et permettre l'amplification ou inhibition de l'expression de groupes de gènes, dépendamment du type cellulaire.

En 2015, Zhou et al. développent DeepSEA [291] pour prédire l'effet des variants non codants à partir de la séquence ADN. Le modèle est entraîné sur un jeu de données regroupant des marques épigénétiques dans un grand nombre de types cellulaires. DanQ est également développé pour prédire les fonctions du génome non-codant *de novo* à partir de la séquence, seule l'architecture du modèle change par rapport à DeepSEA [212]. En 2017, Angermueller et al. développent DeepCpG pour comprendre l'influence de la composition de la séquence ADN, et notamment de petits motifs et leurs mutations, sur la variabilité de la méthylation chez l'homme et la souris [4], à partir de données de séquençage de cellules uniques (*single-cell*).

### Utiliser la séquence ADN pour prédire l'expression des gènes

Comme nous venons de le voir, de nombreux papiers ont été publiés, d'un côté pour prédire l'expression des gènes à partir de TFs et marques épigénétiques et de l'autre pour relier les motifs et compositions de séquence aux marques épigénétiques, ce qui suggère la possibilité de prédire l'expression des gènes à partir de la séquence. Avant l'année 2018, peu d'études se sont intéressées directement au lien entre la séquence et l'expression des gènes ; il en existe cependant quelques unes. En 2005, Raghava et al. développent un modèle pour prédire le niveau d'expression des gènes à partir de la composition en acides aminés et dipeptides de la protéine, chez la levure *Saccharomyces cerevisiae* [216]. Leur modèle est non paramétrique et se base

sur un SVM (*Support vector machine*). Il permet d'obtenir une corrélation de plus de 70% entre le niveau d'expression et la composition protéique, ce qui souligne déjà l'importance de la séquence en tant que régulateur de l'expression. En 2004, un modèle est également publié chez *Saccharomyces cerevisiae* pour identifier les motifs régulateurs et établir les éventuels liens et contraintes de position entre ces éléments [11]. Les séquences utilisées sont d'une longueur de 800 pb et situées en amont du début des gènes. Ce modèle est aussi appliqué à l'organisme *Caenorhabditis elegans* et permet d'identifier des règles de combinaisons d'éléments régulateurs contrôlant l'expression de différentes classes de gènes : facteurs de transcription, histones et gènes spécifiques aux lignées germinales. Cependant, ce papier a été réexaminé en 2007 et montre que le modèle surestime les prédictions faites sur l'expression des gènes et que les règles de régulation établies ne sont pas toujours pertinentes [285]. Toutefois, la capacité des motifs à expliquer une partie de l'expression des gènes n'est pas remise en question.

Nos travaux ont été motivés par ces résultats, afin d'établir directement un lien entre composition nucléotidique de la séquence d'ADN brute et expression des gènes, à partir d'un modèle statistique paramétrique. D'autres modèles s'appuyant sur le *deep learning* et concomitants au développement de notre modèle ont été publiés en 2018. Les 3 modèles utilisent des méthodes de CNN pour prédire l'expression des gènes dans différents types cellulaires, à partir de la séquence ADN. Ainsi, ExPecto [290] est un modèle construit à partir de la séquence de régions larges autour du TSS (40 kb) qui prédit, en plus de l'expression des gènes, l'effet des mutations sur l'expression. Un modèle similaire, Xpresso développé par Agarwal et al. [1], permet également de prédire l'expression des gènes à partir de séquences promotrices relativement large (-7 kb/+3,5 kb) autour des TSSs mais qui intègre aussi des variables relatives à la demi-vie des ARNm comme la longueur des introns, la composition en C/G ou la densité exonique. Enfin, Kelley et al. ont développé Basenji [129], modèle basé également sur la séquence ADN uniquement sur des portions du génome très étendues (131 kb) pour capturer de l'information des séquences régulatrices éloignées. Ce modèle est utilisé pour prédire diverses marques épigénétiques et profils transcriptionnels à partir de données de CAGE.

### 2.1.3 Résultats et contribution

L'approche proposée nous permet d'expliquer l'expression des gènes pour chaque patient, seulement à partir des informations liées à la séquence ADN. Nous allons voir dans cette présente partie ma contribution au projet et les résultats principaux

que j'ai obtenus.

**Modèle prédictif : promoteur.** Dans un premier temps nous avons évalué l'impact du promoteur seul sur l'expression des gènes. Plusieurs définitions sont possibles pour cette région régulatrice et nous montrons que la segmentation du promoteur en promoteurs distaux et promoteur proximal améliore les performances prédictives. Nous comparons également les promoteurs autours des différents TSS annotés et celui donnant les meilleures performances est défini autour du second TSS [30]. Nous avons également évalué l'apport des scores de motifs par rapport aux compositions nucléotidiques ainsi que la contribution des *DNAshapes*. Les scores de motifs reflètent la possibilité de liaison des différents TFs sur le promoteur et les *DNAshapes* traduisent les conformations locales de l'ADN en donnant un score pour 4 conformations différentes et à chaque 5-mer (séquence de 5 bases) [292]. Ces deux types de variables permettent d'améliorer légèrement le modèle prédictif mais sont redondantes avec les compositions nucléotidiques qui, seules, permettent déjà d'expliquer une partie importante de l'expression des gènes.

#### Comparaison avec des modèles basés sur des données expérimentales.

Nous avons comparé notre modèle basé sur les scores de motifs et compositions nucléotidiques à celui développé par Li et al. (RACER [158]) où ils intègrent des ChIP-seq de TFs en tant que variables prédictives. Nous montrons que les performances sont équivalentes mais que les modèles basés sur les ChIP-seq de TFs sont fortement influencés par l'ouverture de la chromatine. En effet, un modèle construit à partir de ces variables expérimentales mais permutées aléatoirement par gène présente les mêmes performances que le modèle de base. Avec un modèle comme le nôtre intégrant seulement des informations basées sur la séquence (compositions nucléotidiques et scores de motifs), nous nous affranchissons du biais d'ouverture de la chromatine et les performances sur un modèle construit à partir des variables permutées par gène fait chuter les corrélations. Ce biais d'ouverture de la chromatine est également observé sur un autre modèle prenant comme variables prédictives des scores de motifs couplés à des données expérimentales d'ouverture de la chromatine (TEPIC [231]).

**Contribution des autres régions régulatrices.** La région promotrice est la plus connue et la plus exploitée en termes de régulations transcriptionnelles. Cependant, ce n'est pas la seule région impliquée dans la régulation de l'expression des gènes et nous avons intégré à notre modèle d'autres régions régulatrices décrites

dans l'article. Nous avons évalué l'impact de chacune d'elles dans le modèle prédictif et mis en évidence l'importance du corps du gène dans la régulation de l'expression des gènes, et notamment des introns.

### 2.1.4 Conclusion

Le modèle développé confirme l'importance de la séquence ADN dans la régulation de l'expression des gènes et montre le haut pouvoir prédictif de variables simples comme les fréquences en dinucléotides. Un modèle intégrant une taille de motif supérieure (tri-nucléotides) a également été testé mais ne permet d'améliorer que faiblement les performances. Nous avons pu observer l'importance du corps du gène, et notamment des introns, dans la régulation de l'expression des gènes. Notre modèle capture ainsi des compositions assez larges. Les gènes situés dans les mêmes TADs partagent des compositions nucléotidiques similaires et nous définissons ici une signature pour chaque TAD. Notre modèle a été comparé à d'autres modèles paramétriques (Elastic Net [294]) et non-paramétriques (Forêts aléatoires [294]) (May Taha, doctorante). Les performances sont similaires, le choix s'est donc porté sur le modèle Lasso qui est plus facile à interpréter et adapté à notre jeu de données important, dont les variables prédictives ne sont pas trop fortement corrélées. Pour aller plus loin dans l'interprétation du modèle, une étape de stabilité de sélection a été effectuée (May Taha, doctorante). Ainsi, pour chaque condition et chacune des variables stables associées, nous pouvons déterminer leur effet positif (activateur) ou répressif (inhibiteur).

Notre modèle reste cependant un modèle global, et il ne permet pas de prédire de la même manière tous les gènes. En utilisant les données d'expression de GTEx [36], nous avons pu calculer un coefficient de Gini pour chacun des gènes considérés dans notre modèle. Le coefficient de Gini traduit ici une expression ubiquitaire quand le coefficient est proche de 0 ou à l'inverse tissu-spécifique quand le coefficient est proche de 1. Une corrélation de 30% a été observée entre le coefficient de Gini et les erreurs absolues des gènes. Il existe donc une tendance pour notre modèle à mieux prédire les groupes de gènes ubiquitaires.

## RESEARCH ARTICLE

# Probing instructions for expression regulation in gene nucleotide compositions

Chloé Bessière<sup>1,2\*</sup>, May Taha<sup>1,2,3\*</sup>, Florent Petitprez<sup>1,2</sup>, Jimmy Vandel<sup>1,4</sup>, Jean-Michel Marin<sup>1,3</sup>, Laurent Bréhélin<sup>1,4‡\*</sup>, Sophie Lèbre<sup>1,3,5‡\*</sup>, Charles-Henri Lecellier<sup>1,2‡\*</sup>

**1** IBC, Univ. Montpellier, CNRS, Montpellier, France, **2** Institut de Génétique Moléculaire de Montpellier, University of Montpellier, CNRS, Montpellier, France, **3** IMAG, Univ. Montpellier, CNRS, Montpellier, France, **4** LIRMM, Univ. Montpellier, CNRS, Montpellier, France, **5** Univ. Paul-Valéry-Montpellier 3, Montpellier, France

\* These authors contributed equally to this work.

‡ LB, SL, and CHL also contributed equally to this work.

\* [brehelin@lirmm.fr](mailto:brehelin@lirmm.fr) (LB); [sophie.lebre@umontpellier.fr](mailto:sophie.lebre@umontpellier.fr) (SL); [charles.lecellier@igmm.cnrs.fr](mailto:charles.lecellier@igmm.cnrs.fr) (CHL)



## OPEN ACCESS

**Citation:** Bessière C, Taha M, Petitprez F, Vandel J, Marin J-M, Bréhélin L, et al. (2018) Probing instructions for expression regulation in gene nucleotide compositions. PLoS Comput Biol 14(1): e1005921. <https://doi.org/10.1371/journal.pcbi.1005921>

**Editor:** Zhaoe Zhang, University of Toronto, CANADA

**Received:** July 11, 2017

**Accepted:** December 10, 2017

**Published:** January 2, 2018

**Copyright:** © 2018 Bessière et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are within the paper, its Supporting Information files, and at <http://www.univ-montp3.fr/miap/~lebre/IBCRegulatoryGenomics>.

**Funding:** The work was supported by funding from CNRS, Plan d'Investissement d'Avenir #ANR-11-BINF-0002 Institut de Biologie Computational (young investigator grant to CHL and post-doctoral fellowship to JV), Labex NUMEV (post-doctoral fellowship to JV), INSERM-ITMO Cancer project "LIONS" BIO2015-04. MT is a recipient of a CBS2-

## Abstract

Gene expression is orchestrated by distinct regulatory regions to ensure a wide variety of cell types and functions. A challenge is to identify which regulatory regions are active, what are their associated features and how they work together in each cell type. Several approaches have tackled this problem by modeling gene expression based on epigenetic marks, with the ultimate goal of identifying driving regions and associated genomic variations that are clinically relevant in particular in precision medicine. However, these models rely on experimental data, which are limited to specific samples (even often to cell lines) and cannot be generated for all regulators and all patients. In addition, we show here that, although these approaches are accurate in predicting gene expression, inference of TF combinations from this type of models is not straightforward. Furthermore these methods are not designed to capture regulation instructions present at the sequence level, before the binding of regulators or the opening of the chromatin. Here, we probe sequence-level instructions for gene expression and develop a method to explain mRNA levels based solely on nucleotide features. Our method positions nucleotide composition as a critical component of gene expression. Moreover, our approach, able to rank regulatory regions according to their contribution, unveils a strong influence of the gene body sequence, in particular introns. We further provide evidence that the contribution of nucleotide content can be linked to co-regulations associated with genome 3D architecture and to associations of genes within topologically associated domains.

## Author summary

Identifying a maximum of DNA determinants implicated in gene regulation will accelerate genetic analyses and precision medicine approaches by identifying key gene features. In that context decoding the sequence-level instructions for gene regulation is of prime importance. Among global efforts to achieve this objective, we propose a novel approach

I2S joint doctoral fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

able to explain gene expression in each patient sample using only DNA features. Our approach, which is as accurate as methods based on epigenetics data, reveals a strong influence of the nucleotide content of gene body sequences, in particular introns. In contrast to canonical regulations mediated by specific DNA motifs, our model unveils a contribution of global nucleotide content notably in co-regulations associated with genome 3D architecture and to associations of genes within topologically associated domains. Overall our study confirms and takes advantage of the existence of sequence-level instructions for gene expression, which lie in genomic regions largely underestimated in regulatory genomics but which appear to be linked to chromatin architecture.

## Introduction

The diversity of cell types and cellular functions is defined by specific patterns of gene expression. The regulation of gene expression involves a plethora of DNA/RNA-binding proteins that bind specific motifs present in various DNA/RNA regulatory regions. At the DNA level, transcription factors (TFs) typically bind 6–8bp-long motifs present in promoter regions, which are close to transcription start site (TSS). TFs can also bind enhancer regions, which are distal to TSSs and often interspersed along considerable physical distance through the genome [1]. The current view is that DNA looping mediated by specific proteins and RNAs places enhancers in close proximity with target gene promoters (for review [2–5]). High-resolution chromatin conformation capture (Hi-C) technology identified contiguous genomic regions with high contact frequencies, referred to as topologically associated domains (TADs) [6]. Within a TAD, enhancers can work with many promoters and, on the other hand, promoters can contact more than one enhancer [5, 7]. Several large-scale data derived from high-throughput experiments (such as ChIP-seq [8], SELEX-seq [9], RNAcompete [10]) can be used to highlight TF/RBP binding preferences and build Position Weight Matrixes (PWMs) [11]. The human genome is thought to encode ~2,000 TFs [12] and >1,500 RBPs [13]. It follows that gene regulation is achieved primarily by allowing the proper combination to occur i.e. enabling cell- and/or function-specific regulators (TFs or RBPs) to bind the proper sequences in the appropriate regulatory regions. In that context, epigenetics clearly plays a central role as it influences the binding of the regulators and ultimately gene expression [14]. Provided the variety of regulatory mechanisms, deciphering their combination requires mathematical/computational methods able to consider all possible combinations [15]. Several methods have recently been proposed to tackle this problem [16–19]. Although these models appear very efficient in predicting gene expression and identifying key regulators, they mostly rely on experimental data (ChIP-seq, methylation, DNase hypersensitivity), which are limited to specific samples (often to cell lines) and which cannot be generated for all TFs/RBPs and all cell types. These technological features impede from using this type of approaches in a clinical context in particular in precision medicine. In addition, we show here that, although these approaches are accurate, their biological interpretation can be misleading. Finally these methods are not designed to capture regulation instructions that may lie at the sequence-level before the binding of regulators or the opening of the chromatin. There is indeed a growing body of evidence suggesting that the DNA sequence *per se* contains information able to shape the epigenome and explain gene expression [20–25]. Several studies have shown that sequence variations affect histone modifications [21–23]. Specific DNA motifs can be associated with specific epigenetic marks and the presence of these motifs can predict the epigenome in a given cell type [24]. Quante and Bird proposed that proteins able to “read” domains of

relatively uniform DNA base composition may modulate the epigenome and ultimately gene expression [20]. In that view, modeling gene expression using only DNA sequences and a set of predefined DNA/RNA features (without considering experimental data others than expression data) would be feasible. In line with this proposal, Raghava and Han developed a Support Vector Machine (SVM)-based method to predict gene expression from amino acid and dipeptide composition in *Saccharomyces cerevisiae* [26].

Here, we built a global regression model per sample to explain the expression of the different genes using their nucleotide compositions as predictive variables. The idea beyond our approach is that the selected variables (defining the model) are specific to each sample. Hence the expression of a given gene may be predicted by different variables in different samples. This approach was tested on several independent datasets: 2,053 samples from The Cancer Genome Atlas (1,512 RNA-sequencing data and 582 microarrays) and 3 ENCODE cell lines (RNA sequencing). When restricted to DNA features of promoter regions our model showed accuracy similar to that of two independent methods based on experimental data [17, 19]. We confirmed the importance of nucleotide composition in predicting gene expression. Moreover the performance of our approach increases by combining the contribution of different types of regulatory regions. We thus showed that the gene body (introns, CDS and UTRs), as opposed to sequences located upstream (promoter) or downstream, had the most significant contribution in our model. We further provided evidence that the contribution of nucleotide composition in predicting gene expression is linked to co-regulations associated with genome architecture and TADs.

## Materials and methods

### Datasets, sequences and online resources

RNA-seq V2 level 3 processed data were downloaded from the TCGA Data Portal. Our training data set contained 241 samples randomly chosen from 12 different cancers (20 cancerous samples for each cancer except 21 for LAML). Our model was further evaluated on an additional set of 1,270 tumors from 14 cancer types. We also tested our model on 582 TCGA microarray data. The TCGA barcodes of the samples used in our study have been made available at <http://www.univ-montp3.fr/miap/~lebre/IBCRegulatoryGenomics>.

Isoform expression data (.rsem.isoforms.normalized\_results files) were downloaded from the Broad TCGA GDAC (<http://gdac.broadinstitute.org>) using firehose\_get. We collected data for 73599 isoforms in 225 samples of the 241 initially considered. All the genes and isoforms not detected (no read) in any of the considered samples were removed from the analyses. Expression data were log transformed.

All sequences were mapped to the hg38 human genome and the UCSC liftover tool was used when necessary. Gene TSS positions were extracted from GENCODEv24. UTR and CDS coordinates were extracted from ENSEMBL Biomart. To assign only one 5UTR sequence to one gene, we merged all annotated 5UTRs associated with the gene of interest using Bedtools merge [27] and further concatenated all sequences. The same procedure was used for 3UTRs and CDSs. Intron sequences are GENCODEv24 genes to which 5UTR, 3UTR and CDS sequences described above were subtracted using Bedtools subtract [27]. These sequences therefore corresponded to constitutive introns. The intron sequences were concatenated per gene. The downstream flanking region (DFR) was defined as the region spanning 1kb after GENCODE v24 gene end. Fasta files were generated using UCSC Table Browser or Bedtools getfasta [27].

TCGA isoform TSSs were retrieved from [https://webshare.bioinf.unc.edu/public/mRNaseq\\_TCGA/unc\\_hg19.bed](https://webshare.bioinf.unc.edu/public/mRNaseq_TCGA/unc_hg19.bed) and converted into hg38 coordinates with UCSC liftover.

For other regulatory regions associated to transcript isoforms (UTRs, CDS, introns and DFR), we used GENCODE v24 annotations.

### Nucleotide composition

The nucleotide ( $n = 4$ ) and dinucleotide ( $n = 16$ ) percentages were computed from the different regulatory sequences where:

$$\text{percentage}(N, s) = \frac{\sharp N}{l}$$

is the percentage of nucleotide  $N$  in the regulatory sequence  $s$ , with  $N$  in  $\{A, C, G, T\}$  and  $l$  the length of sequence  $s$ , and

$$\text{percentage}(NpM, s) = \frac{\sharp NpM}{l - 1}$$

is the  $NpM$  dinucleotide percentage in the regulatory sequence  $s$ , with  $N$  and  $M$  in  $\{A, C, G, T\}$  and  $l$  the length of sequence  $s$ .

### Motif scores

Motif scores in core promoters were computed using the method explained in [11] and Position Weight Matrix (PWM) available in JASPAR CORE 2016 database [28]. Let  $w$  be a motif and  $s$  a nucleic acid sequence. For all nucleotide  $N$  in  $\{A, C, G, T\}$ , we denoted by  $P(N|w_j)$  the probability of nucleotide  $N$  in position  $j$  of motif  $w$  obtained from the PWM, and by  $P(N)$  the prior probability of nucleotide  $N$  in all sequences.

The score of motif  $w$  at position  $i$  of sequence  $s$  is computed as follows:

$$\text{score}(w, s, i) = \sum_{j=0}^{|w|-1} \log \frac{P(s_{i+j}|w_j)}{P(s_{i+j})}$$

with  $|w|$  the length of motif  $w$ ,  $s_{i+j}$  the nucleotide at position  $i + j$  in sequence  $s$ . The score of motif  $w$  for sequence  $s$  is computed as the maximal score that can be achieved at any position of  $s$ , i.e.:

$$\text{score}(w, s) = \max_{i=0}^{l-|w|} \text{score}(w, s, i),$$

with  $l$  the length of sequence  $s$ .

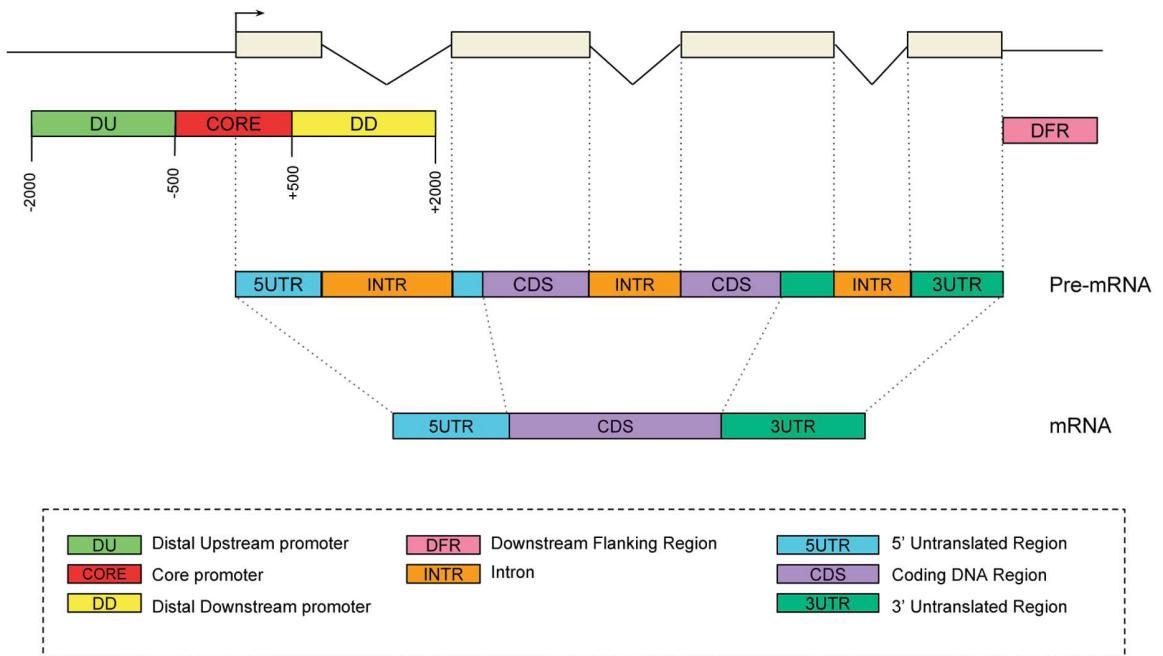
Models were also built on sum scores as:

$$\text{scoreSum}(w, s) = \sum_{i=0}^{l-|w|} \text{score}(w, s, i),$$

and further compared to models built on mean scores (S1 Fig). Taking mean or sum scores per region yielded similar results (Wilcoxon test p-value = 0.68).

### DNAshape scores

DNA shape scores were computed using DNAshapeR [29]. Briefly, provided nucleotide sequences, DNAshapeR uses a sliding pentamer window to derive the structural features corresponding to minor groove width (MGW), helix twist (HelT), propeller twist (ProT) and Roll from all-atom Monte Carlo simulations [29]. Thus, for each DNA shape, a score is given to



**Fig 1. Genomic regions considered for gene expression prediction.** An illustrative transcript is shown as example.

<https://doi.org/10.1371/journal.pcbi.1005921.g001>

each base of each sequence considered (DU, CORE and DD—see Fig 1). We then computed the mean of these scores for each sequence providing 12 additional variables per gene.

### Enhancers

The coordinates of the enhancers mapped by FANTOM on the hg19 assembly [7] were converted into hg38 using UCSC liftover and further intersected with the different regulatory regions. We computed the density of enhancers per regulatory region ( $R$ ) by dividing the sum, for all genes, of the intersection length of enhancers with gene  $i$  ( $L_{enh_i}$ ) by the sum of the lengths of this regulatory region for all genes:

$$enhDensity_{(R)} = \frac{\sum_i (L_{enh_i} \text{ in } R_i)}{\sum_i length(R_i)}$$

### Copy Number Variation (CNV)

Processed data were downloaded from the firehose Broad GDAC (<https://gdac.broadinstitute.org/>). We used the genome-wide SNP array data and the segment mean scores. In order to assign a CNV score to each gene, the coordinates (hg19) of the probes were intersected with that of GENCODE v19 genes using Bedtools intersect [27] and an overlap of 85% of the gene total length. The corresponding segment mean value was then assigned to the intersecting genes. In case no intersection was detected, the gene was assigned a score of 0. We next computed Spearman correlations between genes absolute error (lasso model) and genes absolute segment mean score for each of the 241 samples of the training set.

## Expression quantitative trait loci and single nucleotide polymorphisms

The v6p GTex *cis*-eQTLs were downloaded from the GTEx Portal (<http://www.gtexportal.org/home/>). The hg19 *cis*-eQTL coordinates were converted into hg38 using UCSC liftover and further intersected with the different regulatory regions. We restricted our analyses to *cis*-eQTLs impacting their own host gene. We computed the density of *cis*-eQTL per regulatory region ( $R$ ) by dividing the sum, for all genes, of the number of *cis*-eQTLs of gene  $i$  ( $eQTLs_i$ ) located in the considered region for gene  $i$  ( $R_i$ ) by the sum of the lengths of this regulatory region for all genes:

$$eQTLdensity_{(R)} = \frac{\sum_i \#(eQTLs_i \text{ in } R_i)}{\sum_i length(R_i)}$$

Likewise we computed the density of SNPs in core promoters and introns by intersecting coordinates of these two regions (liftovered to hg19) with that of SNPs detected on chromosomes 1, 2 and 19 ([ftp://ftp.ncbi.nih.gov/snp/organisms/human\\_9606\\_b150\\_GRCh37p13/BED/](ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b150_GRCh37p13/BED/)):

$$SNPdensity_{(R)} = \frac{\sum_i \#(SNP_i \text{ in } R_i)}{\sum_i length(R_i)}$$

## Methylation

Illumina Infinium Human DNA Methylation 450 level 3 data were downloaded from the Broad TCGA GDAC (<http://gdac.broadinstitute.org>) using firehose\_get. The coordinates of the methylation sites (hg18) were converted into hg38 using the UCSC liftover and further intersected with that of the core promoters (hg38). For each gene, we computed the median of the beta values of the methylation sites present in the core promoter and further calculated the median of these values in 21 LAML and 17 READ samples with both RNA-seq and methylation data. We compared the overall methylation status of the core promoters in LAML and READ using a wilcoxon test.

## Gini coefficient

We used 8,556 GTEx RNA-seq libraries (<https://www.gtexportal.org/home/datasets>) to compute the Gini coefficient for 16,134 genes on the 16,294 considered in our model. Gini coefficient measures statistical dispersion and can be used to measure gene ubiquity: value 0 represents genes expressed in all samples while value 1 represents genes expressed in only one sample. To compute Gini coefficient we used R package `ineq`. We then computed, for the 241 samples, Spearman correlation between Gini coefficients and model gene absolute errors. Similar analyses were performed with 1,897 FANTOM 5 CAGE libraries to compute the Gini coefficients for 15,904 genes.

## Functional enrichment

Gene functional enrichments were computed using the database for annotation, visualization and integrated discovery (DAVID) [30].

## Linear regression with $\ell_1$ -norm penalty (Lasso)

We performed estimation of the linear regression model (1) via the lasso [31]. Given a linear regression with standardized predictors and centered response values, the lasso solves the

$\ell_1$ -penalized regression problem of finding the vector coefficient  $\beta = \{\beta_i\}$  in order to minimize

$$\text{Min} \left( \|y^c(g) - \sum_i \beta_i x_{i,g}^s\|^2 + \lambda \sum_i |\beta_i| \right),$$

where  $y^c(g)$  is the centered gene expression for all gene  $g$ ,  $x_{i,g}^s$  is the standardized DNA feature  $i$  for gene  $g$  and  $\sum_i |\beta_i|$  is the  $\ell_1$ -norm of the vector coefficient  $\beta$ . Parameter  $\lambda$  is the tuning parameter chosen by 10 fold cross validation. The higher the value of  $\lambda$ , the fewer the variables. This is equivalent to minimizing the sum of squares with a constraint of the form  $\sum_i |\beta_i| \leq s$ . Gene expression predictions are computed using coefficient  $\beta$  estimated with the value of  $\lambda$  that minimizes the mean square error. Lasso inference was performed using the function `cv.glmnet` from the R package `glmnet` [32]. The LASSO model was compared to two non parametric approaches: Regression trees (CART) [33] and Random forest [34]. [S1 Table](#) summarizes accuracy and computing time of each approach. Regression trees achieved significantly lower accuracy than the two other approaches (Wilcox test p-values  $< 2e^{-16}$ ), while linear model and random forest yielded similar results (p-value 0.18). Moreover, computing time for linear model was much lower than that of random forest. These results emphasize the merits of linear model such as LASSO in their interpretability and efficiency.

### Variable stability selection

We used the stability selection method developed by Meinshausen *et al.* [35], which is a classical selection method combined with lasso penalization. Consistently selected variables were identified as follows for each sample. First, the lasso inference is repeated 500 times where, for each iteration, (i) only 50% of the genes is used (uniformly sampled) and (ii) a random weight (uniformly sampled in [0.5;1]) is attributed to each predictive variable. Second, a variable is considered as stable if selected in more than 70% of the iterations, using the method proposed in [36] to set the value of lasso penalty  $\lambda$ . One of the advantage of this method is that the variable selection frequency is computed globally for all the variables by attributing a random weight to each variable at each iteration, thus taking into account the dependencies between the variables. This variable stability selection procedure was implemented using functions `stabpath` and `stabsel` from the R package `C060` for `glmnet` models [36].

### Regression trees

Regression trees were implemented with the `rpart` package in R [32]. In order to avoid overfitting, trees were pruned based on a criterion chosen by cross validation to minimize mean square error. The minimum number of genes was set to 100 genes per leaf.

### TAD enrichment

We considered TADs mapped in IMR90 cells [6] containing more than 10 genes (373 out of 2243 TADs with average number of genes = 14). The largest TAD had 76 associated genes. First, for each TAD and for each region considered, the percentage of each nucleotide and dinucleotide associated to the embedded genes were compared to that of all other genes using a Kolmogorov-Smirnov (KS) test. For a given dinucleotide (for example CpG), we applied KS tests to assess whether the CpG frequency distribution in genes in one specific TAD differs from the distribution in genes in other TADs. Correction for multiple tests was applied using the False Discovery Rate (FDR)  $< 0.05$  [37] and the R function `p.adjust` [32]. Second, for each of the 967 groups of genes (identified by the regression trees, with mean error  $<$  mean error of the 1st quartile), the over-representation of each TAD within each group was tested

using the R hypergeometric test function `phyper` [32]. Correction for multiple tests was applied using  $FDR < 0.05$  [37].

## Availability of data and materials

The matrices of predicted variables (log transformed RNA seq data) and predictive variables (nucleotide and dinucleotide percentages, motifs and DNA shape scores computed for all genes as described above) as well as the TCGA barcodes of the 241 samples used in our study have been made available at <http://www.univ-montp3.fr/miap/~lebre/IBCRegulatoryGenomics>.

## Results

### Mathematical approach to model gene expression

We built a global linear regression model to explain the expression of genes using DNA/RNA features associated with their regulatory regions (e.g. nucleotide composition, TF motifs, DNA shapes):

$$y(g) = a + \sum_i b_i x_{i,g} + e(g) \quad (1)$$

where  $y(g)$  is the expression of gene  $g$ ,  $x_{i,g}$  is feature  $i$  for gene  $g$ ,  $e(g)$  is the residual error associated with gene  $g$ ,  $a$  is the intercept and  $b_i$  is the regression coefficient associated with feature  $i$ .

The advantage of this approach is that it allows to unveil, into a single model, the most important regulatory features responsible for the observed gene expression. The relative contribution of each feature can thus be easily assessed. It is important to note that the model is specific to each sample. Hence the expression of a given gene may be predicted by different variables depending on the sample. Our computational approach was based on two steps. First, a linear regression model (1) was trained with a lasso penalty [31] to select sequence features relevant for predicting gene expression. Second, the performances of our model was evaluated by computing the mean square of the residual errors, and the correlation between the predicted and the observed expression for all genes. This was done in a 10 fold cross-validation procedure. Namely, in all experiments hereafter, the set of genes was randomly split in ten parts. Each part was alternatively used for the test (i.e. for comparing observed and predicted values) while the remaining genes were used to train the model. This ensures that the model used to predict the expression of a gene has not been trained with any information relative to this gene. Our approach was applied to a set of RNA sequencing data from TCGA. We randomly selected 241 gene expression data from 12 cancer types (see <http://www.univ-montp3.fr/miap/~lebre/IBCRegulatoryGenomics> for the barcode list). For each dataset (i.e sample), a regression model was learned and evaluated. See [Materials and methods](#) for a complete description of the data, the construction of the predictor variables and the inference procedure. We further evaluated our model on 3 independent ENCODE RNA-seq, 1,270 TCGA RNA-seq and 582 microarrays datasets (see below).

### Contribution of the promoter nucleotide composition

We first evaluated the contribution of promoters, which are one of the most important regulatory sequences implicated in gene regulation [38]. We extracted DNA sequences encompassing  $\pm 2000$  bases around all GENCODE v24 TSSs and looked at the percentage of dinucleotides along the sequences ([S2 Fig](#)). Based on these distributions, we segmented the promoter into three distinct regions: -2000/-500 (referred here to as distal upstream promoter, DU), -500/+500 (thereafter called core promoter though longer than the core promoter traditionally

considered) and +500/+2000 (distal downstream promoter, DD)([Fig 1](#)). We computed the nucleotide ( $n = 4$ ) and dinucleotide ( $n = 16$ ) relative frequencies in the three distinct regions of each gene. For each sample, we trained one model using the 20 nucleotide/dinucleotide relative frequencies from each promoter segment separately, and from each combination of promoter segments. We observed that the core promoter had the strongest contribution compared to DU and DD ([Fig 2A](#)). Considering promoter as one unique sequence spanning -2000/+2000 around TSS achieved lower model accuracy than combining different promoter segments ([Fig 2A](#)). The highest accuracy was obtained combining all three promoter segments ([Fig 2A](#)).

Promoters are often centered around the 5' most upstream TSS (i.e. gene start). However genes can have multiple transcriptional start sites. The median number of alternative TSSs for the 19,393 genes listed in the TCGA RNA-seq V2 data is 5 and only 2,753 genes harbor a single TSS ([S3 Fig](#)). We therefore evaluated the performance of our model comparing different promoters centered around the first, second, third and last TSS ([Fig 2B](#)). In the absence of second TSS, we used the first TSS and likewise the second TSS in the absence of a third TSS. The last TSS represents the most downstream TSS in all cases. We found that our model achieved higher predictive accuracy with the promoters centered around the second TSS ([Fig 2B](#)), in agreement with [16]. As postulated by Cheng *et al.* [16] in the case of TFs, the nucleotide composition around the first TSS may be linked to the recruitment of chromatin remodelers and thereby prime the second TSS for gene expression. Dedicated experiments would be required to assess this point.

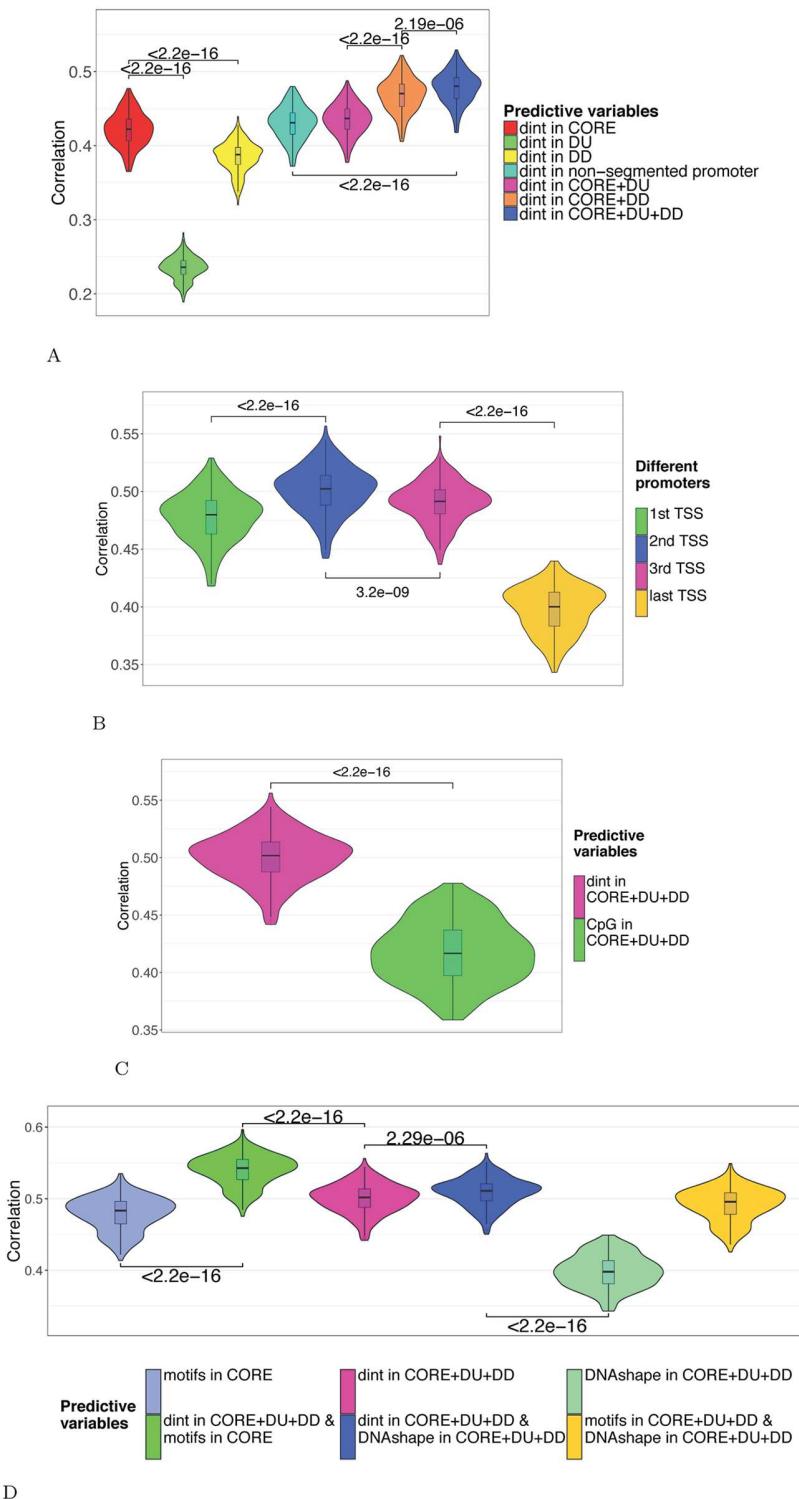
We noticed that incorporating the number of TSSs associated with each gene drastically increased the performance of our model ([S4 Fig](#)). Multiplying TSSs may represent a genuine mechanism to control gene expression level. On the other hand this effect may merely be due to the fact that the more a gene is expressed, the more its different isoforms will be detected (and hence more TSSs will be annotated). Because the number of known TSSs results from annotations deduced from experiments, we decided not to include this variable into our final model.

### Contribution of specific features associated with promoters

Provided the importance of CpGs in promoter activity [38], we first compared our model with a model built only on promoter CpG content. We confirmed that CpG content had an important contribution in predicting gene expression (median  $R = 0.417$ , [Fig 2C](#)). However considering other dinucleotides achieved better model performances, indicating that dinucleotides other than CpG contribute to gene regulation. This is in agreement with results obtained by Nguyen *et al.*, who showed that CpG content is insufficient to encode promoter activity and that other features might be involved [[39](#)].

We integrated TF motifs considering Position Weight Matrix scores computed in the core promoter and observed a slight but significant increase of the regression performance (median  $r = 0.543$  with motif scores vs.  $r = 0.502$  without motif scores, [Fig 2D](#)). As DNA sequence is intrinsically linked to three-dimensional local structure of the DNA (DNA shape), we also computed, for each promoter segment (DU, CORE and DD), the mean scores of the four DNA shape features provided by DNashapeR [29] (helix twist, minor groove width, propeller twist, and Roll), adding 12 variables to the model. Although the difference between models with and without DNA shapes is also significant, the increase in performance is more modest than when including TF motif scores ([Fig 2D](#)).

Our model suggested that nucleotide composition had a greater contribution in predicting gene expression compared to TF motifs and DNA shapes. This is in agreement with the



**Fig 2. A: Contribution of the promoter segments.** The model was built using 20 variables corresponding to the nucleotide (4) and dinucleotide (16) percentages computed in the CORE promoter (red), DU (green) or DD (yellow). These variables were then added in different combinations: CORE+DU (pink, 40 variables); CORE+DD (orange, 40 variables); CORE+DU+DD (light blue, 60 variables). Promoter segments were centered around the first most upstream TSS. For sake of comparison, the model was also built on 20 variables corresponding to the nucleotide and

dinucleotide compositions of the non segmented promoters (-2000/+2000 around the first most upstream TSS)(light blue). All different models were fitted on 19,393 genes for each of the 241 samples considered. The prediction accuracy was evaluated in each sample by evaluating the Spearman correlation coefficients between observed and predicted gene expressions in a cross-validation procedure. The correlations obtained in all samples are shown as violin plots. **B: Prediction accuracy comparing alternative TSSs.** The model was built using the 60 nucleotide/dinucleotide percentages computed in the 3 promoter segments (CORE+DU+DD) centered around 1st, 2nd, 3rd and last TSSs (from left to right). **C: Contribution of CpG.** The model was built using the 60 nucleotide/dinucleotide or only the 3 CpG percentages computed in the 3 promoter segments (CORE+DU+DD) centered around the 2nd TSS. **D: Contribution of motifs and local DNA shapes.** The model was built using (i) 60 nucleotide/dinucleotide percentages computed in the 3 promoter segments (CORE+DU+DD) (“dint”, pink),(ii) 471 JASPAR2016 PWM scores computed in the CORE segment (“motifs”, light blue) and (iii) the 12 DNA shapes corresponding to the 4 known DNAs shapes computed in CORE, DU and DD (“DNAshape”, green). All sequences were centered around the 2nd TSS. These variables were further added in different combinations to build the models indicated: dint+motifs (531 variables, green), dint+DNAs shapes (32 variables, dark blue), motifs+DNAs shapes (483 variables, light green).

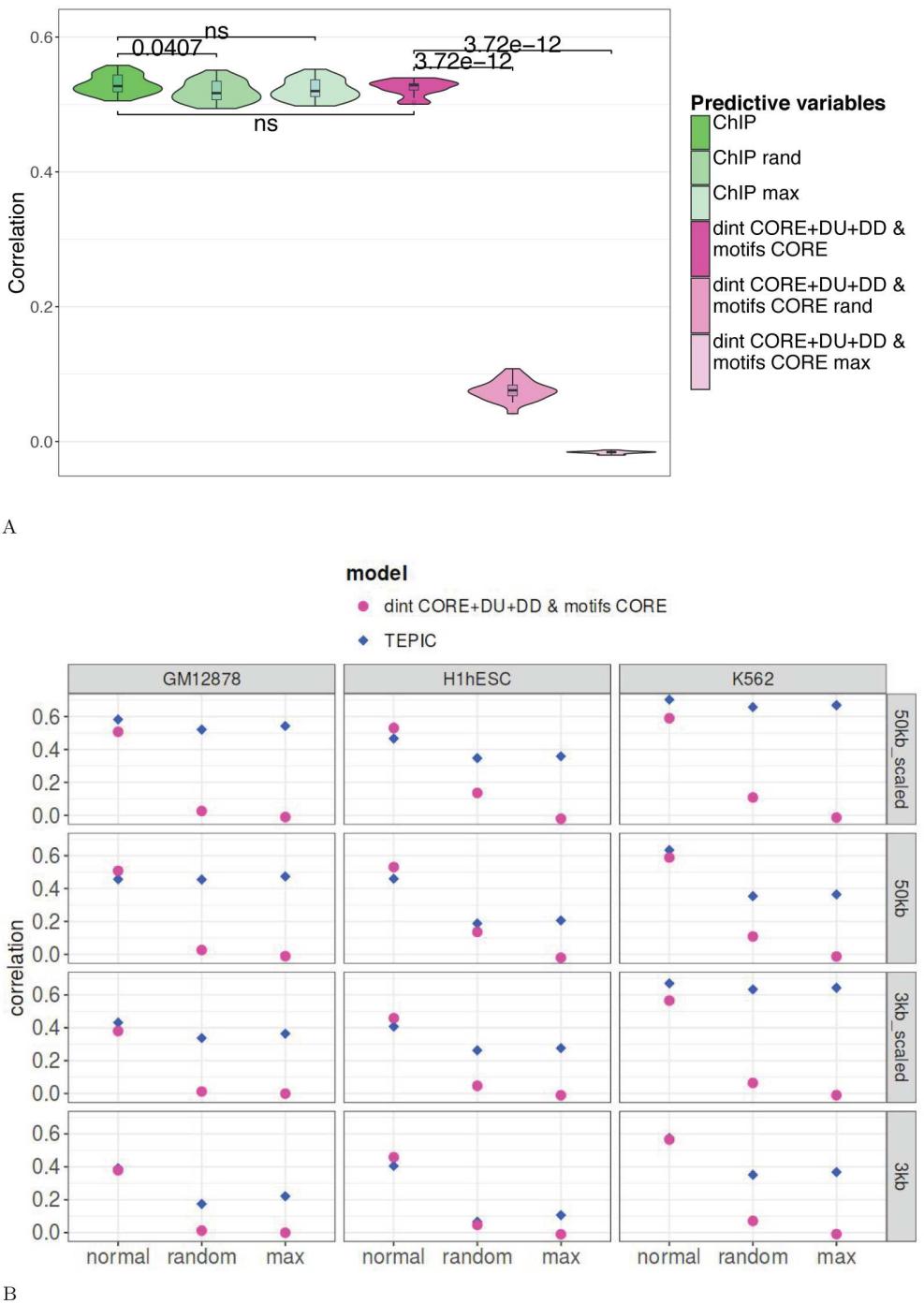
<https://doi.org/10.1371/journal.pcbi.1005921.g002>

findings revealing the influence of the nucleotide environment in TFBS recognition [40]. Note however that nucleotide composition, TF motifs and DNA shapes may be redundant variables. Besides, a linear model may not be optimal to efficiently capture the contributions of TF motifs and/or DNA shapes. The highest performance was achieved by combining nucleotide composition with TF motifs (Fig 2D). In the following analyses, the model was built on both dinucleotide composition and core promoter TF motifs.

### Comparison with models based on experimental data

The wealth of TF ChIP-seq, epigenetic and expression data has allowed the development of methods aimed at predicting gene expression based on differential binding of TFs and epigenetic marks [16–19]. We sought to compare our approach, which does not necessitate such cell-specific experimental data, to these methods. We first compared our results to that of Li *et al.* who used a regression approach called RACER to predict gene expression on the basis of experimental data, in particular TF ChIP-seq data and DNA methylation [17]. Note that, with this model, the contribution of TF regulation in predicting gene expression is higher than that of DNA methylation [17].

We computed the Spearman correlations between expressions observed in the subsets of LAMLs studied in [17] and expressions predicted by our model or by RACER (Fig 3A). For the sake of comparison, we used the RACER model built solely on ChIP-seq data, hereafter referred to as “ChIP-based model”. RACER performance was assessed using the same cross-validation procedure we used for our method. Overall our model was as accurate as ChIP-based model (median correlation  $r = 0.529$  with our model vs. median  $r = 0.527$  with ChIP-based model (Fig 3A)). We then controlled the biological information retrieved by the two approaches by randomly permuting, for each gene, the values of the predictive variables (dinucleotide counts/motif scores in our model and ChIP-seq signals in the ChIP-based model). This creates a situation where the links between the combination of predictive variables and expression is broken, while preserving the score distribution of the variables associated with each gene. For example, genes associated with numerous ChIP-seq peaks will also have numerous ChIP-seq peaks in random data. In such situation, a regression model is expected to poorly perform. Surprisingly, the accuracy of ChIP-based model was not affected by the randomization process (median  $r = 0.517$ , Fig 3A) while that of our model was severely impaired (median  $r = 0.076$ , Fig 3A). We built another control model using a single predictive variable per gene corresponding to the maximum value of all predictive variables initially considered. Here again the ChIP-based model was not affected by this process (median  $r = 0.520$ , Fig 3A) while our model failed to accurately predict gene expression with this type of control variable (median  $r = -0.016$ , Fig 3A).



**Fig 3.** **A:** Comparison with model integrating TF-binding signals. The model was built using 531 variables corresponding to the 60 nucleotide/dinucleotide percentages and the 471 motif scores computed in the 3 promoter segments (CORE, DU, DD) centered around the 2nd TSS (pink). A model built on ChIP-seq data [17] was used for comparison (green). Both models were fitted on the same gene set ( $n = 16,298$ ) for 21 LAML samples and assessed by cross-validation. The correlations obtained with ChIP-based RACER and our model were compared using Wilcoxon test but no significant difference was observed ( $p\text{-value} = 0.425$ ). The two models were also built on randomized values of predictive variables (rand) and on the maximum value of all predictive variables (max). **B:** Comparison with model integrating open-chromatin signals. The linear model was built using the 531 variables (nucleotide/dinucleotide percentages and motif scores in CORE, DU and DD) and the expression data obtained in K562, hESC and GM12878 [19]. TEPIIC was built as described in [19], within a 3 kb or a 50 kb window around TSSs. The scaled version of TEPIIC

incorporates the abundance of open-chromatin peaks in the analyzed sequences. All types of TEPIC models were tested (3kb, 3kb-scaled, 50kb and 50kb-scaled) by cross-validation. In each case, our model was built on the set of genes considered by TEPIC. TEPIC uses 12 conditions making hard to compute Wilcoxon tests. A direct comparison showed that, in “normal” conditions (first column of each panel), our model and TEPIC give overall very similar results (our model being as accurate as TEPIC in 2 conditions and slightly better in 5 out of the 10 remaining conditions). Models were further built on randomized values of predictive variables (rand) and on the maximum value of all predictive variables (max). Overall, absence of effect of the randomization procedure suggests that RACER and TEPIC mainly capture the level of chromatin opening rather than the TF combinations responsible for gene expression.

<https://doi.org/10.1371/journal.pcbi.1005921.g003>

ChIP-seq data are probably the best way to measure the activity of a TF because binding of DNA reflects the output of RNA/protein expression as well as any appropriate post-translational modifications and subcellular localizations. However this type of data also reflects chromatin accessibility (i.e. most TFs bind accessible genomic regions) and TFs tend to form clusters on regulatory regions [41]. The binding of one TF in the promoter region is therefore likely accompanied by the binding of others. Hence, rather than inferring the TF combination responsible for gene expression, linear models based of ChIP-seq data predominantly captures the quantity of TFs (i.e. the opening of the chromatin) in the promoter region of each gene, which explains their good accuracy on randomized or maximized variables.

We indeed observed a similar bias in the results obtained by TEPIC [19], a regression method that predicts gene expression from PWM scores and open-chromatin data. Specifically, TEPIC computes a TF-affinity score for each gene and each PWM by summing up the TF affinities in all open-chromatin peaks (DNaseI-seq) within a close (3,000 bp) or large (50,000 bp) window around TSSs. This scoring takes into account the scores of PWMs in the open-chromatin peaks but is also influenced by the number of open-chromatin peaks in the analyzed sequences and the abundance of open-chromatin peaks (“scaled” version). As a result, genes with many open-chromatin peaks tend to get higher TF-affinity scores than genes with low number of open-chromatin peaks. We trained linear models on three cell-lines using either the four TEPIC affinity-scores or our variables and compared the results (Fig 3B). As for the ChIP-based models, we observed that our model was approximately as accurate as TEPIC score model, validating our approach with an independent dataset. Applying the random permutations on the TEPIC scores did not significantly impact the accuracy of the approach in most cases, especially for the scaled versions (Fig 3B). Hence, as for the ChIP-based model, the TEPIC score model seems to mainly capture the level of chromatin opening rather than the TF combinations responsible for gene expression. Conversely, our model solely built on DNA sequence features is not influenced by the chromatin accessibility and thus can yield relevant combinations of explanatory features (see the randomized control in Fig 3A and 3B). Note that the non-scaled version of TEPIC did show a loss of accuracy for cell-line H1-hESC (as well as a moderate loss for K562, but none for GM12878) when randomizing or maximizing the variables (Fig 3B). This result indicates that, although taking the abundance of open-chromatin peaks in the analyzed sequences does increase expression prediction accuracy, it might generate more irrelevant combinations of explanatory features than non-scaled versions.

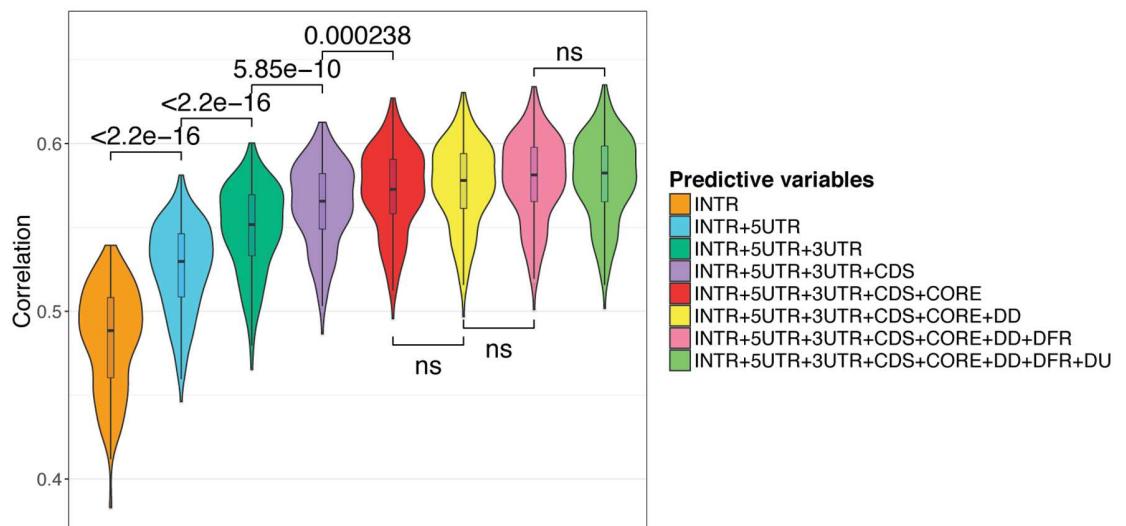
## Contribution of additional genomic regions

Additional genomic regions were integrated into our model. We first thought to consider enhancer sequences implicated in transcriptional regulation. We used the enhancer mapping made by the FANTOM5 project, which identified 38,554 human enhancers across 808 samples [7]. This mapping uses the CAGE technology, which captures the level of activity for both promoters and enhancers in the same samples. It is then possible to predict the potential target genes of the enhancers by correlating the activity levels of these regulatory regions over

hundreds of human samples [7]. However FANTOM5 enhancers are only assigned to 11,359 genes from the TCGA data, which correspond to the most expressed genes across different cancers (S5 Fig). Provided that the detection of enhancers relies on their activity, it is expected that enhancers are better characterized for the most frequently expressed genes. Because considering only the genes with annotated enhancers would considerably reduce the number of genes and including enhancers features only when available would introduce a strong bias in the performance of our model, we decided not to include these regulatory regions.

Second we analyzed the contribution of regions defined at the RNA level, namely 5'UTR, CDS, 3'UTR and introns, which can be responsible for post-transcriptional regulations [13, 17, 26, 42–50] (Fig 1). For all genes, we extracted all annotated 5'UTRs, 3'UTRs and CDSs, which were further merged and concatenated to a single 5'UTR, a single CDS, and a single 3'UTR per gene. Introns were defined as the remaining sequence (Fig 1). We also tested the potential contribution of the 1kb region located downstream the gene end, called thereafter Downstream Flanking Region (DFR, Fig 1). Our rationale was based on reports showing the presence of transient RNA downstream of polyadenylation sites [51], the potential presence of enhancers [7] and the existence of 5' to 3' gene looping [52].

We used a forward selection procedure by adding one region at a time: (i) all regions were tested separately and the region leading to the highest Spearman correlation between observed and predicted expression was selected as the ‘first’ seed region, (ii) each region not already in the model was added separately and the region yielding the best correlation was selected (‘second region’), (iii) the procedure was repeated till all regions were included in the model. The correlations computed in a cross-validation procedure at each steps are indicated in S2 Table. As shown in Fig 4, the nucleotide composition of intronic sequences had the strongest contribution in the accuracy of our model, followed by UTRs (5' then 3') and CDS (Fig 4). The



**Fig 4. Contribution of additional genomic regions.** Genomic regions were ranked according to their contribution in predicting gene expression. First, all regions were tested separately. Introns yielded the highest Spearman correlation between observed and predicted expressions (in a cross-validation procedure) and was selected as the ‘first’ seed region. Second, each region not already in the model was added separately. 5'UTR in association with introns yielded the best correlation and was therefore selected as the ‘second’ region. Third, the procedure was repeated till all regions were included in the model. The contribution of each region is then visualized starting from the most important (left) to the less important (right). Note that the distance between the second TSS and the first ATG is > 2000 bp for only 189 genes implying that 5'UTR and DD regions overlap. The correlations computed at each steps are indicated in (S2 Table). ns, non significant.

<https://doi.org/10.1371/journal.pcbi.1005921.g004>

nucleotide composition of core promoter moderately increased the prediction accuracy. In contrast the composition of regions flanking core promoter (DU and DD, [Fig 1](#)) as well as regions located downstream the end of gene (DFR, [Fig 1](#)) did not significantly improve the predictions of our model. Note that combining all regions improved the performance of our model compared to promoter alone (compare Figs [2B](#) and [4](#)).

We compared models built on ssDNA and dsDNA, and ssDNA-based models yielded better accuracy [S6 Fig](#). We also compared models built on percentages of nucleotides ( $n = 4$ ), dinucleotides ( $n = 16$ ) and nucleotides+dinucleotides ( $n = 20$ ). As shown [S7A Fig](#), dinucleotides provided stronger prediction accuracy than nucleotides and the best accuracy was obtained combining both nucleotides and dinucleotides. We also built a model on trinucleotide percentage ( $n = 64$ ) ([S7A Fig](#)). This model did yield better results than model built on nucleotide+dinucleotide. However, the correlation increase was not as important as that observed when adding dinucleotides to nucleotides. Besides, the model built on trinucleotides involves more variables and is computationally demanding. We compared models built on nucleotides+dinucleotides adding individually trinucleotide percentages of each region (i.e. 8 models built on nucleotides+dinucleotides in all regions + trinucleotides in one specific region) ([S7B Fig](#)). This analysis revealed that the correlation increase observed when incorporating trinucleotides was mostly due to the contribution of trinucleotides computed in introns, reinforcing our conclusions regarding the importance of sequence-level instructions located in this region.

Because RNA-associated regions (introns, UTRs, CDSs) had greater contribution to the prediction accuracy compared to DNA regions (promoters, DFR), we compared the accuracy of our model in predicting gene vs. transcript expression. We retrieved the normalized results for gene expression (RNAseqV2 rsem.genes.normalized\_results) and the matched normalized expression signal of individual isoforms (RNAseqV2 rsem.isoforms.normalized\_results) for 225 TCGA samples. Accordingly, we generated a set of predictive variables specific to each isoform (see [Material and methods](#)). We found that models built on isoforms are less accurate than models built on genes (median  $r = 0.35$ , [S8 Fig](#) and [\(S3 Table\)](#)). Focusing on the broad nucleotide composition may not be optimal to model isoform expression and to differentiate expression of one isoform from another. Yet another simple explanation could be that reconstructing and quantifying full-length mRNA transcripts is a difficult task, and no satisfying solution exists for now [53]. Consequently isoform as opposed to gene expression is more difficult to measure and thus to predict.

### Additional validation of the model

In the above sections, our complete model, built on 160 variables corresponding to 4 nucleotide and 16 dinucleotide rates in 8 distinct regions ([Fig 1](#)), was trained with a data set containing 241 RNA-seq samples randomly chosen from 12 different cancers, and on 3 independent ENCODE RNA-seq datasets (see TEPIC comparison). We further evaluated our approach using two independent additional datasets: (a) a set of 1,270 RNA-seq samples collected from 14 cancer types and (b) a set of 582 microarray data. Overall, the RNA-seq and the microarray samples were collected from respectively 109 and 41 source sites and sequenced in 3 analysis centers. Similar accuracy was observed in all datasets ([S9](#) and [S10 Figs](#)). Note that the correlations computed with microarray data were lower than that computed with RNA-seq data but involved lower number of genes (9,791 genes in microarrays vs. 16,294 in RNA-seq). For sake of comparison, we restricted RNA-seq data to the 9,791 microarray genes and we observed similar correlation ([S10 Fig](#)). Because our model was built on human reference genome, we also have computed the Spearman correlations between absolute values of CNV segment

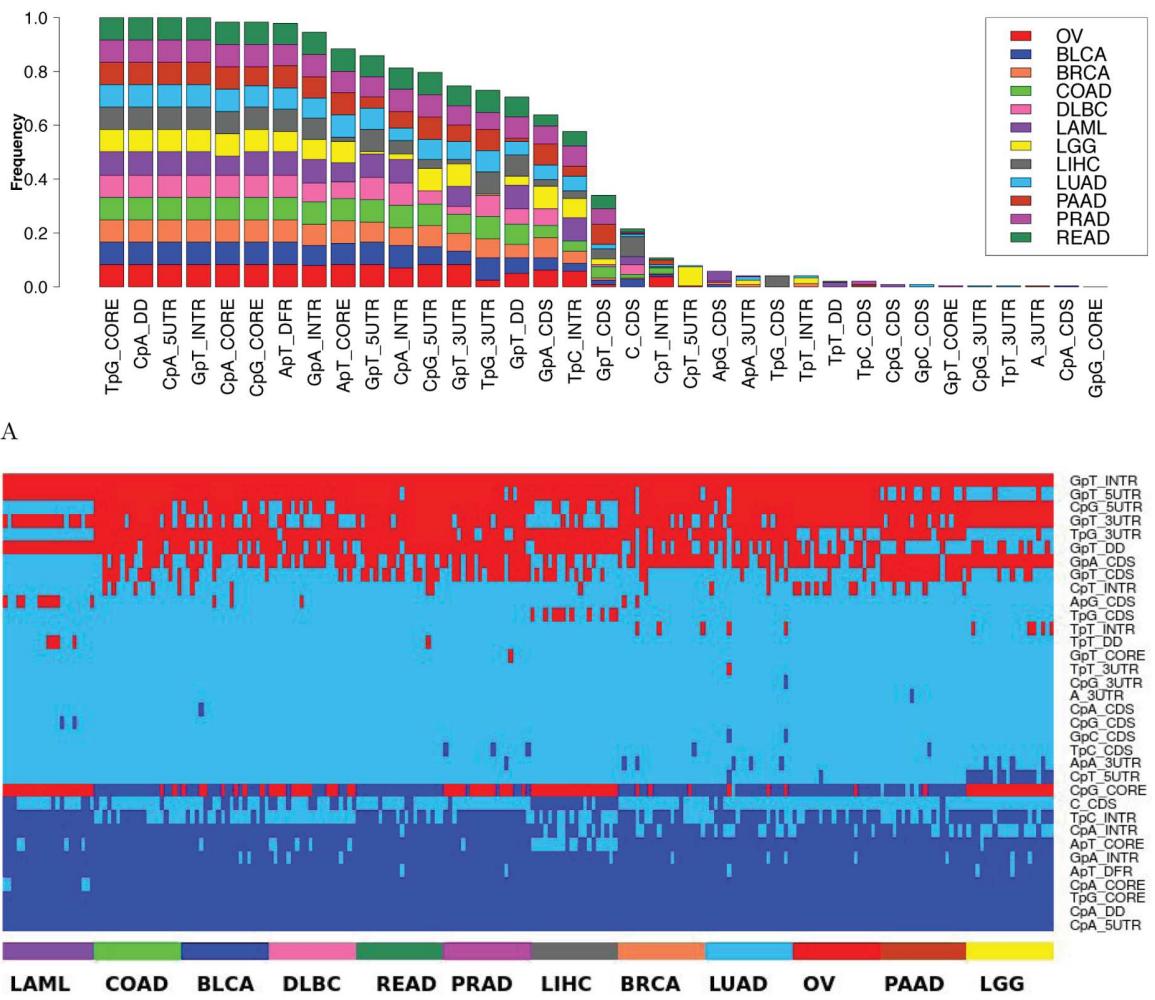
mean scores and model prediction errors calculated for each gene in 241 samples corresponding to 12 cancer types. The median correlation was -0.014, arguing against the model performance being related to CNV-density ([S11 Fig](#)).

### Selecting DNA features related to gene expression

We sought the main DNA features related to gene expression. The complete model built on all 8 regions (160 variables) selected ~ 129 predictive variables per sample. We used the stability selection algorithm developed by Meinshausen *et al.* [35] to identify the variables that are consistently selected after data subsampling (see [Materials and methods](#) for a complete description of the procedure). This procedure selected a median of ~ 16 variables per sample. The barplot in [Fig 5A](#) shows, for each variable, the proportion of samples in which the variable is selected with high consistency (> 70% of the subsets).

We next determined whether stable variables exert a positive (activating) or a negative (inhibiting) effect on gene expression. For each sample, we fitted a linear regression model predicting gene expression using only the standardized variables that are stable for this sample. The activating/inhibiting effect of a variable is then indicated by the sign of its regression coefficient: < 0 for a negative effect and > 0 for a positive effect. The outcome of these analyses for all variables and all samples is shown [Fig 5B](#). With the noticeable exception of CpG in the core promoter, all stable variables had an invariable positive (e.g. GpT in introns) or negative (e.g. CpA in DD and in 5UTR) contribution in gene expression prediction in all samples. In contrast, CpG in the core promoter had an alternating effect being positive in LAML and LGG for instance while negative in READ. It is also the only variable with a regression coefficient close to 0 (absolute value of median = 0.1, see [S12 Fig](#)), providing a partial explanation for the observed changes. As CpG methylation inhibits gene expression [38], we also investigated potential differences in core promoter methylation in LAML (positive contribution of CpG\_CORE) and READ (negative contribution of CpG\_CORE). We used the Illumina Infinium Human DNA Methylation 450 made available by TCGA and focused on the estimated methylation level (beta values) of the sites intersecting with the core promoter. We noticed that core promoters in LAML were overall more methylated (median = 0.85) than in READ (median = 0.69, wilcoxon test p-value < 2.2e-16), opposite to the sign of CpG coefficient in LAML (positive contribution of CpG\_CORE) and READ (negative contribution of CpG\_CORE). This argued against a contribution of methylation in the alternating effect of CpG\_CORE.

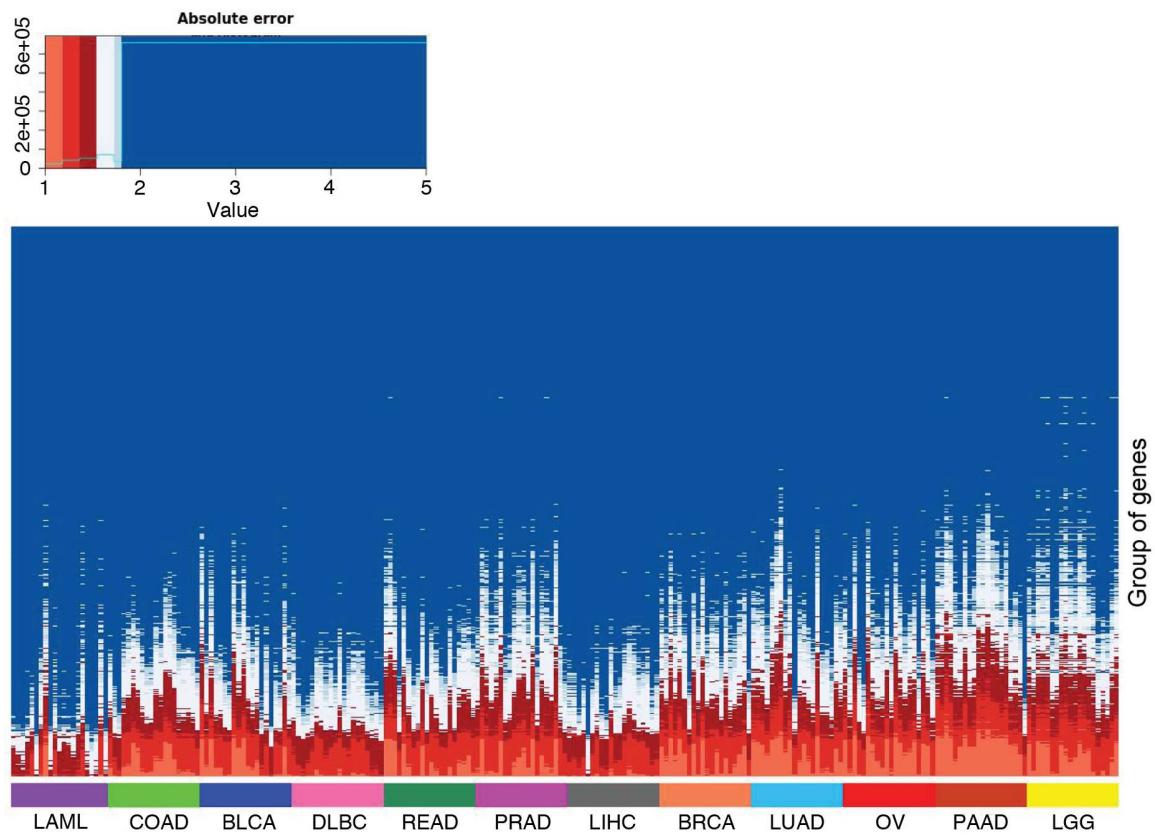
We observed that the accuracy of our model varied between cancer types ([S9 Fig](#)). In order to characterize well predicted genes in each sample, we used a regression tree [54] to classify genes according to the prediction accuracy of our model (i.e. absolute error). The nucleotide and dinucleotide compositions of the various considered regions were used as classifiers. This approach identified groups of genes with similar (di)nucleotide composition in the regulatory regions considered and for which our model showed similar accuracy ([S13 Fig](#)). Implicitly, it identified the variables associated with a better or a poorer prediction. We applied this approach to the 241 linear models. The number of groups built by a regression tree differs from one sample to another (average number = 14). The resulting 3,680 groups can be visualized in the heatmap depicted in [Fig 6](#), wherein each column represents a sample and each line corresponds to a group of genes identified by a regression tree. This analysis showed that our model is not equally accurate in predicting the expression of all genes but mainly fits certain classes of genes (bottom rows of the heatmap, [Fig 6](#)) with specific genomic features ([S13 Fig](#)). Note that the groups well predicted in all cancers presumably correspond to highly and ubiquitously expressed housekeeping genes: groups with low prediction error in all samples and



**Fig 5. A: Consistently selected variables among 12 types of cancer.** For each variable, the fraction of samples in which the variable is considered as stable (i.e. selected in more than 70% of the subsets after subsampling) is shown. Each color refers to a specific type of cancer. Only variables consistently selected in at least one sample are shown (out of the 160 variables). See [Materials and methods](#) for stable variable selection procedure and cancer acronyms. **B: Biological effect of the stable variables.** For each of the 241 samples (columns), a linear model was fitted using the variables (rows) stable for this sample only. The sign of the contribution of each variable in each sample is represented as follows: red for positive contribution, dark blue for negative contribution and sky blue refers to variables not selected (i.e. not stably selected for the considered sample). Only the variables stable in at least one sample are represented. Cancers and samples from the same cancer types are ranked by decreasing mean error of the linear model.

<https://doi.org/10.1371/journal.pcbi.1005921.g005>

cancer types (see [S13 Fig](#) for an example group of 996 genes identified by a regression tree learned in one PRAD sample) are functionally enriched for general and widespread biological processes ([S4 Table](#)). In contrast, groups well predicted in only certain cancers were associated to specific biological function. For instance, a regression tree learned on one PAAD sample identified a group of 1,531 genes, which has low prediction error in LGG and PAAD samples but high error in LAML, LIHC and DLBC samples ([Fig 6](#) and [S13 Fig](#)). Functional annotation of this group showed that, in contrast to the group described above ([S13 Fig](#) and [S4 Table](#)), this group is also linked to specific biological processes ([S5 Table](#)).



**Fig 6. Gene classification according to prediction accuracy.** Columns represent the various samples gathered by cancer type. Samples from the same cancer type are ranked by decreasing mean squared prediction error. Lines represent the 3,680 groups of gene obtained with the regression trees (one tree for each of the 241 samples) ranked by decreasing mean squared prediction error. Groups gathering the top 25% well predicted genes (error < ~ 1.77) are indicated in red and light blue.

<https://doi.org/10.1371/journal.pcbi.1005921.g006>

We further computed Gini coefficient for 16,134 genes using 8,556 GTEx libraries [55]. Gini coefficient measures statistical dispersion which can be used to measure gene expression ubiquity: value 0 represents genes expressed in all samples, while value 1 represents genes expressed in only one sample. We observed that the correlations obtained between Gini coefficient and model errors in each TCGA sample ranged from 0.22 to 0.36. We also compared model errors associated to first and last quartiles of the Gini coefficient distribution using a Wilcoxon test for each of the 241 samples. The test was invariably significant with maximum p-value =  $2.881e^{-7}$ . Likewise analyses were performed with 1,897 FANTOM CAGE libraries [56] considering 15,904 genes. In that case, correlation between models errors and Gini coefficients ranged from 0.25 to 0.4. Overall these analyses suggested that our model better predicts expression of highly and ubiquitously expressed genes. We do not exclude that, when predicting tissue-specific genes, ChIP-seq data collected from the same tissue may add explanatory power to the sequence model. Note, however, that the model performances vary between cancer and cell types implying that part of cell-specific genes are also well predicted by the model (S9 Fig).

### Relationships between selected nucleotide composition and genome architecture

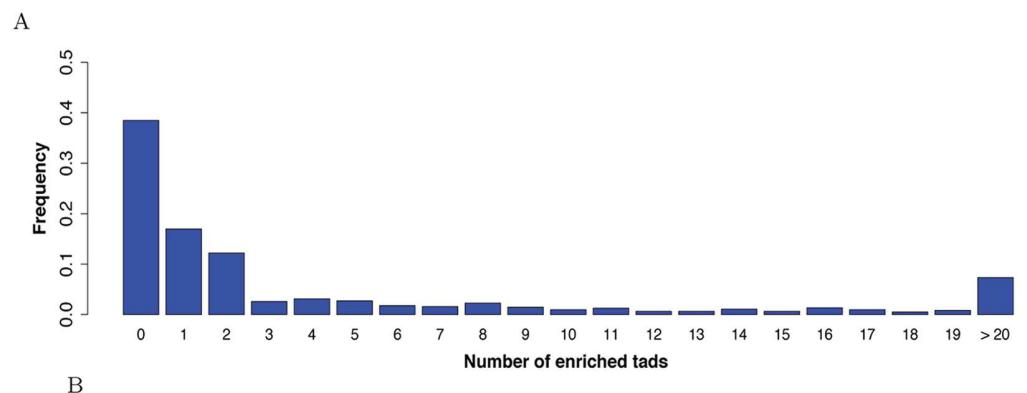
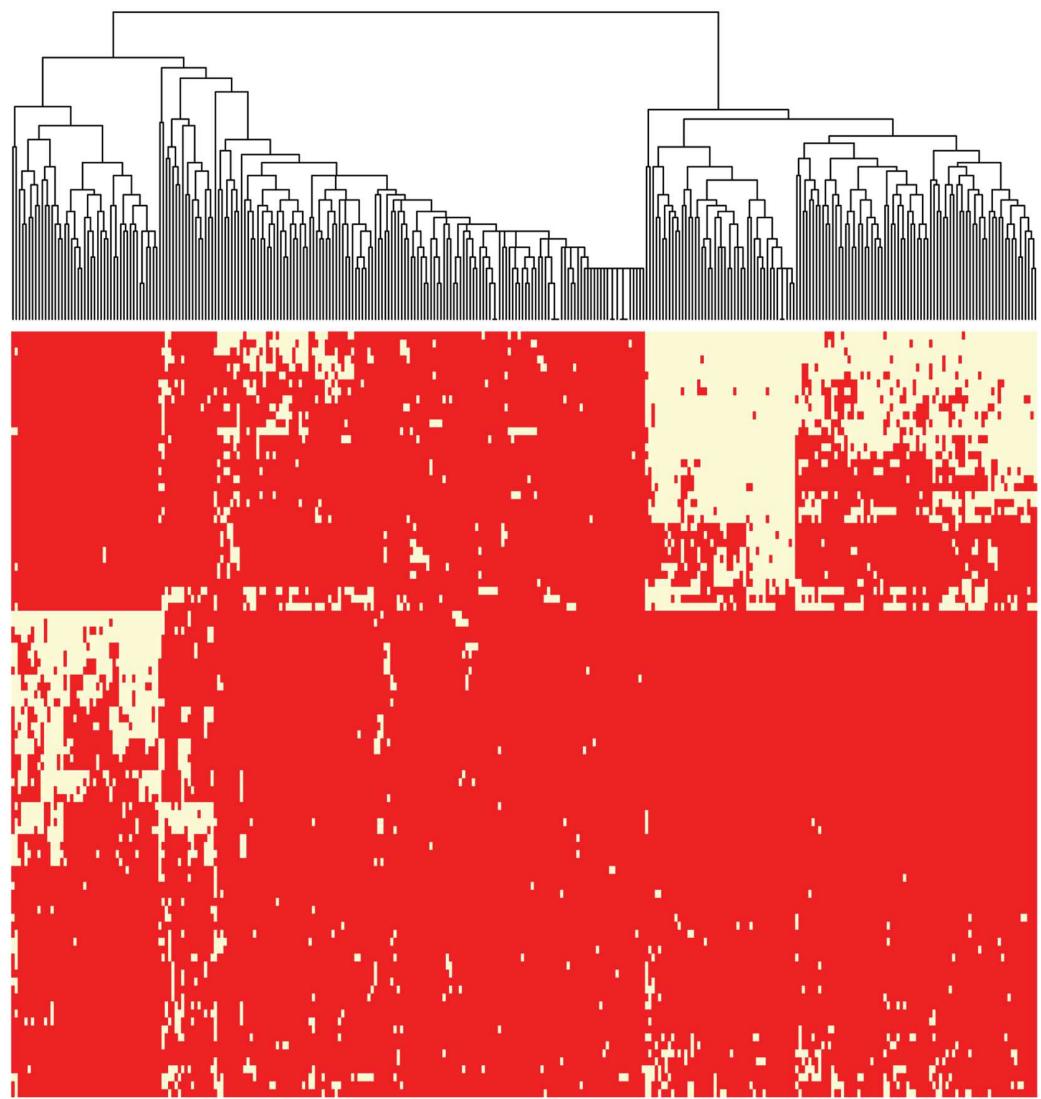
We probed the regulatory activities of the selected regions. We first determined whether introns contained specific regulatory sequence code by assessing the presence of *cis* expression

quantitative trait loci (*cis*-eQTLs). Zhou *et al.* indeed showed that the effect of eQTL SNPs can be predicted from a regulatory sequence code learned from genomic sequences [25]. These findings also implied that *cis*-eQTLs preferentially affect DNA sequences at precise locations (e.g. TF binding sites) rather than global nucleotide composition (i.e. nucleotide/dinucleotide percentages used as variables in our model). We used the v6p GTEx release to compute the average frequencies of *cis*-eQTLs present in the considered genomic regions and directly linked to their host genes ([S6 Table](#)). We noticed that introns contained the smallest density of *cis*-eQTLs (10 times less than any other regions), while containing comparable amount of SNPs ([S7 Table](#)). This result argued against the presence of a regulatory sequence code similar to that observed in promoters for instance [25], despite the presence of enhancers ([S8 Table](#)). These results rather unveiled the existence of another layer of intron-mediated regulation, which involves global nucleotide compositions of larger DNA regions. We then asked whether the groups of genes identified by the regression trees ([Fig 6](#)) correspond to specific TADs. Genes within the same TAD tend to be coordinately expressed [57, 58]. TADs with similar chromatin states tend to associate to form two genomic compartments called A and B: A contains transcriptionally active regions while B corresponds to transcriptionally inactive regions [59]. The driving forces behind this compartmentalization and the transitions between compartments observed in different cell types are not fully understood, but chromatin composition and transcription are supposed to play key roles [5]. Jabbari and Bernardi showed that nucleotide composition along the genome (notably isochores) can help define TADs [60]. As intronic sequences represent ~ 50% of the human genome (1,512,685,844 bp out of 3,137,161,264 according to ENSEMBL merged intron coordinates), the nucleotide composition of introns likely resemble that of neighbor genes and more globally that of the corresponding TAD. We used the 373 TADs containing more than 10 genes mapped in IMR90 cells [6]. For each TAD and each (di)nucleotide, we used a Kolmogorov-Smirnov test to compare the (di)nucleotide distribution of the embedded genes with that of all other genes. We used a Benjamini-Hochberg multiple testing correction to control the False Discovery Rate (FDR), which was fixed at 0.05 (see [Materials and methods](#) section). We found that 324 TADs out of 373 (~ 87%) are characterized by at least one specific nucleotide signature ([Fig 7A](#)). In addition, our results clearly showed the existence of distinct classes of TADs related to GC content (GC-rich, GC-poor and intermediate GC content) ([Fig 7A](#)), in agreement with [60]. We next considered the 967 groups of genes defined in [Fig 6](#) whose expression is accurately predicted by our model (i.e. groups with mean error < mean error of the 1st quartile). We thus focused our analyses on genes for which we did learn some regulatory features. We evaluated the enrichment for specific TADs in each group (considering only TADs containing more than 10 genes) using an hypergeometric test ([Fig 7B](#)). We found that 60% of these groups were enriched for at least one TAD ( $p\text{-value} < 0.05$ ). Hence, several groups of genes identified by the regression trees ([Fig 6](#)) do correspond to specific TADs ([Fig 7B](#)). We concluded that our model, primarily based on intronic sequences, select gene nucleotide compositions that better distinguish active TADs.

## Discussion

In this study, we corroborate the hypothesis that DNA sequence contains information able to explain gene expression [20–25]. We built a global regression model to predict, in any given sample, the expression of the different genes using only nucleotide compositions as predictive variables. Overall our model provided a framework to study gene regulation, in particular the influence of regulatory regions and their associated nucleotide composition.

A surprising result of our study is that sequence-level information is highly predictive of gene expression and in some occasions comparable to reference ChIP-seq data alone [17, 19].



**Fig 7. A: Nucleotide compositions of resident genes distinguish TADs.** For each TAD and for each region considered, the percentage of each nucleotide and dinucleotide associated to the embedded genes were compared to that of all other genes using a Kolmogorov-Smirnov test. Red indicates FDR-corrected  $p$ -value  $\geq 0.05$  and yellow FDR-corrected  $p$ -value  $< 0.05$ . TAD clustering was made using this binary information. Only TADs with at least one  $p$ -value  $< 0.05$  are shown (i.e. 87% of the TADs containing at least 10 genes). y-axis from top to bottom: G\_INTR, GpC\_INTR, CpC\_INTR, CpC\_3UTR,

GpC\_3UTR, G\_3UTR, GpC\_CDS, CpC\_CDS, G\_CDS, G\_DFR, CpC\_DFR, GpC\_INTR, CpG\_3UTR, CpG\_CDS, CpG\_DFR, G\_DU, GpC\_DD, CpG\_DU, CpG\_DD, GpC\_DU, CpC\_DD, G\_DD, GpC\_5UTR, CpG\_5UTR, G\_5UTR, GpC\_CORE, CpG\_CORE, CpC\_CORE, G\_CORE, CpC\_5UTR, CpT\_3UTR, CpT\_CDS, CpT\_INTR, ApT\_INTR, TpA\_INTR, A\_INTR, ApA\_INTR, TpA\_3UTR, ApT\_3UTR, A\_3UTR, ApA\_3UTR, ApA\_CDS, A\_CDS, ApT\_CDS, TpA\_CDS, A\_DD, ApA\_DD, ApT\_DD, TpA\_DD, TpA\_DU, ApT\_DU, ApA\_DU, A\_DU, TpA\_DFR, ApT\_DFR, A\_DFR, ApA\_DFR, ApA\_CORE, A\_CORE, ApT\_CORE, TpA\_CORE, ApA\_5UTR, ApT\_5UTR, A\_5UTR, TpA\_5UTR, ApC\_DFR, ApC\_DD, ApC\_DU, TpC\_DFR, ApC\_CORE, CpA\_DU, CpA\_DFR, CpA\_CDS, ApC\_CDS, ApC\_3UTR, TpC\_CDS, TpC\_CORE, CpT\_5UTR, TpC\_5UTR, CpT\_CORE, TpC\_DD, CpA\_CORE, ApC\_5UTR, CpA\_5UTR, ApC\_INTR, CpA\_DD, CpT\_DFR, CpT\_DD, CpT\_DU, TpC\_3UTR, TpC\_INTR, CpA\_INTR, CpA\_3UTR.

**B: TAD enrichment within groups of genes whose expression is accurately predicted by our model.** The enrichment for each TAD (containing more than 10 genes) in each gene group accurately predicted by our model (i.e. groups with mean error < mean errors of the 1st quartile) was evaluated using an hypergeometric test. The fraction of groups with enriched TADs (p-value < 0.05) is represented.

<https://doi.org/10.1371/journal.pcbi.1005921.g007>

The similar accuracy of models built on real and randomly permuted experimental data indicated that, though the experimental data are biologically relevant, their interpretation through a linear model, in particular inference of TF combinations, is not straightforward as randomization of experimental data did not show the expected loss of accuracy (Fig 3). An interesting perspective would be to devise a strategy to infer TF combinations from experimental data without being influenced by the opening of the chromatin.

The accuracy of our model confirmed that DNA sequence *per se* and basic information like dinucleotide frequencies have very high predictive power. It remains to determine the exact nature of these sequence-level instructions. Interestingly, nucleotide environment contributes to prediction of TF binding sites and motifs bound by a TF have a unique sequence environment that resembles the motif itself [40]. Hence, the potential of the nucleotide content to predict gene expression may be related to the presence of regulatory motifs and TFBSSs. However, we showed that the gene body (introns, CDS and UTRs), as opposed to sequences located upstream (promoter) or downstream (DFR), had the most significant contribution in our model. Moreover, *cis*-eQTL frequencies argue against the presence of a regulatory sequence code in introns similar to that observed in promoters, suggesting the existence of another layer of regulation implicating the nucleotide composition of large DNA regions.

Gene nucleotide compositions vary across the genome and can even help define TAD boundaries [60]. In line with [60], we showed that genes located within the same TAD share similar nucleotide compositions, which provides a nucleotide signature for their TADs (Fig 7A). Our model aimed at predicting gene expression, and therefore intimately linked to TAD compartmentalization, appeared to capture these signatures. Several studies have already demonstrated the existence of sequence-level instructions able to determine genomic interactions. Using an SVM-based approach, Nikumbh *et al* demonstrated that sequence features can determine long-range chromosomal interactions [61]. Similar results were obtained by Singh *et al*. using deep learning-based models [62]. Using biophysical approaches, Kornyshev *et al*. showed that sequence homology influences physical attractive forces between DNA fragments [63]. It would be interesting to determine whether the nucleotide signatures identified by our model are directly implicated in DNA folding and 3D genome architecture.

Finally, although sequence-level instructions are—almost—identical in all cells of an individual, their usage must be cell-type specific to allow proper A/B compartmentalization of TADs, gene expression and ultimately diversity of cell functions. At this stage, the mechanisms driving this cell-type specific selection of nucleotide compositions remain to be characterized.

## Supporting information

**S1 Fig. Comparison of models built on maximum or sum PWM motif scores.** The model was built (i) using 60 nucleotide/dinucleotide percentages computed in the 3 promoter

segments (CORE+DU+DD) and 471 JASPAR2016 PWM maximum scores computed in the CORE segment (pink) or (ii) using 60 nucleotide/dinucleotide percentages computed in the 3 promoter segments (CORE+DU+DD) and 471 JASPAR2016 PWM sum scores computed in the CORE segment (green). All sequences were centered around the 2nd TSS and the 2 models were fitted on 16,294 genes for each of the 241 samples.

(PDF)

**S2 Fig. Dinucleotide local distribution around GENCODEv24 TSSs.** Dinucleotide percentages (y-axis) along 140,604 DNA regions centered around GENCODE v24 TSSs  $\pm 2000$  bp (the distance to TSS is shown in the x-axis). Dinucleotide combinations are represented as first nucleotide on left and second nucleotide on top. The promoter segmentation used in this study (Fig 1) is indicated with vertical dashed lines at -500 bp and 500 bp from the TSS.

(PDF)

**S3 Fig. Number of TSSs by gene.** We considered 19,393 TCGA genes listed in TCGA and the TSSs annotated by GENCODE v24.

(PDF)

**S4 Fig. Contribution in the model of the TSS number.** The model is built using 20 variables corresponding to the nucleotide (4) and dinucleotide (16) percentages computed in the CORE promoter (red), DU (green) or DD (yellow) centered around the second TSS as predictive variables (green). Linear models are also built on the number of isoforms (dark pink) and the number of TSSs (dark blue). Finally models are built using the combinations of variables indicated. All different models were fitted on 19,393 genes for each of the 241 samples considered. The prediction accuracy was evaluated in each sample by evaluating the Spearman correlation coefficients between observed and predicted gene expressions. The correlations obtained in all samples are shown as violin plots. These two last plots underscored the importance of these two variables in predicting gene expression.

(PDF)

**S5 Fig. Gene expression distribution and FANTOM5 enhancer association.** The 19,393 genes listed in one LAML sample (TCGA.AB.2939.03A.01T.0740.13\_LAML) (pink) and a subset of 11,359 genes with assigned FANTOM enhancers (green) were considered. The median expression of genes with assigned enhancers is greater than that of all genes (wilcoxon test p-value < 2.2e-16)

(PDF)

**S6 Fig. Accuracies of models built on dsDNA or ssDNA. A:** Models were built using nucleotide and dinucleotide percentages computed on dsDNA (2 nucleotides + 8 dinucleotides; green violin) or on ssDNA (4 nucleotides + 16 dinucleotides; purple violin) in all the regulatory regions (CORE, DU, DD, 5UTR, CDS, 3UTR, INTR, DFR). The 2 models were fitted on 16,294 genes for each of the 241 samples. The prediction accuracy was evaluated in each sample by evaluating the Spearman correlation coefficients. **B:** Same analyses focusing on each of the indicated regions.

(PDF)

**S7 Fig. Model accuracy with different set of nucleotide predictive variables. A:** Models were built using different set of variables including nucleotide (4 x 8 regions), dinucleotide (16 x 8 regions) and/or trinucleotide (64 x 8 regions) percentages computed in all the regulatory regions (CORE, DU, DD, 5UTR, CDS, 3UTR, INTR, DFR). All different models were fitted on 16,280 genes for each of the 241 samples considered. The prediction accuracy was evaluated in each sample by evaluating the Spearman correlation coefficients. **B:** Models were built using

nucleotide (4 x 8 regions) and dinucleotide (16 x 8 regions) percentages computed in all the regulatory regions and trinucleotide (64) percentages computed in each of the indicated region separately.

(PDF)

**S8 Fig. Forward selection procedure with models built on isoform expressions.** The procedure is identical to that described in [Fig 4](#) but models were built on isoform-specific variables and correlations were computed between observed and predicted isoform expression, not gene expression.

(PDF)

**S9 Fig. Model accuracy in different cancer types.** The model with 160 variables (20 (di)nucleotide rates in 8 regions) was built on 16,294 genes in 241 samples corresponding to the initial training set corresponding to 12 cancer types (**A**) and in an additional set of 1,270 samples corresponding to 14 different cancer types (**B**). The prediction accuracy was evaluated in each sample by evaluating the Spearman correlation coefficients between observed and predicted gene expressions. The correlations obtained in all samples of each data sets are shown as violin plots in **A** (training set) and **B** (additional set). The color code indicates the cancer types. The horizontal dashed lines indicates the median correlation (**A**, 0.582; **B**, 0.577).

(PDF)

**S10 Fig. Comparison on models built on RNA-seq or microarray data.** The model with 160 variables (20 (di)nucleotide rates in 8 regions) was built on 9,791 genes in 582 samples with matched RNA-seq and microarray data. The prediction accuracy was evaluated in each sample by evaluating the Spearman correlation coefficients between observed and predicted gene expressions. The correlations obtained in all samples with RNA-seq- or microarray-built models are shown as violin plots.

(PDF)

**S11 Fig. Spearman correlations between CNV segment mean score and model prediction error.** CNV absolute segment mean scores were computed for each as explained in Materials and Methods section. Model prediction absolute error for each gene are given by our predictive model using nucleotide and dinucleotide percentages computed in all the regulatory regions. Models were fitted on 16,294 genes for each of the 234 on 241 samples having CNV TCGA data available. The median correlation for the 234 samples is -0.014.

(PDF)

**S12 Fig. Absolute values of the regression coefficients.** A linear regression model was built, for each sample, on standardized stable variables only. The boxplots show absolute values of the corresponding coefficients in all samples for each variable considered. Color code as in [Fig 5](#). CpG in the core promoter is highlighted in white. Purple line represents the median of CpG\_CORE coefficients.

(PDF)

**S13 Fig. Example of regression trees learned on two linear models. A: Regression tree leading to a group of genes well predicted in all samples.** This tree has been learned on the sample TCGA.FC.A5OB.01A.11R.A29R.07\_PRAD using all nucleotide composition in all regions. The red path defines a group of 996 genes which has low Lasso error in all samples and cancer types. This group was used for functional annotation ([S4 Table](#)). **B: Regression tree leading to a group of genes well predicted in LGG and PPAD samples.** This tree has been learned on the sample TCGA.IB.7646.01A.11R.2156.07\_PAAD using all nucleotide composition in all

regions. The red path defines a group of 1,531 genes which has low Lasso error in LGG and PAAD samples but high error in LAML, LIHC and DLBC samples. This group was used for functional annotation ([S5 Table](#)).

(PDF)

**S1 Table. Model comparison.** Each model is fitted for each tumor, using all the variables over all regions (160 variables among 8 regulatory regions). First and second columns are median correlation and mean square error over all the tumors. The third column represents mean computing time per tumor (in minutes) on a standard laptop.

(PDF)

**S2 Table. Contributions of additional genomic regions.** Genomic regions were ranked according to their contribution in predicting gene expression. First, all regions were tested separately. Introns yielded the highest Spearman correlation between observed and predicted expressions and was selected as the ‘first’ seed region. Second, each region not already in the model was added separately. 5UTR in association with introns yielded the best correlation and was therefore selected as the ‘second’ region. Third, the procedure was repeated till all regions were included in the model. The contribution of each region is then visualized starting from the most important (left) to the less important (right). The correlations computed at each steps are indicated.

(PDF)

**S3 Table. Correlations between observed and predicted isoform expression.** The procedure is identical to that described in [S2 Table](#) but models were built on isoform-specific variables and correlations were computed between observed and predicted isoform expression, not gene expression.

(PDF)

**S4 Table. Functional enrichment of a group of genes well predicted in all samples.** The group of 996 genes is obtained by fitting a regression tree on the sample TCGA.FC.A5OB.01A.11R.A29R.07\_PRAD using all the nucleotide composition in all regions. These genes are well predicted (mean error < 1st quartile) for all samples of different type cancers. This group of genes was further annotated using the DAVID functional annotation tool. Only the top 5 biological processes indicated by DAVID is shown. The GO term yielded by this analysis corresponded to general and widespread biological processes indicating that these genes likely corresponded to housekeeping genes.

(PDF)

**S5 Table. Functional enrichment of a group of genes well predicted in LGG and PAAD.** The group of 1,531 genes is obtained by fitting a regression tree on the sample TCGA.IB.7646.01A.11R.2156.07\_PAAD using all the nucleotide composition in all regions. These genes are well predicted (mean error < 1st quartile) for all LGG and PAAD samples but not that of LAML, DBLC and LIHC. This group of genes was further annotated using the DAVID functional annotation tool. Only the top 5 biological processes indicated by DAVID is shown. The GO term “Nervous system development” indicates that these genes can be involved in specific biological processes.

(PDF)

**S6 Table. Frequencies of *cis*-eQTLs in the genomic regions considered.** We computed the density of *cis*-eQTL per regulatory region by dividing the sum of *cis*-eQTLs intersecting with the region considered for all genes by the sum of the lengths of the same regulatory region of

all genes. see [Material and methods](#) for details.  
(PDF)

**S7 Table. Frequencies of SNPs in CORE and INTRON regions.** We computed the density of SNPs per regulatory region by dividing the sum of SNPs intersecting with the region considered for all genes by the sum of the lengths of the same regulatory region of all genes. We only considered SNPs detected on chromosomes 1, 2 and 19. see [Material and methods](#) for details.  
(PDF)

**S8 Table. Intersection between enhancers and the genomic regions considered.** We computed the density of enhancers per regulatory region by dividing the total length of the intersection between the enhancers and the region considered for all genes by the sum of the lengths of the same regulatory region of all genes. see [Material and methods](#) for details.  
(PDF)

## Acknowledgments

We thank Mohamed Elati, Mathieu Lajoie, Anthony Mathelier and Cédric Notredame for insightful discussions and suggestions. We also thank Yue Li, Zhaolei Zhang, Florian Schmidt and Marcel H. Schulz for sharing data. We are indebted to the researchers around the globe who generated experimental data and made them freely available. C-H.L. is grateful to Marc Piechaczyk, Edouard Bertrand, Anthony Mathelier and Wyeth W. Wasserman for continued support.

## Author Contributions

**Conceptualization:** Laurent Bréhélin, Sophie Lèbre, Charles-Henri Lecellier.

**Formal analysis:** Chloé Bessière, May Taha, Florent Petitprez, Jimmy Vandel.

**Funding acquisition:** Jean-Michel Marin, Laurent Bréhélin, Sophie Lèbre, Charles-Henri Lecellier.

**Investigation:** Chloé Bessière, May Taha, Florent Petitprez, Jimmy Vandel, Laurent Bréhélin, Sophie Lèbre, Charles-Henri Lecellier.

**Methodology:** Laurent Bréhélin, Sophie Lèbre, Charles-Henri Lecellier.

**Project administration:** Laurent Bréhélin, Sophie Lèbre, Charles-Henri Lecellier.

**Supervision:** Laurent Bréhélin, Sophie Lèbre, Charles-Henri Lecellier.

**Validation:** Chloé Bessière, May Taha.

**Writing – original draft:** Laurent Bréhélin, Sophie Lèbre, Charles-Henri Lecellier.

**Writing – review & editing:** Chloé Bessière, May Taha, Florent Petitprez, Jimmy Vandel, Jean-Michel Marin, Laurent Bréhélin, Sophie Lèbre, Charles-Henri Lecellier.

## References

1. Andersson R, Sandelin A, Danko CG. A unified architecture of transcriptional regulatory elements. *Trends in genetics: TIG*. 2015; 31(8):426–433. <https://doi.org/10.1016/j.tig.2015.05.007> PMID: 26073855
2. Babu D, Fullwood MJ. 3D genome organization in health and disease: emerging opportunities in cancer translational medicine. *Nucleus (Austin, Tex)*. 2015; 6(5):382–393.

3. Ea V, Baudement MO, Lesne A, Forné T. Contribution of Topological Domains and Loop Formation to 3D Chromatin Organization. *Genes*. 2015; 6(3):734–750. <https://doi.org/10.3390/genes6030734> PMID: 26226004
4. Gonzalez-Sandoval A, Gasser SM. On TADs and LADs: Spatial Control Over Gene Expression. *Trends Genet.* 2016; <https://doi.org/10.1016/j.tig.2016.05.004> PMID: 27312344
5. Merkenschlager M, Nora EP. CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. *Annu Rev Genomics Hum Genet.* 2016; 17:17–43. <https://doi.org/10.1146/annurev-genom-083115-022339> PMID: 27089971
6. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 2012; 485(7398):376–380. <https://doi.org/10.1038/nature11082> PMID: 22495300
7. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. *Nature*. 2014; 507(7493):455–461. <https://doi.org/10.1038/nature12787> PMID: 24670763
8. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science*. 2007; 316(5830):1497–1502. <https://doi.org/10.1126/science.1141319> PMID: 17540862
9. Slattery M, Riley T, Liu P, Abe N, Gomez-Alcalá P, Dror I, et al. Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*. 2011; 147(6):1270–1282. <https://doi.org/10.1016/j.cell.2011.10.053> PMID: 22153072
10. Ray D, Kazan H, Chan ET, Pena Castillo L, Chaudhry S, Talukder S, et al. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat Biotechnol*. 2009; 27(7):667–670. <https://doi.org/10.1038/nbt.1550> PMID: 19561594
11. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*. 2004; 5(4):276–287. <https://doi.org/10.1038/nrg1315> PMID: 15131651
12. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, et al. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*. 2010; 140(5):744–752. <https://doi.org/10.1016/j.cell.2010.01.044> PMID: 20211142
13. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. *Nat Rev Genet*. 2014; 15(12):829–845. <https://doi.org/10.1038/nrg3813> PMID: 25365966
14. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489(7414):57–74. <https://doi.org/10.1038/nature11247>
15. Lundberg SM, Tu WB, Raught B, Penn LZ, Hoffman MM, Lee SI. ChromNet: Learning the human chromatin network from all ENCODE ChIP-seq data. *Genome Biol*. 2016; 17:82. <https://doi.org/10.1186/s13059-016-0925-0> PMID: 27139377
16. Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res*. 2012; 22(9):1658–1667. <https://doi.org/10.1101/gr.136838.111> PMID: 22955978
17. Li Y, Liang M, Zhang Z. Regression analysis of combined gene expression regulation in acute myeloid leukemia. *PLoS Comput Biol*. 2014; 10(10):e1003908. <https://doi.org/10.1371/journal.pcbi.1003908> PMID: 25340776
18. Jiang P, Freedman ML, Liu JS, Liu XS. Inference of transcriptional regulation in cancers. *Proc Natl Acad Sci USA*. 2015; 112(25):7731–7736. <https://doi.org/10.1073/pnas.1424272112> PMID: 26056275
19. Schmidt F, Gasparoni N, Gasparoni G, Gianmoena K, Cadenas C, Polansky JK, et al. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res*. 2017; 45(1):54–66. <https://doi.org/10.1093/nar/gkw1061> PMID: 27899623
20. Quante T, Bird A. Do short, frequent DNA sequence motifs mould the epigenome? *Nat Rev Mol Cell Biol*. 2016; 17(4):257–262. PMID: 26837845
21. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, et al. Identification of genetic variants that affect histone modifications in human cells. *Science*. 2013; 342(6159):747–749. <https://doi.org/10.1126/science.1242429> PMID: 24136359
22. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*. 2013; 342(6159):744–747. <https://doi.org/10.1126/science.1242463> PMID: 24136355
23. Kasowski M, Kyriazopoulou-Panagiopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, et al. Extensive variation in chromatin states across humans. *Science*. 2013; 342(6159):750–752. <https://doi.org/10.1126/science.1242510> PMID: 24136358

24. Whitaker JW, Chen Z, Wang W. Predicting the human epigenome from DNA motifs. *Nat Methods*. 2015; 12(3):265–272. <https://doi.org/10.1038/nmeth.3065> PMID: 25240437
25. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods*. 2015; 12(10):931–934. <https://doi.org/10.1038/nmeth.3547> PMID: 26301843
26. Raghava GP, Han JH. Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein. *BMC Bioinformatics*. 2005; 6:59. <https://doi.org/10.1186/1471-2105-6-59> PMID: 15773999
27. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics*. 2014; 47:1–34.
28. Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 2016; 44(D1):D110–115. <https://doi.org/10.1093/nar/gkv1176> PMID: 26531826
29. Chiu TP, Comoglio F, Zhou T, Yang L, Paro R, Rohs R. DNaseR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*. 2016; 32(8):1211–1213. <https://doi.org/10.1093/bioinformatics/btv735> PMID: 26668005
30. Jiao X, Sherman BT, Huang daW, Stephens R, Baseler MW, Lane HC, et al. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*. 2012; 28(13):1805–1806. <https://doi.org/10.1093/bioinformatics/bts251> PMID: 22543366
31. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*. 1996; p. 267–288.
32. R Core Team. R: A Language and Environment for Statistical Computing; 2013. Available from: <http://www.R-project.org/>.
33. Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. Monterey, CA: Wadsworth and Brooks; 1984.
34. Breiman L. Random Forests. *Machine Learning*. 2001; 45(1):5–32. <https://doi.org/10.1023/A:1010933404324>
35. Meinshausen N, Bühlmann P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2010; 72(4):417–473. <https://doi.org/10.1111/j.1467-9868.2010.00740.x>
36. Sill M, Hiebscher T, Becker N, Zucknick M, et al. c060: Extended inference with lasso and elastic-net regularized Cox and generalized linear models. *Journal of Statistical Software*. 2015; 62(5).
37. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society Series B (Methodological)*. 1995; p. 289–300.
38. Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet*. 2012; 13(4):233–245. PMID: 22392219
39. Nguyen TA, Jones RD, Snavely A, Pfenning A, Kirchner R, Hemberg M, et al. High-throughput functional comparison of promoter and enhancer activities. *Genome Res*. 2016;. <https://doi.org/10.1101/gr.204834.116>
40. Dror I, Golan T, Levy C, Rohs R, Mandel-Gutfreund Y. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res*. 2015; 25(9):1268–1280. <https://doi.org/10.1101/gr.184671.114> PMID: 26160164
41. Diamanti K, Umer HM, Kruczak M, Dąbrowski MJ, Cavalli M, Wadelius C, et al. Maps of context-dependent putative regulatory regions and genomic signal interactions. *Nucleic Acids Res*. 2016;. <https://doi.org/10.1093/nar/gkw800> PMID: 27625394
42. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, et al. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*. 2013; 499(7457):172–177. <https://doi.org/10.1038/nature12311> PMID: 23846655
43. Li X, Quon G, Lipshitz HD, Morris Q. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*. 2010; 16(6):1096–1107. <https://doi.org/10.1261/rna.2017210> PMID: 20418358
44. Auweter SD, Oberstrass FC, Allain FH. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res*. 2006; 34(17):4943–4959.
45. Liu C, Mallick B, Long D, Rennie WA, Wolenc A, Carmack CS, et al. CLIP-based prediction of mammalian microRNA binding sites. *Nucleic Acids Res*. 2013; 41(14):e138. <https://doi.org/10.1093/nar/gkt435> PMID: 23703212
46. Boel G, Letso R, Neely H, Price WN, Wong KH, Su M, et al. Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nature*. 2016; 529(7586):358–363. <https://doi.org/10.1038/nature16509> PMID: 26760206

47. Bazzini AA, Del Viso F, Moreno-Mateos MA, Johnstone TG, Vejnar CE, Qin Y, et al. Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. *EMBO J.* 2016;. <https://doi.org/10.1525/embj.201694699>
48. Presnyak V, Alhusaini N, Chen YH, Martin S, Morris N, Kline N, et al. Codon optimality is a major determinant of mRNA stability. *Cell.* 2015; 160(6):1111–1124. <https://doi.org/10.1016/j.cell.2015.02.029> PMID: [25768907](#)
49. Chorev M, Carmel L. The function of introns. *Front Genet.* 2012; 3:55. <https://doi.org/10.3389/fgene.2012.00055> PMID: [22518112](#)
50. Rose AB. Intron-mediated regulation of gene expression. *Curr Top Microbiol Immunol.* 2008; 326:277–290. PMID: [18630758](#)
51. Schwalb B, Michel M, Zacher B, Fruhauf K, Demel C, Tresch A, et al. TT-seq maps the human transient transcriptome. *Science.* 2016; 352(6290):1225–1228. <https://doi.org/10.1126/science.aad9841> PMID: [27257258](#)
52. Bunting KL, Soong TD, Singh R, Jiang Y, Beguelin W, Poloway DW, et al. Multi-tiered Reorganization of the Genome during B Cell Affinity Maturation Anchored by a Germinal Center-Specific Locus Control Region. *Immunity.* 2016; 45(3):497–512. <https://doi.org/10.1016/j.jimmuni.2016.08.012> PMID: [27637145](#)
53. Hayer KE, Pizarro A, Lahens NF, Hogenesch JB, Grant GR. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics.* 2015; 31(24):3938–3945. <https://doi.org/10.1093/bioinformatics/btv488> PMID: [26338770](#)
54. Breiman L, et al. Classification and Regression Trees. New York: Chapman & Hall; 1984.
55. Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. Human genomics. The human transcriptome across tissues and individuals. *Science.* 2015; 348(6235):660–665. <https://doi.org/10.1126/science.aaa0355> PMID: [25954002](#)
56. Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, et al. A promoter-level mammalian expression atlas. *Nature.* 2014; 507(7493):462–470. <https://doi.org/10.1038/nature13182> PMID: [24670764](#)
57. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature.* 2012; 485(7398):381–385. <https://doi.org/10.1038/nature11049> PMID: [22495304](#)
58. Fanucchi S, Shibayama Y, Burd S, Weinberg MS, Mhlanga MM. Chromosomal contact permits transcription between coregulated genes. *Cell.* 2013; 155(3):606–620. <https://doi.org/10.1016/j.cell.2013.09.051> PMID: [24243018](#)
59. Lieberman-Aiden E, Berkum NLv, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science.* 2009; 326(5950):289–293. <https://doi.org/10.1126/science.1181369> PMID: [19815776](#)
60. Jabbari K, Bernardi G. An Isochore Framework Underlies Chromatin Architecture. *PLoS ONE.* 2017; 12(1):e0168023. <https://doi.org/10.1371/journal.pone.0168023> PMID: [28060840](#)
61. Nikumbh S, Pfeifer N. Genetic sequence-based prediction of long-range chromatin interactions suggests a potential role of short tandem repeat sequences in genome organization. *BMC Bioinformatics.* 2017; 18(1):218. <https://doi.org/10.1186/s12859-017-1624-x> PMID: [28420341](#)
62. Singh S, Yang Y, Poczos B, Ma J. Predicting Enhancer-Promoter Interaction from Genomic Sequence with Deep Neural Networks. *BioRxiv.* 2016;
63. Kornyshev AA, Leikin S. Sequence recognition in the pairing of DNA duplexes. *Phys Rev Lett.* 2001; 86(16):3666–3669. <https://doi.org/10.1103/PhysRevLett.86.3666> PMID: [11328049](#)

## 2.2 Caractérisation d'une nouvelle classe de longs ARNs non-codants introniques

Dans la partie précédente (voir partie 2.1), nous avons montré l'importance de la séquence ADN dans la régulation de l'expression des gènes et notamment de la composition des introns en comparaison aux régions régulatrices connues comme les promoteurs. Toujours dans une optique de compréhension des mécanismes de régulation de l'expression des gènes, je me suis intéressée, dans une deuxième partie de ma thèse, à la caractérisation de nouveaux motifs régulateurs dans la partie non-codante du génome. Le consortium FANTOM présenté dans la partie 1.2 a annoté un grand nombre de TSSs à partir de données de CAGE (voir partie 1.3.3), générées dans plus de 1800 conditions. Parmi les TSSs détectés, une minorité intersecte avec les annotations connues, la plus grande partie se trouvant dans les régions non-codantes du génome et dont les fonctions ne sont pas encore déterminées. A partir de ces données, nous avons cherché à identifier de potentiels motifs régulateurs qui nous aideraient à comprendre la présence des signaux de CAGE dans les régions non-codantes et à les annoter.

### 2.2.1 Signal de transcription associé à un motif poly-T

Chez l'homme, 1 048 124 pics de CAGE (TSSs) ont été localisés par FANTOM5 à partir de 1829 conditions différentes. Nous avons utilisé HOMER [100], un outil de recherche de motifs *de novo*, pour trouver les motifs de 21 pb enrichis dans une région de 21 pb centrée autour de ces TSSs, ce qui nous a permis d'identifier une séquence poly-T en amont d'environ 6% des signaux de CAGE. Cette dernière est localisée précisément 2 bases en amont des TSSs et est conservée chez la souris et d'autres mammifères.

### 2.2.2 Caractéristiques des CAGEs associés à un motif poly-T

**Signaux de CAGE associés au motif poly-T : artefact technique versus véritable signal de transcription.** La technique de séquençage Heliscope utilisée par FANTOM pour séquencer les *reads* de CAGE (appelés tags) et caractériser les TSSs peut aussi détecter des séquences poly-T internes aux ARNs séquencés (voir partie 1.3.3) [123]. Il peut donc potentiellement y avoir de nombreux faux positifs. Nous avons étudié les éléments qui sont en faveur de l'hypothèse d'un vrai signal de départ de transcription par rapport à l'hypothèse d'un artefact technique.

Plusieurs points montrent que certains des signaux de CAGE associés à un motif poly-T pourraient provenir d'un artefact technique :

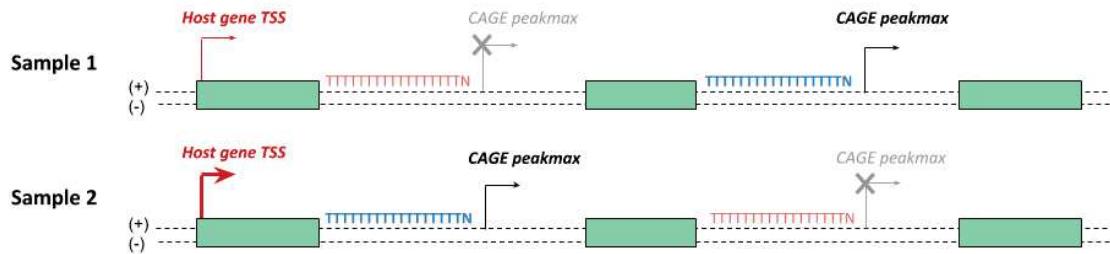
1. Les régions autour des poly-T associés à un signal de CAGE possèdent des caractéristiques spécifiques (régions pauvres en CpG, absence de TATA-box, pas ou peu de transcription antisens). On retrouve cependant le motif INR (Initiator element) caractéristique des promoteurs conventionnels.
2. Les CAGEs associés à un poly-T sont faiblement exprimés.
3. Les CAGEs associés à un poly-T sont majoritairement localisés dans des régions intragéniques et plus particulièrement dans les introns. De plus, ils sont préférentiellement localisés à proximités des débuts de gènes ce qui pourrait favoriser le biais de détection interne.
4. Les CAGEs associés à un poly-T sont majoritairement détectés dans le noyau.
5. Enfin, les CAGEs associés à un poly-T sont préférentiellement détectés dans des gènes fortement exprimés et leur expression est corrélée à celle de leurs gènes hôtes.

Ces différents points convergent vers l'idée d'une synthèse de ces ARNs à partir du pré-ARNm.

D'un autre côté, il a été noté auparavant que, lors du séquençage par la technologie Illumina, la coiffe présente en 5' des ARNs transcrits par la polymérase II pouvait être reconnue comme un nucléotide G pendant l'étape de transcription inverse. Un biais en G est ainsi introduit dans les données de séquençage, ce qui n'est pas le cas avec le séquençage Heliscope. En analysant les données de CAGE générées par ENCODE via le séquenceur Illumina, nous observons un biais en G au niveau de l'extrémité 5' des ARNs, ce qui est comparable aux observations faites pour les gènes codants pour des protéines et est donc en faveur de l'hypothèse d'un vrai signal de transcription, au moins pour une bonne partie des signaux détectés.

#### **Caractéristiques épigénétiques des CAGEs associés au motif poly-T.**

En comparant l'environnement chromatinien et épigénétique des CAGEs associés à un motif poly-T exprimés dans un type cellulaire donné par rapport à ceux qui ne sont pas exprimés, nous avons noté des différences significatives, notamment un enrichissement des marques épigénétiques H3K36me3 et H3K4me3 pour les CAGEs associés à un motif poly-T exprimés (voir Figure 2.2 pour la définition des CAGEs associés à un motif poly-T exprimés et non-exprimés).



**FIGURE 2.2 – Représentation schématique des CAGEs associés à un motif poly-T exprimés versus non-exprimés dans une condition donnée.** Les CAGEs associés à un motif poly-T que l'on appelle exprimés sont caractérisés par un signal de CAGE détecté dans l'échantillon considéré. De plus, on ne les prend en compte que s'ils sont localisés dans un gène contenant au moins 1 CAGE associé à un motif poly-T non-exprimé. D'un autre côté, les CAGEs associés à un motif poly-T non-exprimés sont caractérisés par un signal de CAGE qui n'est pas détecté dans l'échantillon considéré.

H3K4me3 est enrichie au niveau des promoteurs et marque les gènes actifs et H3K36me3 est enrichie dans le cœur des gènes et est associée à l'elongation de la transcription [187]. De plus, la combinaison de ces deux marques, dénotée 'K4-K36 domain', a été utilisée auparavant pour décrire les longs ARNs non-codants intergéniques (*long intergenic non-coding RNA*, lincRNA) [90]. .

**Une partie des CAGEs associés au motif poly-T correspondent à des TSSs de lncRNAs.** D'après les observations précédentes, certains des CAGEs associés à un motif poly-T pourraient correspondre à des TSSs de lncRNAs. Nous avons vérifié cette hypothèse en intersectant les coordonnées des pics de CAGEs associés à un motif poly-T avec celles des débuts de transcrits annotés par FANTOM5. Parmi ces transcrits, environ 3000 coïncident avec un pic de CAGE associé à un motif poly-T. Ces transcrits sont longs (médiane de 2 374 nt) et correspondent à 1 300 gènes différents. D'autres données de séquençage comme celles des longs *reads* de MinION (Nanopore Consortium) viennent confirmer ces observations.

### 2.2.3 Expression des CAGEs associés à un motif poly-T et lien avec leurs gènes hôtes

Pour mieux caractériser ces CAGEs associés à un motif poly-T, nous avons étudié leur expression à travers différents tissus, ainsi que celle des gènes les contenant. Nous avons tout d'abord observé que, pour un gène donné, son expression à travers les différents types cellulaires dépend des CAGEs associés à un motif poly-T qui sont exprimés dans ce tissu. Ainsi, pour un gène ayant des expressions similaires

dans deux tissus différents, le sous-ensemble des CAGEs associés à un motif poly-T exprimés dans ce gène sera très similaire. A l'inverse, les CAGEs associés à un motif poly-T exprimés auront tendance à être différents pour des expressions du gène très éloignées. Cette caractéristique reflète en partie la tissu-spécificité des CAGEs associés à un motif poly-T qui sont exprimés chacun dans un nombre très limité de types cellulaires. De plus, il existe une hiérarchie dans l'expression de ces CAGEs associés à un motif poly-T.

**Interactions entre CAGEs associés à un motif poly-T et promoteur de leur gène hôte.** Grâce à des données d'interaction de ChIA-PET dirigées contre l'ARN Pol II (voir partie 1.1.2), nous avons étudié les interactions impliquant des CAGEs associés à un motif poly-T. Les régions en interaction avec les CAGEs associés à un motif poly-T exprimés sont enrichies dans les prédictions de régions promotrices et TSSs obtenues par les modèles de segmentation du génome ChromHMM/Segway (voir partie 1.2). A l'inverse, les régions en interaction avec les CAGEs associés à un motif poly-T non-exprimés correspondent majoritairement à des régions prédites comme réprimées. De plus, nous observons un enrichissement des interactions des CAGEs associés à un motif poly-T avec le promoteur de leur gène hôte.

**Environnement des gènes contenant des CAGEs associés à un motif poly-T.** En étudiant les gènes contenant des CAGEs associés à un motif poly-T, en comparaison aux autres gènes (annotés dans FANTOM CAT), nous avons mis en évidences des caractéristiques spécifiques. En effet, ces gènes sont plutôt ubiquitaires, malgré la tissu-spécificité des T-motifs qu'ils contiennent, et sont de ce fait enrichis dans des mécanismes biologiques très généraux (biogénèse, organisation des composés cellulaires). De plus, ils sont conservés et sont pauvres en CpG bien qu'ils contiennent plus de TSSs et des promoteurs larges.

**La présence de CAGEs associés à un motif poly-T est sans doute causale à l'expression du gène hôte.** Bien que les CAGEs associés à un motif poly-T et l'expression de leurs gènes hôtes soient liés, les résultats présentés ci-dessus ne nous permettent pas de déterminer la cause et la conséquence. Un exemple de gène contenant un motif poly-T associé à un signal de CAGE est TOMM40. Le variant rs10524523 associé au motif poly-T est localisé dans un intron du gène TOMM40 [199, 162]. Ce *short tandem repeat* (STR) a auparavant été corrélé au déclin cognitif et à la maladie d'Alzheimer et il semblerait que la longueur du motif poly-T soit à

l’origine des variations d’expression du gène hôte observées dans un type cellulaire particulier [199, 162]. Une expérience de clonage avec le gène de la luciférase utilisé comme gène rapporteur confirme l’influence de la longueur du motif poly-T sur l’expression du gène TOMM40 [162]. Les travaux sur l’étude à l’échelle du génome (*genome-wide*) des variants de ces STRs et leur impact sur l’expression du gène hôte sont en cours et effectués en collaboration avec Diego Garrido Martín du *Centre for Genomic Regulation* (CRG, Barcelona, Roderic Guigo Lab).

## 2.2.4 Conclusion

Le motif poly-T associé à un signal de CAGE découvert dans cette analyse est conservé dans de nombreux organismes et il est généralement localisé dans les introns des gènes codants. Il semblerait que ce motif marque une classe de *long non-coding RNAs* agissant directement sur l’expression de leurs gènes hôtes. Pour mieux caractériser ces poly-T associés à un signal de CAGE, des travaux de développement sont en cours pour construire un modèle permettant de discriminer les poly-T associés à un signal de CAGE de ceux qui ne le sont pas, ainsi qu’un modèle pour prédire, à partir du type cellulaire donné, si le poly-T associé à un signal de CAGE est réellement exprimé. Le cas d’étude du poly-T contenu dans l’intron 6 du gène TOMM40 supporte l’idée des variations de la longueur du poly-T entraînant des modifications de l’expression du gène hôte. Des expériences complémentaires seraient requises pour confirmer la présence de ce signal de transcription et son rôle direct sur l’expression du gène hôte. Toutefois, les analyses effectuées sont des indices en faveur de cette hypothèse.

## Article en cours de rédaction

Le présent projet fait l’objet d’une publication en cours de rédaction, dont la version préliminaire est présentée ci-dessous.

# **PolyT tracts can initiate transcription of sense intronic long non-coding RNAs.**

Chloé Bessière<sup>1,2</sup>, Christophe Menichelli<sup>1,3</sup>, Diego Martin Garrido<sup>4</sup>, FANTOM consortium, Roderic Guigo<sup>4</sup>, Michiel J.L. de Hoon<sup>5,6</sup>, Laurent Bréhélin<sup>1,3,7</sup> & Charles-Henri Lecellier<sup>1,2,7</sup>

<sup>1</sup>*IBC, Montpellier, France*

<sup>2</sup>*Institut de Génétique Moléculaire de Montpellier, University of Montpellier, CNRS, Montpellier, France*

<sup>3</sup>*LIRMM, CNRS, Univ. Montpellier, Montpellier, France*

<sup>4</sup>*Centre de Regulacio Genomica, Universitat Pompeu Fabra, Barcelona, Spain*

<sup>5</sup>*Division of Genomic Technologies, RIKEN Center for Life Science Technologies, Yokohama, Japan.*

<sup>6</sup>*RIKEN Omics Science Center (OSC), Yokohama, Japan.*

<sup>7</sup>*correspondence: brehelin@lirmm.fr, charles.lecellier@igmm.cnrs.fr*

**The FANTOM5 consortium provided a comprehensive map of numerous human and mouse transcription start sites (TSSs) using the Cap Analysis of Gene Expression technology. Strikingly, most of them could not be assigned to a specific gene and/or initiate at unconventional regions, outside promoters or enhancers. Here, to determine whether these unconventional TSSs, sometimes referred to 'transcriptional noise or junk', are functionally relevant, we looked for novel and conserved regulatory motifs located in their vicinity. We thereby showed that, in all mammalian species studied, a significant fraction of CAGE tags initiate at**

**short tandem repeats (STRs) of thymidines (Ts). Biochemical and genetic evidence further demonstrate that a significant portion of these CAGEs correspond to TSSs of sense intronic lncRNAs expressed in a precise cell-specific and combinatorial manner, which may exert enhancer activity on their host genes. Together, our results extend the repertoire of mammalian lncRNAs and provide molecular insights on how some STRs can affect gene expression.**

## **Introduction**

RNA polymerase II (RNAP-II) transcribes many loci outside annotated protein-coding gene (PCG) promoters<sup>1,2</sup> to generate a diversity of RNAs, including for instance enhancer RNAs<sup>3</sup> and long non-coding RNAs<sup>4</sup>. Although many more functional non-coding RNAs are yet to be discovered, some authors suggest that many of these transcripts, often faintly expressed, can simply be 'noise' or 'junk'<sup>5</sup>. On the other hand many non annotated RNAP-II transcribed regions correspond to open chromatin<sup>1</sup> and *cis*-regulatory modules (CRMs) bound by transcription factors (TFs)<sup>6</sup>. Besides, genome-wide association studies showed that trait-associated loci, including those linked to human diseases, can be found outside canonical gene regions<sup>7-9</sup>. Together, these findings suggest that the noncoding regions of the human genome harbor a plethora of functionally significant elements which can drastically impact genome regulations and functions<sup>9,10</sup> but remain to discover. In that context, using the cap analysis of gene expression (CAGE) technology<sup>11</sup>, the FANTOM5 consortium provided one the most comprehensive maps of TSSs in human and mouse<sup>2</sup>. Integrating multiple collections of transcript models with FANTOM CAGE datasets, Hon *et al.* built an atlas of 27,919 human lncRNAs, among them 19,175 potentially functional RNAs, and provided

a new annotation of the human genome (FANTOM5 CAGE Associated Transcriptome, FANTOM CAT) <sup>4</sup>. Despite this annotation, many CAGE tags can not be assigned to a specific gene and/or initiate at unconventional regions, outside promoters or enhancers, providing an unprecedented mean to further characterize novel non-coding RNAs encoded by the genome 'dark matter' <sup>10</sup> and to decode part of the transcriptional noise. Here, we probed CAGE data collected from various mammalian species by the FANTOM5 consortium <sup>2</sup> from another perspective and specifically looked for novel and conserved regulatory motifs able to classify these CAGE tags. We observe that, in all mammalian species studied, a significant fraction of CAGE tags (between 2.22% in rat and 6.45% in macaque) initiate at polyT repeats i.e. STRs of Ts. Biochemical and genetic evidence demonstrate that a significant portion of these CAGEs do not correspond to technical artifacts <sup>12</sup> but rather encode sense intronic lncRNAs. The expression of these lncRNAs occurs in a precise cell-specific and combinatorial manner, in close relationship with the expression of their host genes, suggestive of enhancer activity. Together, our results further support the existence of novel functional lncRNAs <sup>4</sup> and provide insights on how some STRs can affect gene expression <sup>35</sup>.

## Results

**A significant fraction of CAGE tags initiates at polyT tracts in various mammals.** We first classified FANTOM5 CAGE peaks (1,048,124 human CAGE tags and 158,969 mouse CAGE tags obtained in 1,829 human and 1,073 mouse libraries) according to their DNA sequence. We used HOMER <sup>13</sup> to look for 21bp-long motifs in 21bp-long sequences centered around the CAGE peak summit. As shown in Figure 1, the first motif identified in both human and mouse is the canonical

initiator element INR<sup>14,15</sup>, demonstrating the relevance of our strategy to unveil specific sequence-level and TSS-associated features. A second motif corresponding to a poly-thymidine (polyT) tract starting precisely at -2 bases (0 being the maximum CAGE signal) is identified. This motif is present in 61,907 human and 8,274 mouse CAGE tags, thereafter called polyT-associated CAGEs (Figure 1A and B). For 391 mouse CAGEs, the polyT tract do not start at -2, explaining that, in contrast to human, a strict T at -2 was not detected in the weblogo motif (Figure 1B). Though this observation may reflect genuine biological difference between mouse and human, the simplest explanation could be a slight difference in the computation of the peak summit for these 391 CAGEs. We further looked at CAGE tags collected in dog, chicken, macaque and rat (Supplementary Figure 1) and this polyT motif is invariably detected by HOMER (sometimes even before INR motif). In human, the median size of the polyT tract is 17 bp with a minimum of 9 bp and a maximum size of 64 bp. Similar results are obtained in mouse (median = 21 bp, minimum = 8 bp and maximum = 58 bp). A third G-rich motif is also identified with strong similarities in both human and mouse (Figure 1A and B). However, this motif is not conserved in dog, chicken, macaque and rat (Supplementary Figure 1). Though this third motif may represent a biologically relevant signal in human and mouse, the absence of conservation in other species makes it less relevant for our study. We then focused our analyses on polyT-associated CAGEs .

**PolyT-associated CAGEs are linked to host gene expression.** We first questioned the biological relevance of these CAGE tags. Because Heliscope sequencing used by FANTOM5 may be internally primed at polyT repeats (Supplementary Figure 2), polyT-associated CAGEs may indeed represent technical artifacts<sup>12</sup>. Several features support this idea: (i) Most polyT-associated

CAGEs (91.2%) fall into the 'no TSS' class established by FANTOM5<sup>11</sup>. Accordingly, almost all polyT-associated CAGEs (i.e. 99%) are TATA- and CpG-less, in contrast to gene-assigned CAGEs, which are CpG-rich (49%) but TATA-less (98%). Looking at antisense transcription, we observed that the median distance between the closest CAGE tag located upstream polyT-associated CAGEs on the opposite strand is 12,649 bp while only 540 bp for gene-assigned CAGEs. Accordingly, the strand bias of CAGE signal (i.e. directionality) computed by Hon *et al.*<sup>4</sup> indicates a median of 0.7867 (Supplementary Figure 3A). These two results argue against the presence of divergent (i.e. upstream antisense) RNAP-II transcription associated with polyT-associated CAGEs , as widely observed for canonical TSSs<sup>16</sup>. The CAGE exosome sensitivity scores previously computed by FANTOM<sup>4</sup> indicates that polyT-associated CAGEs are stable transcripts, as defined in<sup>17</sup> (median sensitivity score = 0.116, Supplementary Figure 3B). (ii) polyT-associated CAGEs are faintly expressed (median of median CAGE TPM expression in all samples = 0.3320) compared to all CAGEs detected (median = 1.117, Wilcoxon test p-value < 2.2e-16) or to CAGEs assigned to gene TSS (median = 2.396, Wilcoxon test < 2.2e-16) Figure 2A). (iii) Some polyT-associated CAGEs are enriched at gene starts, although many are located away from gene starts, in contrast to all repeats of more than 9 Ts present in the human genome (called hereafter T repeats, n = 1,337,561, Supplementary Figure 3C, Wilcoxon test p-value < 2.2e-16). This observation is in accordance with the fact that internal random priming during Heliscope sequencing would favor polyT tracts close to TSS. (iv) Using FANTOM CAT<sup>4</sup>, we observed that > 99% of polyT-associated CAGEs are intragenic with > 80% of them located in protein coding genes (PCG) (Table 1). The median number of polyT-associated CAGEs per gene is 2. In contrast, only 52.4% of all T repeats

are located in FANTOM CAT genes (with a median number of T repeats per gene of 7). Similar results were obtained with the GENCODEv19 annotation (Supplementary Table S1). Moreover 48,411 out of 61,907 polyT-associated CAGEs are located in introns (> 78%), while only 40% of all T repeats are intronic (535,206 out of 1,337,561, Fisher's exact test p-value < 2.2e-16). Likewise, in mouse, 6,162 out of 8,274 polyT-associated CAGEs (~ 74%) but 204,328 out of 834,954 T repeats (~ 24%) are located in introns. Hence, polyT-associated CAGEs could arise from introns of messenger RNAs and not being an independent transcription unit. (iv) In line with this, polyT-associated CAGEs are mostly detected in the nuclear compartment (Figure 2B), in contrast to all CAGEs detected, PCG-assigned CAGEs or CAGEs associated with the third motif shown in Figure 1 (Supplementary Figure S4A-C). PolyT-associated CAGEs are even enriched in chromatin RNAs (Supplementary Figure 4D). Note that a similar profile, with a smaller peak, is observed for T repeats located within genes containing polyT-associated CAGEs ('co-localized' T repeats, Supplementary Figure 4D and see below). We considered T repeats co-localized within the same genes to limit the influence of genomic context and associated epigenetics modifications. (v) Finally, polyT-associated CAGEs are preferentially detected in highly expressed genes (Figure 2C and Supplementary Figure S5). The median Spearman correlation coefficient between expression of polyT-associated CAGEs and that of CAGEs assigned to host genes across 1,829 FANTOM5 samples was 0.2694. Though not very high, this coefficient is greater than that observed with random pairs (median = 0.08830, Figure 2D). The correlation is even more evident when considering host gene expression (not assigned CAGEs individually as in Figure 2D) (median = 0.4006, Supplementary Figure S6A) or considering temporal expression (Supplementary

Figure S6B). The ratio between polyT-associated CAGEs truly associated with a CAGE signal (i.e. 'expressed' polyT-associated CAGEs) and all polyT-associated CAGEs detected in a given gene is correlated to the expression of this host gene (median = 0.4771 while median = 0.2277 with random pairs, Wilcoxon test p-value < 2.2e-16) (Figure 2E). All these features show that the expression of polyT-associated CAGEs is intimately linked to that of their host genes thereby suggesting that polyT-associated CAGEs could be artificially generated during sequencing from immature coding RNAs due to the presence of internal T repeats.

In that scenario, provided similar expression profiles, genes containing polyT-associated CAGEs should contain more T repeats than genes devoid of polyT-associated CAGEs. To test this hypothesis, we created, for distinct libraries, two sets of genes corresponding to genes containing or not polyT-associated CAGEs but exhibiting a similar expression profile (Supplementary Figure S7). Only polyT-associated CAGEs truly associated with a CAGE signal in each library were considered. For each of these libraries, we evaluated the density of T repeats within each gene i.e. number of T repeats divided by the length of the gene. A significant difference exists in all libraries tested (Figure 2F) but the density of T repeats in genes devoid of polyT-associated CAGEs is sufficiently high to permit internal priming. The density of internal T repeats is therefore not sufficient to explain the presence of CAGE signal in one group of genes and not in the other. We noticed a striking difference though in gene length (Supplementary Figure S8), that may facilitate internal priming in one case and not in the other. To clarify that point, we further investigated whether at least some polyT-associated CAGEs are associated with canonical features of genuine TSSs.

**Several polyT-associated CAGE tags are truly capped.** We used a strategy described by de Rie *et al.*<sup>18</sup>, which compares CAGE sequencing data obtained by Illumina (ENCODE) vs. Heliscope (FANTOM) technologies. Briefly, the 7-methylguanosine cap at the 5' end of CAGE tags produced by RNA polymerase II can be recognized as a guanine nucleotide during reverse transcription. This artificially introduces mismatched Gs at Illumina tag 5' end, which is not detected with Heliscope<sup>18</sup>. Such G bias is indeed observed at the 3' end of polyT tract (position -2 from FANTOM CAGE summit) using Illumina ENCODE CAGE data produced in Hela-S3 nuclei (Figure 3A), indicating that the 5' end of polyT-associated CAGEs is located 1 bp after the 3' T of the polyT tracts (or 1 bp before the CAGE peak summit). This bias is also observed in other cell types and is comparable to that observed with CAGE tags assigned to gene TSSs (Figure 3B and Supplementary Figure S9). Conversely, most CAGE tag 5' ends perfectly match the sequences of pre-miRNA 3'end, as previously reported<sup>18</sup>, or that of 61,907 randomly chosen genomic positions (Figure 3B and Supplementary Figure S9). Mismatched Gs at the 3' end of all T repeats co-localized with polyT-associated CAGEs are also detected (Figure 3B and Supplementary Figure S9), though the abundance of tags is higher at polyT-associated CAGEs (Figure 3B). These analyses show that polyT-associated CAGEs are truly capped. However, CAGE tags can capture not only TSSs but also the 5' ends of post-transcriptionally processed RNAs<sup>19</sup>. We then determined whether polyT-associated CAGEs are associated with typical epigenetics marks associated with transcription at the chromatin level.

**Several polyT-associated CAGEs are associated with promoter-related epigenetics marks.** First we found that the INR motif is observed in 44% of polyT-associated CAGEs (27,384 / 61,907)

similar to all permissive CAGEs (467,023 / 1,048,124) (Fisher's exact test p-value = 0.1158) but more frequently than in the case of co-localized T repeats (208,753 / 508,385, Fisher's exact test p-value < 2.2e-16). Using the ENCODE DNaseI Hypersensitive Site (DHS) Master List, we noticed that ~ 11% of polyT-associated CAGEs lie in DHSs, while this is true for only ~ 6% of co-localized T repeats (Fisher's exact test p-value < 2.2e-16). We further studied this potential transcription at the gene level in order to preclude the effect of global chromatin/gene environment. We created two sets of 'expressed' and 'non-expressed' polyT-associated CAGEs : polyT-associated CAGEs are considered as 'expressed' if (i) associated with a detectable CAGE signal in the sample considered and (ii) located in a gene containing at least one 'non-expressed' polyT-associated CAGEs . Conversely, 'non-expressed' polyT-associated CAGEs are (i) not detected in the sample considered but detected in other samples and (ii) located in a gene containing at least one 'expressed' polyT-associated CAGEs . Note that there is no difference in the mean distance between 'expressed'/'non-expressed' CAGEs according to the gene start in the 4 libraries considered (Wilcoxon test p-value = 0.3029 in CNhs12325, p-value = 0.7416 in CNhs12331, p-value = 0.03794 in CNhs12334 and p-value = 0.9523 in CNhs10722).

RAMPAGE data obtained in K562 cells <sup>20</sup> confirmed 1% of 'expressed' polyT-associated CAGEs (12 out of 1,309). No signal was detected at 'non expressed' polyT-associated CAGEs (though being more abundant than 'expressed' polyT-associated CAGEs (0 out of 1,647). The fraction of 'expressed' polyT-associated CAGEs validated by RAMPAGE is low but this may simply be explained by the faint expression levels of these particular CAGEs and the depth of RAMPAGE sequencing.

Genome segmentation provided by combined ChromHMM and Segway<sup>21,22</sup> shows that 'expressed' polyT-associated CAGEs are systematically more enriched in regions corresponding to predicted transcribed regions than 'non expressed' polyT-associated CAGEs (Figure 4A, Fisher's exact test p-value < 2.2e-16 in HeLa-S3 and GM12878, p-value = 2.053e-13 in K562), despite being in the same genes and chromatin environment. Note that the fact that 'non-expressed' polyT-associated CAGEs can be located in expressed (i.e. transcribed) genes likely explain why many of these CAGEs lie in transcribed regions.

We then looked at epigenetics marks individually and used Roadmap epigenetics data obtained in H1 embryonic stem cells to compare the epigenetics status of 'expressed' vs. 'non expressed' polyT-associated CAGEs . We observed that, in three replicates of untreated H1 cells CAGE libraries, H3K36me3 and H3K4me3 are invariably enriched in 'expressed' polyT-associated CAGEs (Figure 4B and Supplementary Figure S10). Similar profiles are obtained with ENCODE ChIP-seq data, although less pronounced in GM12878 (Supplementary Figure S11). H3K36me3 is a histone modification mark enriched on the gene body region and associated with transcription elongation<sup>23</sup>. H3K4me3 is a mark classically associated with active or poised transcription start sites<sup>23</sup>. Hence, 'expressed' polyT-associated CAGEs , as opposed to 'non expressed' ones, are associated with H3K4me3/H3K36me3 domains. Interestingly, this type of 'K4-K36' domains have been previously used to discover lncRNAs<sup>24</sup>. Note that, in general, H3K4me3 levels at lncRNAs are rather low<sup>23</sup>. Together our results show that polyT-associated CAGEs can harbor typical transcription-associated chromatin marks. The presence of 'K4-K36' domains further suggest that polyT-associated CAGEs can act as TSSs for long non-coding RNAs.

**Several polyT-associated CAGEs correspond to long non-coding RNA TSSs.** Using FANTOM CAT robust transcript annotation <sup>4</sup>, we noticed that 2,003 summits of polyT-associated CAGEs (3.2%) are located within a window of < 5bp centered around the start of the first exon of 3,163 transcripts (with a median size of 2,374nt, min = 94 and max = 129,890). These 3,163 transcripts correspond to 1,302 genes (4 examples are shown Figure 5). In contrast, only 0.3% (1,305 / 455,141) of co-localized T repeats (end + 2bp) are located < 5bp of a first exon start. Among these 1,302 genes, 698 correspond to genes with only one single transcript (Supplementary Table S2), with 35 known ENSG genes and 663 novel CATG FANTOM genes. ~ 95% of these genes do not have an assigned gene type in FANTOM CAT ('\_na'). These 698 genes are almost exclusively constrained within introns (only 6 genes overlap with a splice donor and 7 with a splice acceptor) and, looking at the stop codon frequency, we could not detect Open Reading Frames within these 698 transcripts (Supplementary Figure S12).

In order to confirm the existence of these sense intronic lncRNAs, we used GM12878 MInION data made available by the Nanopore consortium. We looked for long reads whose starts correspond to polyT-associated CAGEs ( $\pm$  5bp). First we noticed that 1,454 of 61,907 summits of polyT-associated CAGEs (2.3%) are located within < 5bp of the start of MInION long reads, with only 0.7% for co-localized T repeats (3,172 out of 455,141, Fisher's exact test p-value < 2.2e-16). As for RAMPAGE validation, the small fraction of polyT-associated CAGEs validated by MInION may simply be explained by the faint expression levels of these particular CAGEs and the depth of MInION sequencing. Furthermore, focusing on 'expressed' and 'non expressed' polyT-associated CAGEs in GM12878, we found that 6.4% (90/1,387), 13.6% (157/1155) and 15.4% (164/1,064)

of 'expressed' polyT-associated CAGEs in CNhs12331, CNhs12332, CNhs12333 respectively are located within < 5bp of the start of MInION long reads while the same location is found for only 2.2% (28/1,238), 4.4% (79/1,761) and 3.3% (57/1,731) of 'non expressed' polyT-associated CAGEs in the same samples (Fisher's exact test p-value = 1.275e-07 for CNhs12331, < 2.2e-16 for CNhs12332 and CNhs12333).

Together these results confirmed that several polyT-associated CAGEs are genuine transcriptional products and are not mere technical artifacts.

**Specific combinations of polyT-associated CAGEs are linked to host gene expression.** Computing the ratio, for each gene, between the number of embedded polyT-associated CAGEs and that of all T repeats indicates that a median of only 12.9% of T repeats present in FANTOM CAT genes are associated with CAGE tags (Figure 6A, left), despite close proximity between polyT-associated CAGEs and T repeats not associated with CAGE (median distance in human = 1,1183 bp) (Supplementary Figure S13). Likewise only 10% of all T repeats present in mouse GENCODE M1 annotated genes harbor a CAGE signal (Figure 6A, right). PolyT-associated CAGEs are associated with poly-T tracts longer than 'co-localized' T repeats (Supplementary Figure S14, Wilcoxon test p-value < 2.2e-16). Hence not all T repeats can initiate CAGE tags.

We further assessed the tissue-specificity in polyT-associated CAGE expression and examined the expression distribution across the 1,829 FANTOM libraries by computing the Gini coefficient of each CAGE (Figure 6B). The Gini coefficient measures the degree of variation in gene expression profiles in different samples. It ranges from 0 (no variation i.e. expression is identical

in all samples) to 1 (extreme variation, i.e. all expression values are contained in one individual sample). This analysis shows that no polyT-associated CAGE is ubiquitously expressed (i.e. Gini coefficients  $> 0.5$ ). This is in contrast to Gini coefficients computed for all detected CAGEs or PCG-assigned CAGEs (Figure 6B).

Because polyT-associated CAGEs are expressed in a tissue-specific manner, we wondered whether host gene expression is associated with specific non-redundant combinations of polyT-associated CAGEs . We first compared, for each gene, combinations of polyT-associated CAGEs truly associated with a CAGE signal (i.e. 'expressed') in two distinct samples (1,671,706 pairwise comparisons for 2,368 genes containing more than one polyT-associated CAGEs ). We selected for this study only genes with at least two polyT-associated CAGEs and measured, for each gene, the similarity between the sets of 'expressed' polyT-associated CAGEs in two samples, by computing the size of the symmetric difference of these two sets (Figure 6C). We then plotted the sizes computed in all pairwise comparisons distinguishing the situation when gene expression in two samples is similar or different. As shown in Figure 6C, the sets of 'expressed' polyT-associated CAGEs are more similar in two samples when host gene expression is identical (median size of symmetric difference = 0.07046) than when host gene expression is different (median = 0.92869, Wilcoxon test p-value  $< 2.2\text{e-}16$ ).

Because (i) the ratio of 'expressed' polyT-associated CAGEs correlates with host gene expression (Figure 2E) and (ii) the sets of 'expressed' polyT-associated CAGEs tend to be similar with identical host gene expression, we further tested the existence of a hierachal expression.

Namely, given two sets of 'expressed' polyT-associated CAGEs observed in two samples for the same gene, we asked whether one set is included in the other as exemplified in Figure 6D. We specifically computed, in pairs of samples, the ratio between the size of the intersection (pink area in Figure 6D) and the size of the smallest set ( $s_2$  in the toy example of Figure 6D). In that case, if a hierarchy exists, this ratio should be close to 1 (i.e. the smallest set should be entirely included in the largest one). Note that we restricted the analysis to genes containing more than 1 polyT-associated CAGE and for which not all polyT-associated CAGEs are expressed (total number of comparison = 870,703). We also considered random associations between genes and polyT-associated CAGEs as control (total number of comparisons = 823,276). As shown in Figure 6D, the median ratio is 1 for genuine associations but 0.5 for random control, supporting the existence of a selective hierarchy in polyT-associated CAGEs expression. Note that this non random expression also argues against random internal priming during Heliscope sequencing.

Next we computed the ratio between the number of 'expressed' T motifs and the maximum rank of 'expressed' polyT-associated CAGEs within each gene (Supplementary Figure S15A). For this analysis, we did not consider genes without 'non expressed' polyT-associated CAGEs (i.e. genes whose polyT-associated CAGEs are all 'expressed'). As mentioned earlier, if polyT-associated CAGEs were technical artifacts, polyT tracts close to TSS would be favored for internal priming. In that case, the more polyT-associated CAGEs are 'expressed', the more likely priming will start at a polyT tract distant from TSS. For instance, if three polyT-associated CAGEs were expressed within the same gene, the most distant from TSS will be ranked 3rd and, hence, the computed ratio should equal 1 (Supplementary Figure S15A). However, we observed that the median

ratio is 0.57 (Supplementary Figure S15B), meaning that the polyT-associated CAGEs closest to TSS is not invariably the first one 'expressed'. Similar analysis performed shuffling positions of polyT-associated CAGEs among host genes yields a ratio of 0.6 (Supplementary Figure S15). We then concluded in the absence of bias towards polyT tracts close to TSS.

**DNA looping between polyT-associated CAGEs and host gene promoters.** Provided the intimate link between polyT-associated CAGE expression and that of their host gene, we evaluated potential contact between polyT-associated CAGEs and host gene promoters. We used ENCODE Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) directed against RNAP-II in K562 cell line<sup>25</sup>. As in Figure 4, we selected only polyT-associated CAGEs in genes that host at least one 'expressed' polyT-associated CAGEs and one 'non-expressed' polyT-associated CAGEs . We first noticed that 'expressed' polyT-associated CAGEs were enriched in ChIA-PET data compared to 'non-expressed' polyT-associated CAGEs (Figure 7A) , confirming the results depicted in Figure 4. Second, using combined chromHMM/Segway, we observed that regions interacting with 'non-expressed' polyT-associated CAGEs mostly correspond to 'Predicted Repressed or Low Activity' region (Figure 7B). On the other hand, regions interacting with 'expressed' polyT-associated CAGEs mostly correspond to 'Predicted promoter region including TSS' regions (Figure 7B). Third we used the FANTOM5 CAGE classification<sup>2</sup> and showed that CAGEs interacting with 'expressed' polyT-associated CAGEs mostly correspond to 'true' and 'weak' TSSs (Figure 7C). We next calculated how many times 'expressed' polyT-associated CAGEs are associated with their host gene in ChIA-PET and further assessed whether this number, coined n, could have been obtained by chance using an empirical test. Namely, we randomly shuffled the

associations between polyT-associated CAGEs and host genes and computed the number of polyT-associated CAGEs associated with these random host genes in ChIA-PET. This procedure was repeated 10,000 times. We observed that, in all K562 replicates (3 CAGE replicates for each anti-RNAP-II ChIA-PET), n could never be reached with shuffled pairs (empirical test p-value < 1e-4). This test reflects a genuine enrichment for pairs of polyT-associated CAGEs and CAGE assigned to host gene in ChIA-PET data. We also noticed that, among all CAGEs assigned to genes and interacting with 'expressed' polyT-associated CAGEs in K562, CAGEs assigned to host genes showed the strongest expression correlations with polyT-associated CAGEs (computed in all 1,829 samples) (Figure 7D).

**Regulatory landscape of genes hosting polyT-associated CAGEs .** We next asked whether genes containing polyT-associated CAGEs exhibit specific features that distinguish them from other genes. We first compared the Gini coefficients of genes containing polyT-associated CAGEs ( $n = 16,805$  referred hereafter to as 'with' class) vs. gene without polyT-associated CAGEs ( $n = 36,415$  referred hereafter to as 'without' class). We found that the distribution of expression of genes from the 'with' class is bimodal with genes with a Gini coefficient around 0.55 and another population of genes around 0.9 (Figure 8A). Strikingly, the gene population around 0.55 is almost absent in the 'without' class (Figure 8A), indicating that overall the 'with' class is more widely expressed than the 'without' one. In accordance with the expression dispersion measured by the Gini coefficient, gene ontology (GO) term enrichment indicated that genes from the 'with' class (restricted to known ENSG genes,  $n = 13,826$ , Supplementary Table S3A) are linked to general biological processes (e.g. 'cellular component organization or biogenesis' ; Supplementary Table

S3C). Conversely, the 'without' class ( $n = 17,855$ , Supplementary Table S3B) is linked to more specialized cell-specific processes (e.g 'keratinization', 'antimicrobial humoral response' ; Supplementary Table S3D). According to ENSEMBL database, genes from the 'with' class have more orthologs in *Caenorhabditis elegans*, *Galus galus*, *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Danio rerio* (2,447 out of 13,826) than 'without' genes (943 out of 17,847, Fisher's exact test p-value < 2.2e-16). Together these results suggest that genes containing polyT-associated CAGEs play very fundamental biological functions.

The regulatory landscape of both sets of genes was further compared at the chromatin/DNA level. For sake of comparison, because gene expression is associated with specific features (e.g. presence of CpG islands<sup>26</sup>, number of targeting enhancers<sup>27,28</sup>), we used the gene sets defined in Figure 2F corresponding to genes containing or not polyT-associated CAGEs but harboring similar expression profiles in HeLa-S3, GM12878 and K562 cells. We first counted the number of associated canonical enhancers and used sample specific and LASSO-based associations computed by Cao *et al.*<sup>27</sup>. No difference is observed for the two sets of genes (Supplementary Figure S16). Second, we counted the number of TSSs (i.e. CAGE DPI cluster) assigned to each gene of the two sets<sup>2</sup>. Genes containing polyT-associated CAGEs are associated with a greater number of TSSs than genes devoid of polyT-associated CAGEs in all libraries tested (Figure 8B), suggesting that genes containing polyT-associated CAGEs fall into the 'broad promoter' class defined in<sup>26</sup>. These TSSs are more often CpG-less than that of genes devoid of polyT-associated CAGEs (Figure 8C). No difference in TATA box was observed between the two sets of genes and most TSSs, no matter the set of genes they are assigned to, are TATA-less (Figure 8C).

Together these analyses show that, compared to genes without polyT-associated CAGEs, genes containing these CAGEs are more widely expressed, implicated in fundamental biological processes and more conserved. Their promoters are more broader, in accordance with their expression profile<sup>26</sup>, but strikingly CpG islands are not overrepresented.

## Discussion

We have looked for novel and conserved motifs associated with FANTOM CAGE tags and showed that a significant portion of them initiate at polyT tracts in several mammalian species. Although we cannot exclude at this stage that some CAGE tags initiating at polyT tracts are generated by internal priming during Heliscope sequencing, we provide strong evidence that several represent genuine TSSs of sense intronic lncRNAs. Notably we showed that (i) polyT-associated CAGEs are truly capped and associated with typical epigenetic marks, (ii) several of them correspond to annotated long non-coding transcripts detected by various technologies in different laboratories and (ii) their expression is not random but tightly controlled in a cell-specific manner. It is worth noticing that several features insinuating that polyT-associated CAGEs are technical artifacts constitute specific properties of lncRNAs. For instance, the majority of lncRNAs are enriched in the nucleus<sup>40</sup>. Likewise co-expression has been shown to provide valuable insight into lncRNA function<sup>40</sup>. Li *et al.* suggested that CAGEs detected at polyA/T tracts correspond to technical artifacts because they are not detected by GRO-cap<sup>12</sup>. However, in comparison to CAGE, GRO-cap generates fewer reads mapping to introns<sup>41</sup>, making hard to confirm the existence of faintly expressed and intronic CAGEs, such as polyT-associated CAGEs, using GRO-cap.

On the other hand, genes containing polyT-associated CAGEs exhibit distinguishing features compared to genes without polyT-associated CAGEs . First they correspond to long genes (Supplementary Figure S8), which might simply be linked to the existence of internal non-coding genes whose length can peak to several thousands bp. Moreover, they are more conserved, implicated in fundamental biological processes and widely expressed, with their promoters being of the 'broad' type. Strikingly, these promoters are more often devoid of CpG islands. This is in contrast to the idea that ubiquitous genes are CpG-rich<sup>26</sup> thereby suggesting that polyT-associated CAGEs may compensate for the lack of CpG islands in order to ensure gene expression robustness.

The expression of polyT-associated CAGEs is indeed intimately linked to that of their host genes suggesting a potential regulatory role. Not all polyT tracts within a gene can initiate transcription and polyT-associated CAGEs appear to be associated with long T repeats (Supplementary Figure S14). Interestingly polyT tracts constitute a particular class of STRs. STR length variations can influence gene expression<sup>35</sup>. One noticeable example is the variant rs10524523 which is located in TOMM40 intron 6 and associated with a CAGE signal in several FANTOM libraries. This STR was linked to the age of onset of cognitive decline, Alzheimer's disease and sporadic inclusion body myositis<sup>29-34</sup>. The TOMM40 mRNA brain expression appears to be linked to the length of the polyT tract with the longer variant the higher expression<sup>29,30</sup>. This result was confirmed cloning this locus in a luciferase expression system<sup>30</sup>.

PolyT tracts in human and mouse genomes can correspond to the polyA tails of short interspersed nuclear elements (SINEs) typically ~ 300 bp in length<sup>42</sup>. SINEs have been implicated in

various aspects of gene expression regulation (for review<sup>43</sup>) including enhancer activity<sup>56,57</sup>. This latter is likely distinct from the potential effect of sense intronic lncRNAs because enhancer SINEs (eSINEs)<sup>57</sup> harbor typical enhancer chromatin marks<sup>56</sup>, while polyT-associated CAGEs are associated with promoter marks. Moreover, according to RepeatMasker, only a fraction of human polyT-associated CAGEs correspond to SINEs (~ 28%). Very few polyT-associated CAGEs (n = 6 out of 8,274) correspond to one of the 1,150 eSINEs identified in mouse<sup>57</sup>. PolyT-associated CAGEs are also detected in chicken (Supplementary Figure 1) where SINE elements are quite rare<sup>58</sup>.

Dedicated experiments are required to formally support a regulatory function of polyT-associated CAGEs in host gene expression. However it is tempting to speculate that these sense intronic lncRNAs constitute a novel class of enhancer RNAs, which would be specific to certain coding - as opposed to non-coding - genes. Their existence may contribute to the difference observed between highly expressed coding genes and lowly expressed non coding genes while both gene biotypes exhibit similar TF regulations<sup>59</sup>.

Finally, provided that the length of polyT tracts influences transcription (Supplementary Figure S14), it would be interesting to assess the extent of length variation at the population level and its consequences on host gene expression. This would provide a molecular explanation as to how some STRs can contribute to gene expression variation in humans<sup>35</sup>.

## Methods

**Datasets and online resources** All files used in this study can be found at the urls provided in Supplementary Table S4. The hg19 and mm9 genome assemblies were used throughout the study. Intron coordinates were downloaded from UCSC table browser ; group: Genes and Gene Prediction; track: ENSEMBL Genes ; table: knownGene ; region: genome ; output format: BED ('Introns plus', no flanking region). Human and mouse RepeatMasker coordinates were downloaded from UCSC table browser ; group: Repeats; track: RepeatMasker ; table: rmsk ; region: genome. Note that, for RNAP-II ChIA-PET, interacting regions within the same chromosome were further extracted using the column name \$4 of the bed file.

**Motif analyses** The HOMER motif analysis tool <sup>13</sup> was used to find 21bp long-motifs in regions spanning 10bp around CAGE peakmax (*findMotifsGenome.pl* with options -len 21, -size given, -noknown and -norevopp). HOMER was further used to identify CAGEs harboring specific motifs (*annotatePeaks.pl* with option -m) in particular INR motif (motif1 in Figure 1) and polyT tract (motif2 in Figure 1).

**Characterization of T repeats in human and mouse** link to Eric's tool

**Evaluating mismatched G bias at Illumina 5'end CAGE reads** Comparison between Heliscope vs. Illumina CAGE sequencing was performed as in de Rie *et al.* <sup>18</sup>. Briefly, ENCODE CAGE data were downloaded as bam file (Supplementary Table S4) and converted into bed file using samtools view <sup>70</sup> and unix awk as follow:

```
samtools view file.bam | awk '{FS="\t"}BEGIN{OFS="\t"}{$2=="0") print $3,$4-1,$4,$10,$11}'}> file.bed
```

The bedtools intersect<sup>71</sup> was further used to identify all CAGE reads mapped at a given position (see Figure 3B and C). The unix awk command was used to count the number and type of mismatches as follow:

```
intersectBed -a positions_of_interest.bed -b file.bed -wa -wb -s | awk '{if (substr($11,1,6)=="MD:Z:0" && $6=="+") print substr($10,1,1)}' | grep -c "N"
```

with N = {A, C, G or T}, positions\_of\_interest.bed being for instance polyT-associated CAGE coordinates and file.bed being the Illumina CAGE tag coordinates.

Absence of mismatch were counted as:

```
intersectBed -a positions_of_interest.bed -b file.bed -wa -wb -s | awk '{if (substr($11,1,6)!="MD:Z:0" && $6=="+") print $0}' | wc -l
```

focusing on the plus strand.

As a control, 61,907 random positions were generated with bedtools random (-l 1 and -n 61907). We also used the 3' end of the pre-miRNAs, which were defined, as in de Rie *et al.*<sup>18</sup>, as the 3' nucleotide of the mature miRNA on the 3' arm of the pre-miRNA (miRBase V21, see Supplementary Table S4), the expected Drosha cleavage site being immediately downstream of

this nucleotide (pre-miR end + 1 base).

**Gini coefficient.** We used 1,829 FANTOM5 samples to compute the Gini coefficient for 61,907 polyT-associated CAGEs , 130,286 CAGEs assigned to genes and 1,048,124 permissive CAGEs. Gini coefficient measures statistical dispersion and can be used to measure gene expression ubiquity: value 0 indicates genes expressed in all samples while value 1 indicates genes expressed in only one sample. The Gini coefficient was computed, for each gene, using the following Python function:

```
def gini(list_of_values):  
    sorted_list = sorted(list_of_values)  
    height, A = 0, 0  
    for value in sorted_list:  
        height += value  
        A += height - value / 2  
    B = height * len(list_of_values) / 2  
    return (B - A) / B
```

with list\_of\_values being the vector of expression of each gene in all 1,829 samples.

**Functional enrichment** The list of genes containing (Supplementary Table S3A) or not (Supplementary Table S3B) with polyT-associated CAGEs were submitted to GOrilla <sup>72</sup> at <http://cbl-gorilla.cs.technion.ac.il/>. We used the two unranked lists option with genes containing polyT-associated CAGEs as targets and genes without polyT-associated CAGEs as background (Supplementary Table S3C). We also ran the opposite analysis with genes without

polyT-associated CAGEs as targets and genes containing polyT-associated CAGEs as background (Supplementary Table S3D)

**Statistical tests** Fisher's exact tests, Wilcoxon tests and Spearman correlations were computed in R (*fisher.test*, *wilcox.test*, *cor*) or Python (*fisher\_exact*, *mannwhitneyu*, *spearmanr* from *scipy.stats*).

**Other bioinformatics tools** Average chromatin CAGE signal around polyT-associated CAGEs (Figure 2D) were computed using the *agg* subcommand of the *bwtool* tool<sup>73</sup>.

1. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
2. Forrest, A. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
3. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
4. Hon, C. C. *et al.* An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543**, 199–204 (2017).
5. Palazzo, A. F. & Lee, E. S. Non-coding RNA: what is functional and what is junk? *Front Genet* **6**, 2 (2015).
6. Cheneby, J., Gheorghe, M., Artufel, M., Mathelier, A. & Ballester, B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* (2017).
7. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22**, 1748–1759 (2012).
8. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).

9. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 6131–6138 (2014).
10. Clark, M. B., Choudhary, A., Smith, M. A., Taft, R. J. & Mattick, J. S. The dark matter rises: the expanding world of regulatory RNAs. *Essays Biochem.* **54**, 1–16 (2013).
11. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 15776–15781 (2003).
12. Li, C., Lenhard, B. & Luscombe, N. M. Integrated analysis sheds light on evolutionary trajectories of young transcription start sites in the human genome. *Genome Res.* **28**, 676–688 (2018).
13. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
14. Smale, S. T. & Baltimore, D. The "initiator" as a transcription control element. *Cell* **57**, 103–113 (1989).
15. Butler, J. E. & Kadonaga, J. T. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.* **16**, 2583–2592 (2002).
16. Wei, W., Pelechano, V., Jarvelin, A. I. & Steinmetz, L. M. Functional consequences of bidirectional promoters. *Trends Genet.* **27**, 267–276 (2011).

17. Andersson, R. *et al.* Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat Commun* **5**, 5336 (2014).
18. de Rie, D. *et al.* An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat. Biotechnol.* **35**, 872–878 (2017).
19. Fejes-Toth, K. *et al.* Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009).
20. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res.* **23**, 169–180 (2013).
21. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9**, 215–216 (2012). URL <http://www.nature.com/nmeth/journal/v9/n3/full/nmeth.1906.html>.
22. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods* **9**, 473–476 (2012). URL <http://www.nature.com/nmeth/journal/v9/n5/full/nmeth.1937.html>.
23. Natoli, G. & Andrau, J. C. Noncoding transcription at enhancers: general principles and functional models. *Annu. Rev. Genet.* **46**, 1–19 (2012).
24. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).

25. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).
26. Sandelin, A. *et al.* Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat. Rev. Genet.* **8**, 424–436 (2007).
27. Cao, Q. *et al.* Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat. Genet.* **49**, 1428–1436 (2017).
28. Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T. & Flórek, P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat Ecol Evol* **2**, 152–163 (2018).
29. Payton, A. *et al.* A TOMM40 poly-T variant modulates gene expression and is associated with vocabulary ability and decline in nonpathologic aging. *Neurobiol. Aging* **39**, 1–7 (2016).
30. Linnertz, C. *et al.* The cis-regulatory effect of an Alzheimer's disease-associated poly-T locus on expression of TOMM40 and apolipoprotein E genes. *Alzheimers Dement* **10**, 541–551 (2014).
31. Maruszak, A. *et al.* TOMM40 rs10524523 polymorphism's role in late-onset Alzheimer's disease and in longevity. *J. Alzheimers Dis.* **28**, 309–322 (2012).
32. Greenbaum, L. *et al.* The TOMM40 poly-T rs10524523 variant is associated with cognitive performance among non-demented elderly with type 2 diabetes. *Eur Neuropsychopharmacol* **24**, 1492–1499 (2014).

33. Bernardi, L. *et al.* Role of TOMM40 rs10524523 polymorphism in onset of alzheimer's disease caused by the PSEN1 M146L mutation. *J. Alzheimers Dis.* **37**, 285–289 (2013).
34. Mastaglia, F. L. *et al.* Polymorphism in the TOMM40 gene modifies the risk of developing sporadic inclusion body myositis and the age of onset of symptoms. *Neuromuscul. Disord.* **23**, 969–974 (2013).
35. Gymrek, M. *et al.* Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **48**, 22–29 (2016).
36. Gymrek, M., Willems, T., Reich, D. & Erlich, Y. Interpreting short tandem repeat variations in humans using mutational constraint. *Nat. Genet.* **49**, 1495–1501 (2017).
37. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).
38. Perera, D. *et al.* Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* **532**, 259–263 (2016).
39. Kaiser, V. B., Taylor, M. S. & Semple, C. A. Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. *PLoS Genet.* **12**, e1006207 (2016).
40. Bergmann, J. H. *et al.* Regulation of the ESC transcriptome by nuclear long noncoding RNAs. *Genome Res.* **25**, 1336–1346 (2015).

41. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).
42. Goodier, J. L. Restricting retrotransposons: a review. *Mob DNA* **7**, 16 (2016).
43. Mita, P. & Boeke, J. D. How retrotransposons shape genome regulation. *Curr. Opin. Genet. Dev.* **37**, 90–100 (2016).
44. Mariner, P. D. *et al.* Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Mol. Cell* **29**, 499–509 (2008).
45. Vasant, G. & Reynolds, W. F. The consensus sequence of a major Alu subfamily contains a functional retinoic acid response element. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 8229–8233 (1995).
46. Piedrafita, F. J. *et al.* An Alu element in the myeloperoxidase promoter contains a composite SP1-thyroid hormone-retinoic acid response element. *J. Biol. Chem.* **271**, 14412–14420 (1996).
47. Polak, P. & Domany, E. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics* **7**, 133 (2006).
48. Tajaddod, M. *et al.* Transcriptome-wide effects of inverted SINEs on gene expression and their impact on RNA polymerase II activity. *Genome Biol.* **17**, 220 (2016).
49. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.* **41**, 563–571 (2009).

50. Sorek, R., Ast, G. & Graur, D. Alu-containing exons are alternatively spliced. *Genome Res.* **12**, 1060–1067 (2002).
51. Lev-Maor, G., Sorek, R., Shomron, N. & Ast, G. The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science* **300**, 1288–1291 (2003).
52. Karijolich, J., Zhao, Y., Alla, R. & Glaunsinger, B. Genome-wide mapping of infection-induced SINE RNAs reveals a role in selective mRNA export. *Nucleic Acids Res.* **45**, 6194–6208 (2017).
53. Dixon, R. J., Eperon, I. C. & Samani, N. J. Complementary intron sequence motifs associated with human exon repetition: a role for intragenic, inter-transcript interactions in gene expression. *Bioinformatics* **23**, 150–155 (2007).
54. Zhang, X. O. *et al.* Complementary sequence-mediated exon circularization. *Cell* **159**, 134–147 (2014).
55. Zhang, X. O. *et al.* Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res.* **26**, 1277–1287 (2016).
56. Su, M., Han, D., Boyd-Kirkup, J., Yu, X. & Han, J. D. Evolution of Alu elements toward enhancers. *Cell Rep* **7**, 376–385 (2014).
57. Policarpi, C. *et al.* Enhancer SINEs Link Pol III to Pol II Transcription in Neurons. *Cell Rep* **21**, 2879–2894 (2017).

58. Gao, B. *et al.* Low diversity, activity, and density of transposable elements in five avian genomes. *Funct. Integr. Genomics* **17**, 427–439 (2017).
59. Vandel, J., Cassan, O., Lebre, S., Lecellier, C.-H. & Brehelin, L. Probing transcription factor combinatorics in different promoter classes and in enhancers. *bioRxiv* (2018). URL <https://www.biorxiv.org/content/early/2018/03/02/197418>. <https://www.biorxiv.org/content/early/2018/03/02/197418.full.pdf>.
60. Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243 (2018).
61. Ruiz-Velasco, M. *et al.* CTCF-Mediated Chromatin Loops between Promoter and Gene Body Regulate Alternative Splicing across Individuals. *Cell Syst* **5**, 628–637 (2017).
62. Ferrigno, O. *et al.* Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. *Nat. Genet.* **28**, 77–81 (2001).
63. Lunyak, V. V. *et al.* Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science* **317**, 248–251 (2007).
64. Canella, D. *et al.* A multiplicity of factors contributes to selective RNA polymerase III occupancy of a subset of RNA polymerase III genes in mouse liver. *Genome Res.* **22**, 666–680 (2012).
65. Oler, A. J. *et al.* Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. *Nat. Struct. Mol. Biol.* **17**, 620–628 (2010).

66. Raha, D. *et al.* Close association of RNA polymerase II and many transcription factors with Pol III genes. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 3639–3644 (2010).
67. Barski, A. *et al.* Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. *Nat. Struct. Mol. Biol.* **17**, 629–634 (2010).
68. White, R. J. Transcription by RNA polymerase III: more complex than we thought. *Nat. Rev. Genet.* **12**, 459–463 (2011).
69. Ahuja, R. & Kumar, V. Stimulation of Pol III-dependent 5S rRNA and U6 snRNA gene expression by AP-1 transcription factors. *FEBS J.* **284**, 2066–2077 (2017).
70. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
71. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841–842 (2010).
72. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
73. Pohl, A. & Beato, M. bwtool: a tool for bigWig files. *Bioinformatics* **30**, 1618–1619 (2014). URL <http://bioinformatics.oxfordjournals.org/content/30/11/1618>.
74. Severin, J. *et al.* Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. *Nat. Biotechnol.* **32**, 217–219 (2014).

75. Gomez, N. C. *et al.* Widespread Chromatin Accessibility at Repetitive Elements Links Stem Cells with Human Cancer. *Cell Rep* **17**, 1607–1620 (2016).

**Acknowledgements** We thank Wyeth W. Wasserman, Oriol Fornes Crespo, Anthony Mathelier, Jean-Christophe Andrau, Cyril Esnault, Cédric Notredame, Charles Plessy and Chung Hon for insightful discussions and suggestions. We are indebted to the researchers around the globe who generated experimental data and made them freely available. C-H.L. is grateful to Marc Piechaczyk and Edouard Bertrand for continued support.

**Funding** The work was supported by funding from CNRS, *Plan d'Investissement d'Avenir #ANR-11-BINF-0002 Institut de Biologie Computationnelle* (young investigator grant to C-H.L.), Labex NUMEV, INSERM-ITMO Cancer project "LIONS" BIO2015-04 and CNRS International Associated Laboratory "miREGEN".

**Competing Interests** The authors declare that they have no competing financial interests.

**Correspondence** Correspondence and requests for materials should be addressed to L.B. and C-H.L.. (email: brehelin@lirmm.fr and charles.lecellier@igmm.cnrs.fr).

gene class	CAGE-associated T repeats (61,387)	T repeats (701,159)
coding mRNA	50,109 (81.62%)	486,956 (69.45%)
lncRNA_antisense	585 (0.95%)	13,492 (1.92%)
lncRNA_divergent	1,833 (2.98%)	40,348 (5.75%)
lncRNA_intergenic	5,506 (8.97%)	117,651 (16.78%)
lncRNA_sense_intronic	1,694 (2.76%)	11,692 (1.67%)
pseudogene	778 (1.27%)	17,038 (2.43%)
sense_overlap_RNA	373 (0.61%)	2,385 (0.34%)
short_ncRNA	18 (0.03%)	258 (0.04%)
small_RNA	157 (0.25%)	1,851 (0.26%)
structural_RNA	4 (0.006%)	121 (0.02%)
uncertain_coding	330 (0.54%)	9,367 (1.33%)

Table 1: **CAGE-associated T-repeats are preferentially located in coding gene.** The coordinates of polyT-associated CAGEs and that of all repeats of more than 9 Ts (n = 1,337,561) were intersected with the annotation provided by the FANTOM5 CAGE Associated Transcriptome (CAT). The total number of intersections polyT-associated CAGEs and polyT tracts is 61,387 and 701,159 respectively. The location of all T repeats are shown for comparison. For each gene class, the number of intersections and the corresponding percentage is shown.

**Figure 1 Motif discovery around CAGE peakmax.** HOMER<sup>13</sup> was used to find 21bp-long motifs in 21bp-long sequences centered around the peak max of FANTOM5 CAGEs.

**A.** 1,048,124 human permissive CAGEs were used. The top 5 is shown. 61,907 CAGEs are associated with polyT tract starting precisely at -2 (0 being the peakmax). **B.** Among 158,966 mouse CAGEs, 8,274 are associated with a polyT tract.

**Figure 2 CAGEs associated with polyT tracts exhibit specific features.** **A.** Expression medians were computed for each permissive CAGE, CAGE assigned to genes and polyT-associated CAGEs in 1,829 samples. The Phase 1 and 2 combined Tag Per Million (TPM) normalized data were used. Note that the y axis was limited to 20.

**B.** Nucleocytoplasmic distribution of polyT-associated CAGEs . For each indicated library, CAGE expression (RLE normalized) was measured in nuclear and cytoplasmic fractions. Each CAGE was then assigned to nucleus, cytoplasm or both compartments. The number of CAGEs in each class is shown for each sample as a fraction of all detected CAGEs.

**C.** Distribution of median expression in 1,829 FANTOM5 samples of genes containing polyT-associated CAGEs (green) or T repeats (blue) (Wilcoxon test p-value < 2.2e-16).

The median expression of PCGs (red) is shown for sake of comparison. Distribution of median expression of all genes is shown Supplementary Figure 5. **D.** Distribution of Spearman correlations computed between the indicated CAGE pairs in 1,829 FANTOM5 samples.

**E.** Distribution of Spearman correlations computed between expression of CAGEs assigned to host gene in 1,829 FANTOM5 samples and the ratio between the

number of polyT-associated CAGEs truly associated with a CAGE in the sample considered (i.e. 'expressed') and all possible polyT-associated CAGEs detected in the gene considered in all samples. Same analysis was performed with random association between polyT-associated CAGEs and host genes (red). **F.** Number of polyT tracts relative to gene length for genes associated (red) or not (blue) with polyT-associated CAGEs in the library considered (CNhs12325, HeLa-S3 ; CNhs12331, GM12878 ; CNhs12334, K562). Wilcoxon tests were performed to assess the significance of the difference observed (p-value = 1.52e-15 for CNhs12325, p-value = 7.033e-12 for CNhs12331, p-value < 2.2e-16 for CNhs12334)

**Figure 3 polyT-associated CAGEs are capped at 5'end** **A.** G bias in ENCODE CAGE reads was assessed at the indicated positions (x-axis) around polyT-associated CAGEs (0 correspond to CAGE summit). The number of intersecting reads is indicating in bracket (only the positive strand is indicated). **B.** Same analyses as in A but considering position -2 only and other CAGEs/positions indicated in x-axis: CAGE peaks assigned to gene, pre-microRNA 3' ends, 61,907 random positions, co-localized T repeat 3' ends. Note that the number of polyT-associated CAGEs on + strand is 32,171 and that of co-localized T repeats 236,049 indicating that the number of reads at polyT-associated CAGEs is relatively higher than that at col-localized T repeats.

**Figure 4 Several polyT-associated CAGEs are associated with promoter-related epigenetics marks** **A.** Distribution of ChromHMM/Segway genome segments containing 'expressed' or 'non-expressed' polyT-associated CAGEs (see text for definitions) in HeLa-S3, GM12878 and K562. Note the absence of 'Predicted promoter flanking region' class in K562. **B.** Intersection of 'expressed'/'non expressed' polyT-associated CAGEs coordinates with that of Roadmap epigenetics data collected in H1 embryonic stem cells. The fraction obtained for 'expressed' and 'non expressed' polyT-associated CAGEs were compared using Fisher's exact test. The color correspond to p-value < 5e-3 (green) or  $\geq 5e-3$  (grey).

**Figure 5 Several polyT-associated CAGEs correspond to TSSs of FANTOM\_CAT robust transcripts.** Four examples of polyT-associated CAGEs initiating  $< 5\text{bp}$  upstream an exon start are shown as Zenbu views<sup>74</sup>. The FANTOM names of polyT-associated CAGEs , single-exon transcripts and corresponding genes are indicated. D provides an example of a gene with a single transcript.

**Figure 6 Specific combinations of polyT-associated CAGEs are linked to host gene expression.** **A.** For each gene, we computed the ratio between polyT-associated CAGEs and T repeats present in the same gene in human (left) and mouse (right). We used human FANTOM CAT and mouse GENCODE M1 gene annotations. The dashed line represents the median fraction. **B.** Gini coefficients were computed for all permissive CAGEs, CAGEs assigned to protein coding genes (PCGs) and T repeat-associated

CAGEs in 1,829 samples. The distribution of the coefficients calculated for each CAGE is shown. Note that a Gini coefficient close to 0 indicates that CAGE tags are detected in almost all samples and, conversely, a Gini coefficient close to 1 indicates that the expression is restricted to few samples. **C.** Size of symmetric differences (pink area) between combinations of polyT-associated CAGEs computed for each gene in pairs of samples with similar (blue) or dissimilar (red) gene expression. **D.** The existence of a potential hierarchy in polyT-associated CAGEs linked to host gene expression was evaluated. A toy example is shown (left). For each gene and for each pairs of samples, the ratio between the intersection of combinations of 'expressed' polyT-associated CAGEs in the two samples and the combination minimal length was computed. the distribution is shown as density plots. The blue dashed line represents the median ration. Same analysis was performed with random association between polyT-associated CAGEs and host genes (red).

### **Figure 7 DNA looping between polyT-associated CAGEs and host gene promoters**

**A.** The number of polyT-associated CAGEs associated with a CAGE signal ('expressed') or not ('non-expressed') in K562 cells located (red) or not (green) in ChIA-PET interacting regions was calculated intersecting their genomic coordinates. Results are shown in the case of CNhs12334 CAGE and RNAP-II ENCODE K562 ChIA-PET replicate 2 data. A Fisher's exact test was computed to assess the statistical significance of the results ( $p$ -value = 1.166e-05). Similar results were observed with other CAGE and /or ChIA-PET replicates. **B.** The coordinates of the regions interacting with 'expressed' or

'non-expressed' polyT-associated CAGEs were intersected with that of the genome segments provided by combined chromHMM/Segway in K562. Fisher's exact tests were performed to assess potential enrichments in the indicated segments (\*,  $< 0.05$ ; \*\*,  $< 0.005$ ). Similar results were observed with other CAGE and /or ChIA-PET replicates. Note however that, in the case of CNhs12336, only 15 regions interact with 'non-expressed' polyT-associated CAGEs . E, 'Predicted enhancer' ; R, 'Predicted Repressed or Low Activity region' ; T, 'Predicted transcribed region' ; TSS, 'Predicted promoter region including TSS'. **C.** The coordinates of the regions interacting with 'expressed' or 'non-expressed' polyT-associated CAGEs were FANTOM5 CAGE coordinates and the number of CAGEs in each class was calculated. Fisher's exact tests were performed to assess potential enrichments in the indicated classes (\*\*,  $< 0.005$ ) D. Expression correlation in 1,829 FANTOM5 samples was computed for all pairs of polyT-associated CAGEs and gene-assigned CAGEs interacting in K562 ChIA-PET data (red and dashed line, median = 0.2289). Similar analyses were performed with pairs of polyT-associated CAGEs and host (green) or non-host (blue) genes (median = 0.28309 and 0.13407 respectively). Medians are indicated as vertical lines. Wilcoxon tests were performed to assess the significance of the results ( $p$ -value = 1.64e-13 between green and red,  $< 2.2e-16$  otherwise).

**Figure 8 Regulatory landscape of genes containing polyT-associated CAGEs A.** Gini coefficients were computed as in Figure 6B for all FANTOM CAT genes (red), genes containing polyT-associated CAGEs (green) and genes without polyT-associated CAGEs (blue). The distribution of the coefficients calculated for each gene is shown. **B.** The

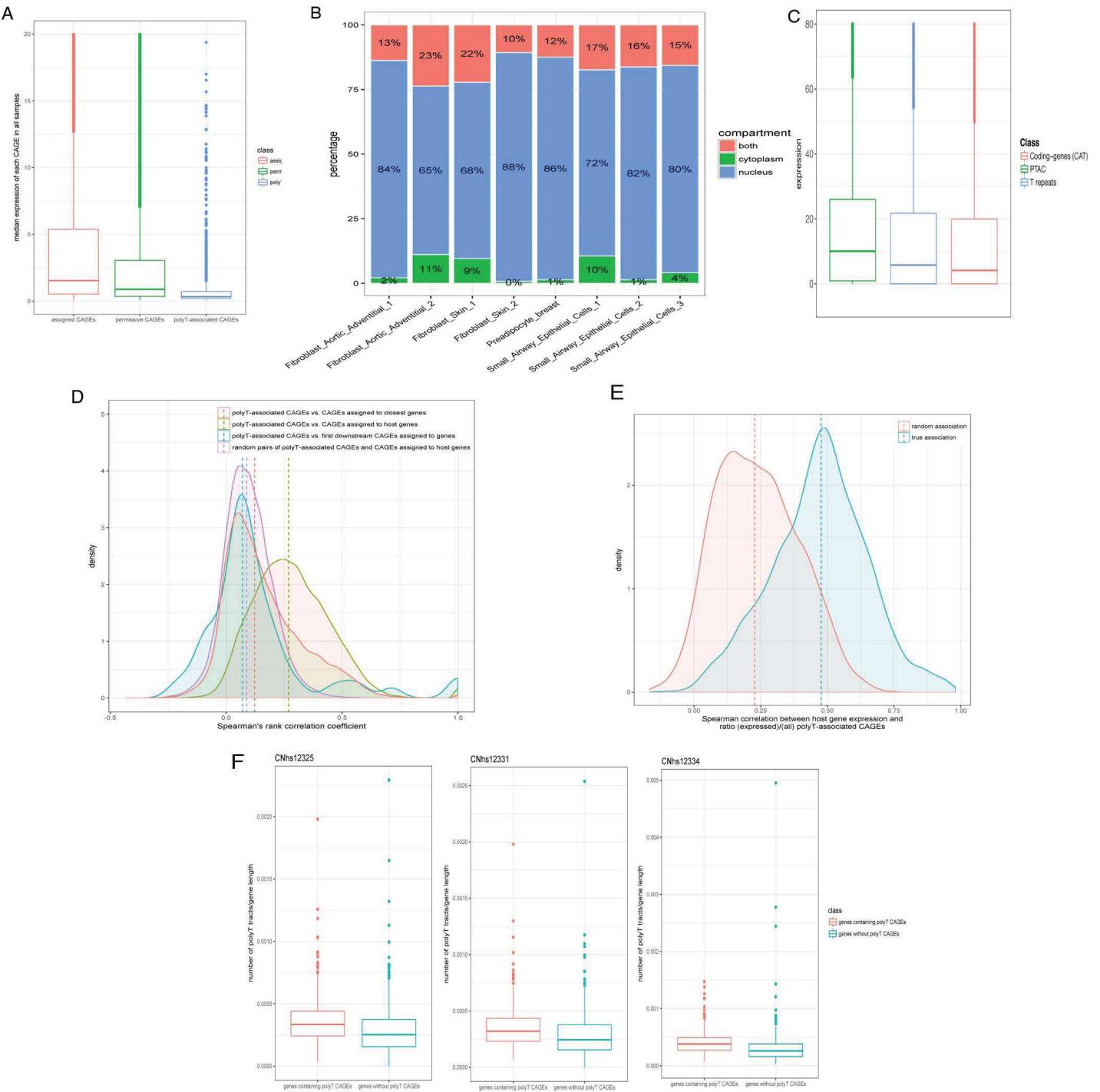
number of TSSs (i.e. CAGE DPI cluster) was calculated per gene using FANTOM CAT. Genes containing polyT-associated CAGEs truly expressed in CNhs12325 (HeLa-S3), CNhs12331 (GM12878) and CNhs12334 (K562) were listed (red). For each of these genes, a gene devoid of polyT-associated CAGEs but exhibiting a similar expression was identified (blue). The distributions of numbers of TSSs per gene are shown as boxplots and the Wilcoxon test p-value is indicated. **C.** The presence of CpG islands and TATAbox was evaluated in the sets of genes defined in B (i.e genes containing ('with') or not ('without') polyT-associated CAGEs ). To annotate genes, we used FANTOM CAGE to gene association and CAGE classification into CpG/CpG-less (top) and TATA/TATA-less (bottom) (see Table S4 for annotation files). Genes containing polyT-associated CAGEs ('with') are mostly associated with CpG-less CAGEs (Wilcoxon test p-value < 2.2e-16). No difference is observed for TATA box (most CAGEs are TATA-less).

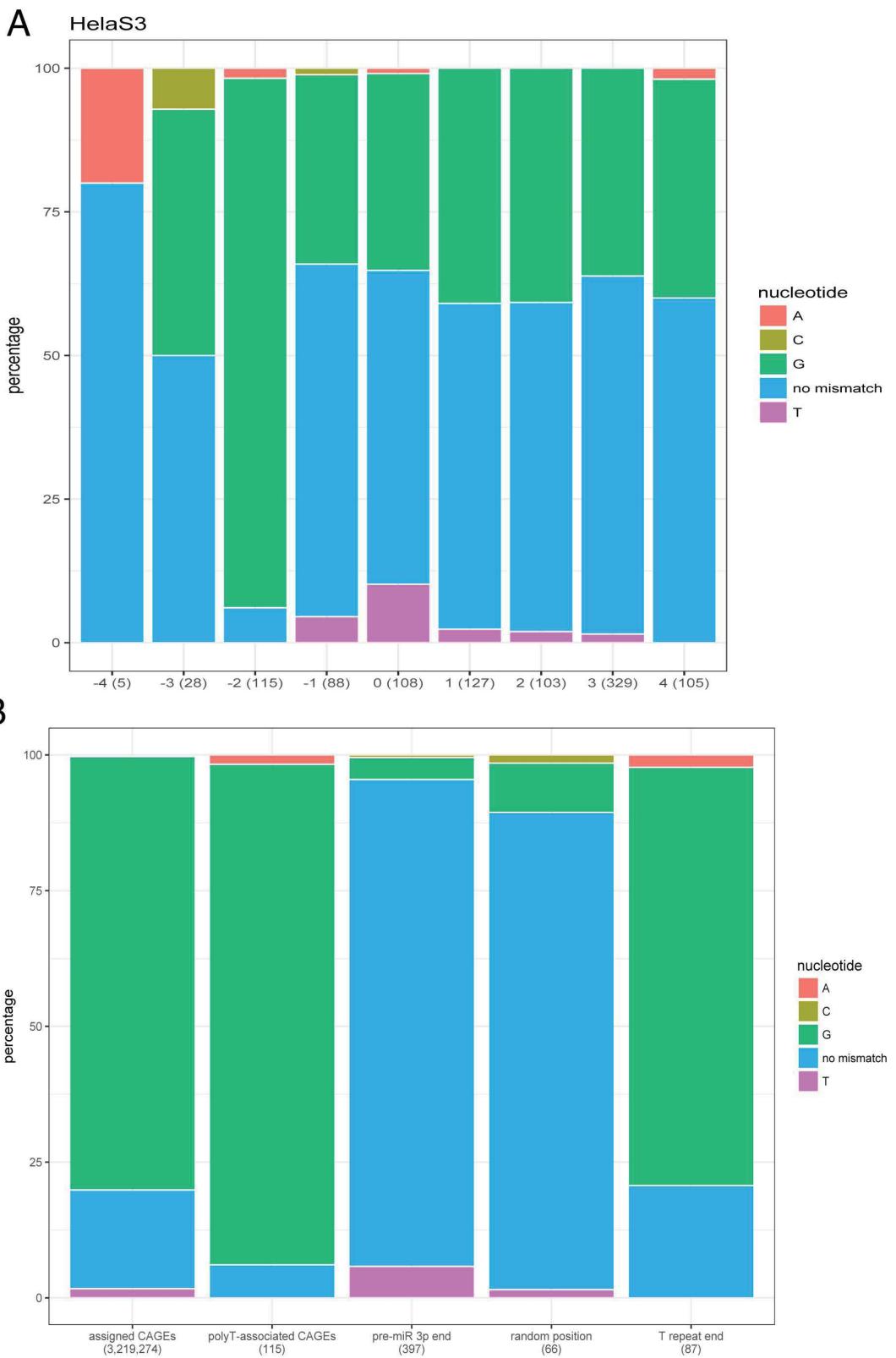
**A**

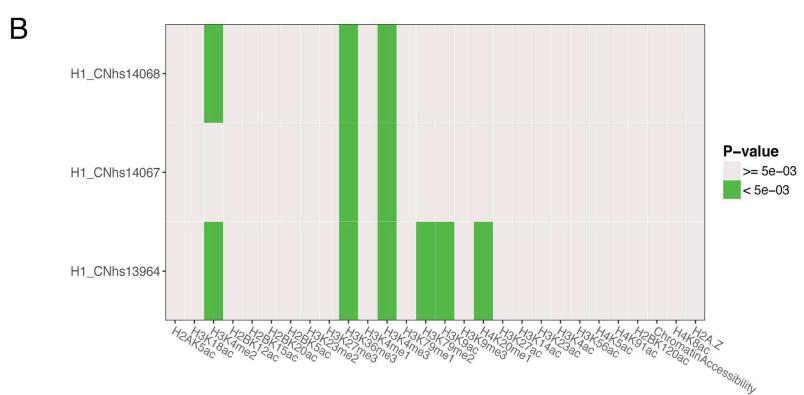
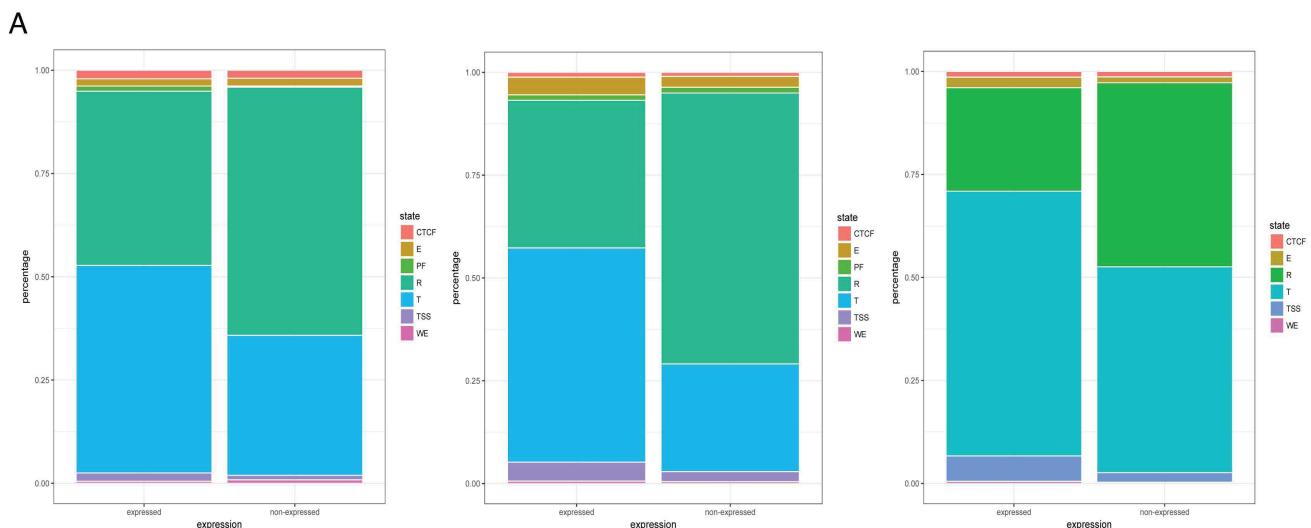
rank	motif	p-value	% in foreground	% in background
1		1e-125974	42.37	12.49
2		1e-99010	5.91	0.06
3		1e-27467	5.39	0.76
4		1e-13220	8.24	3.19
5		1e-10229	10.53	5.21
-10	position relative to CAGE peakmax			

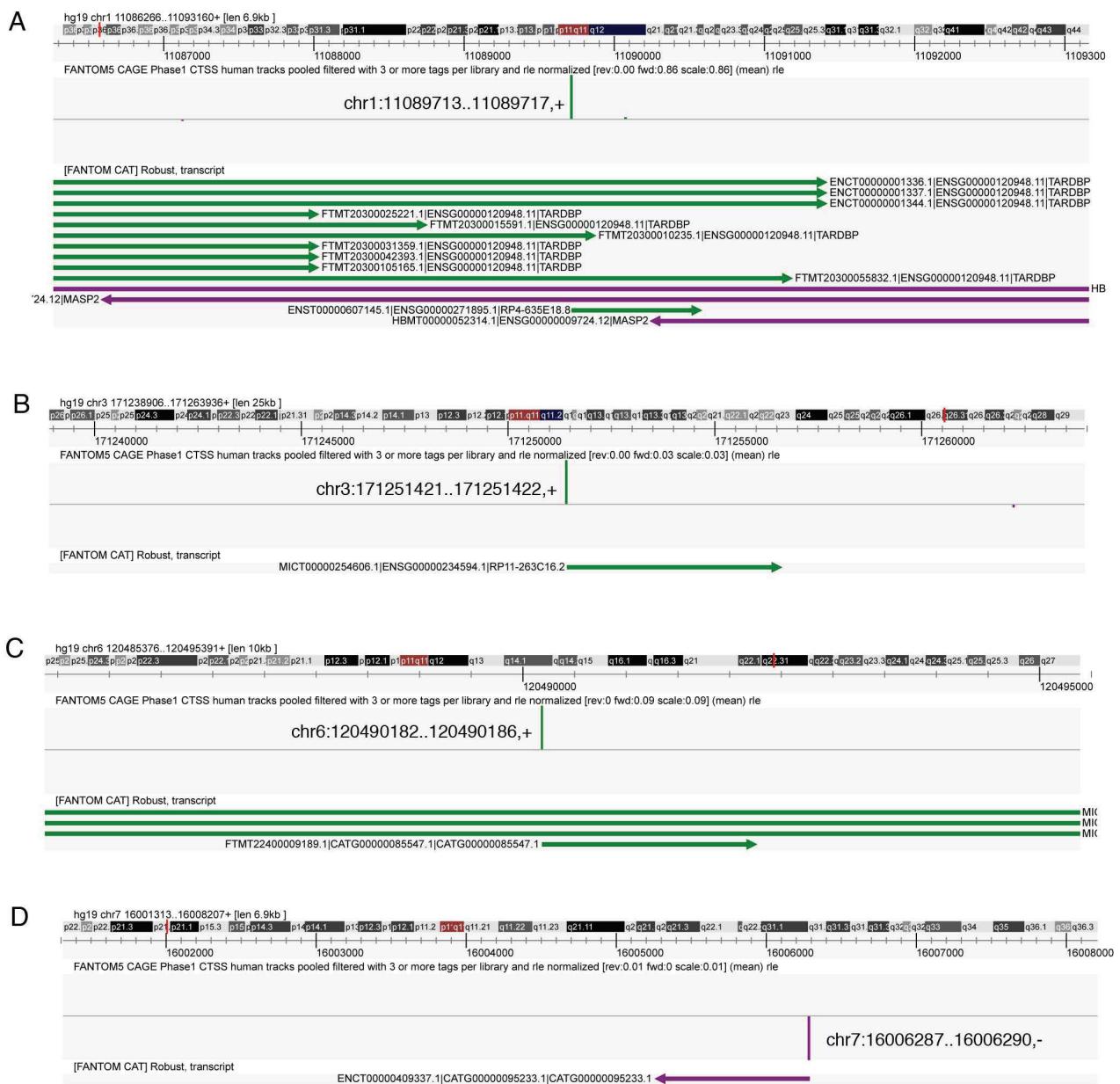
**B**

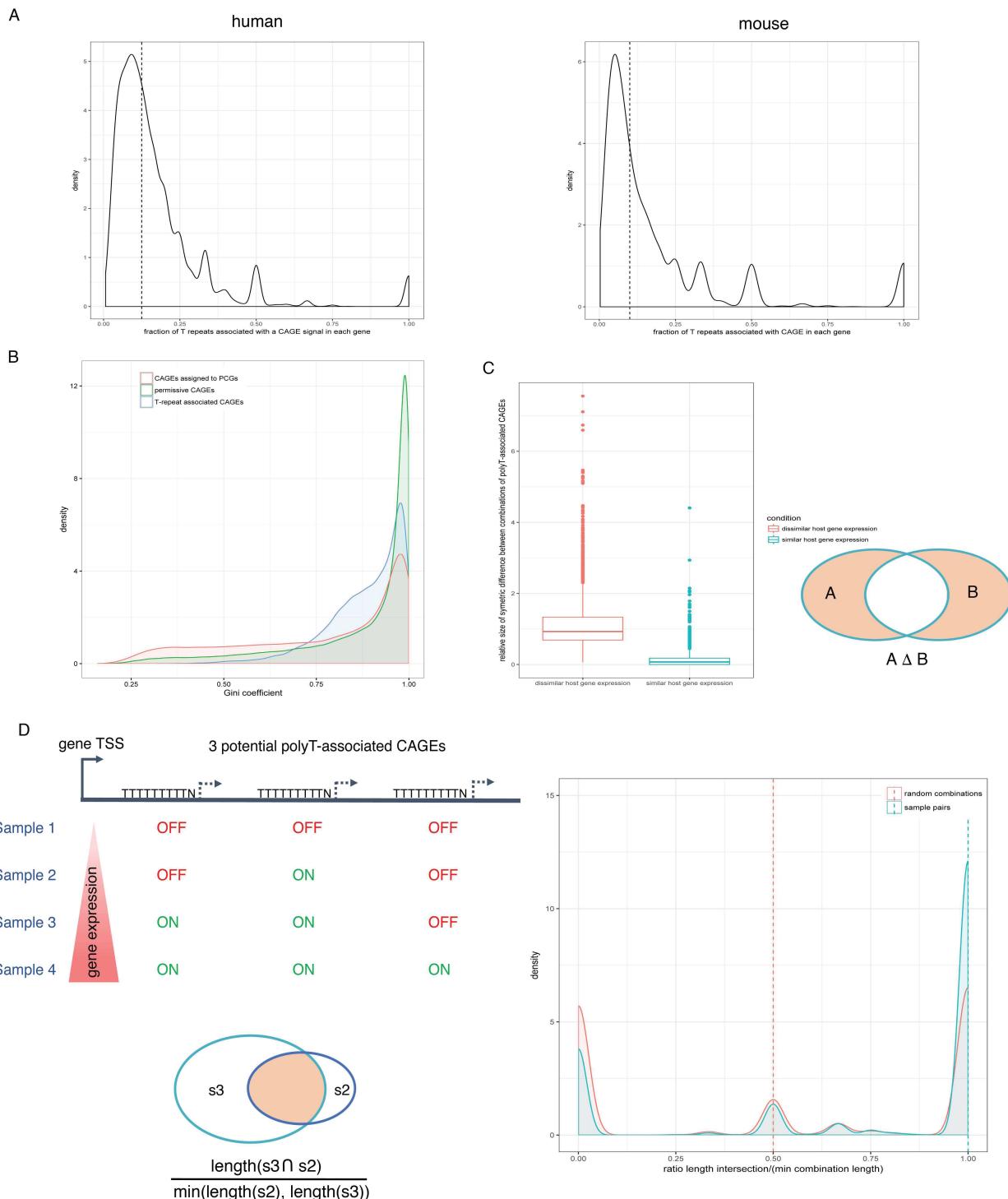
rank	motif	p-value	% in foreground	% in background
1		1e-19871	53.17	19.13
2		1e-9244	5.20	0.15
3		1e-6244	7.45	0.95
4		1e-405	0.26	0.01
5		1e-70	0.03	0.00
-10	position relative to CAGE peakmax			

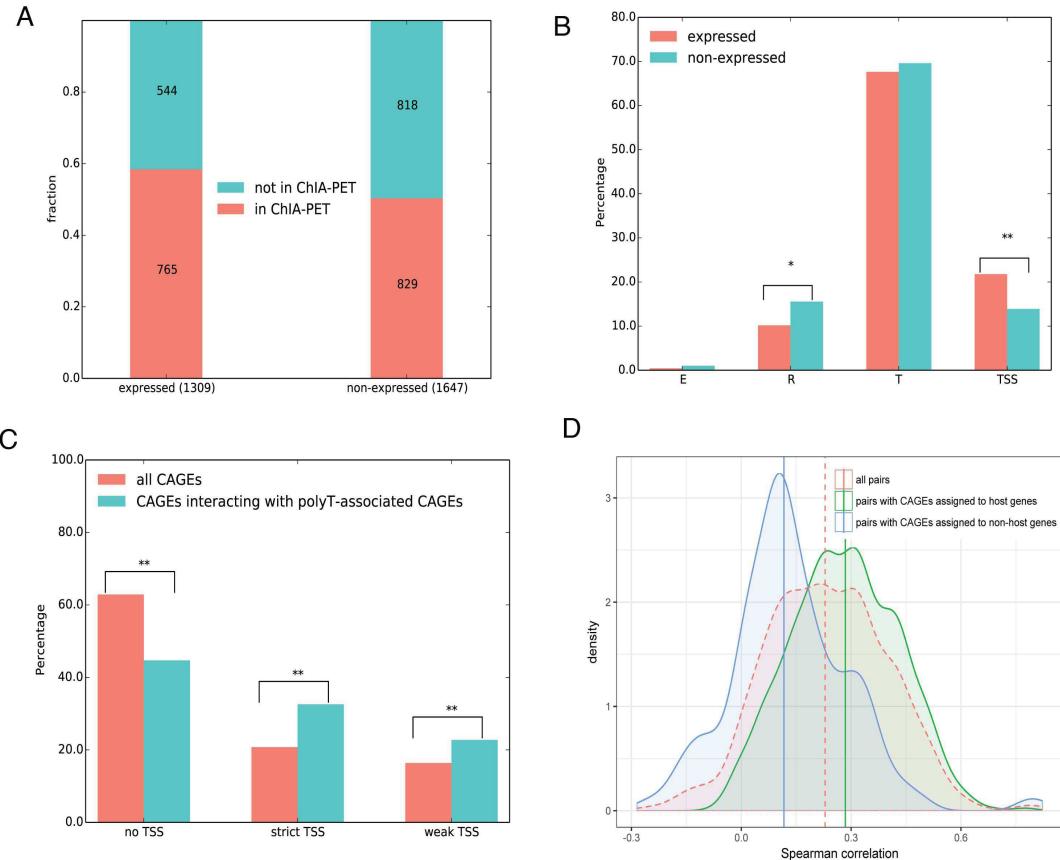


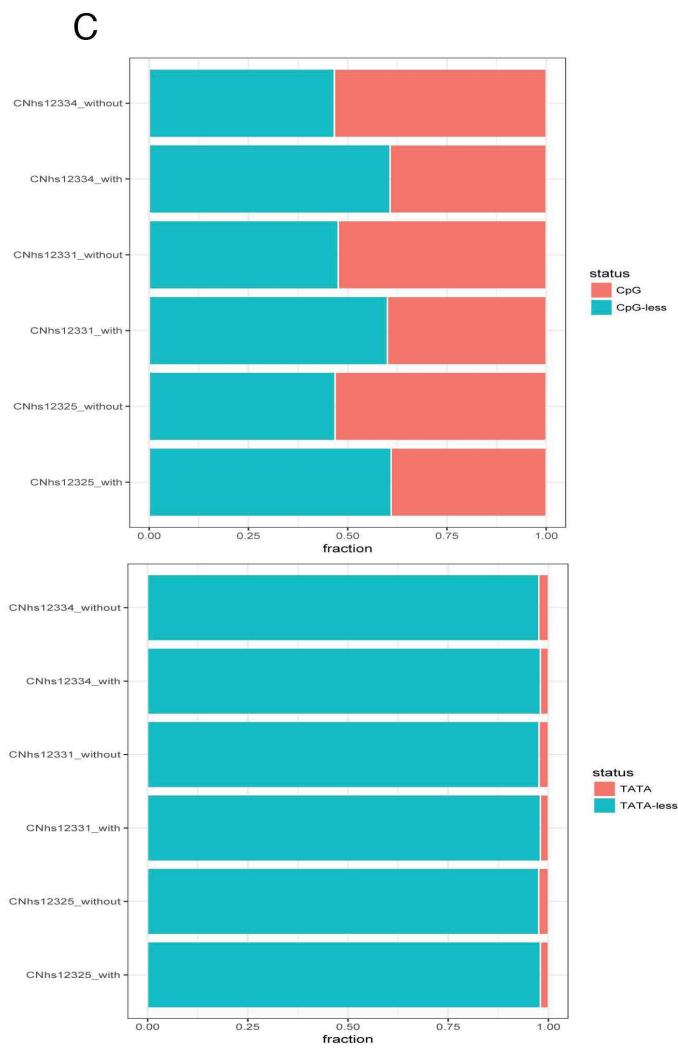
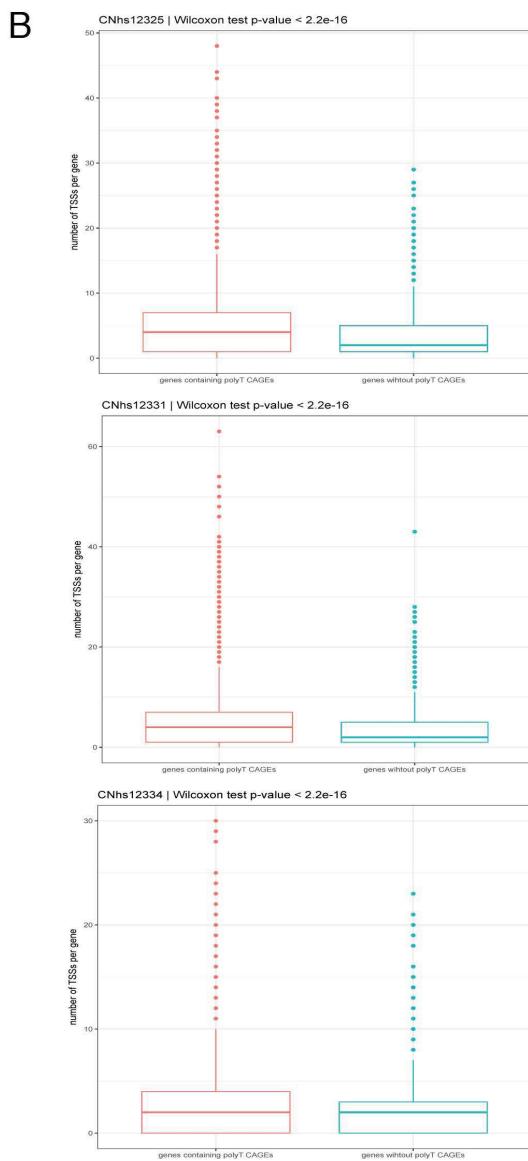
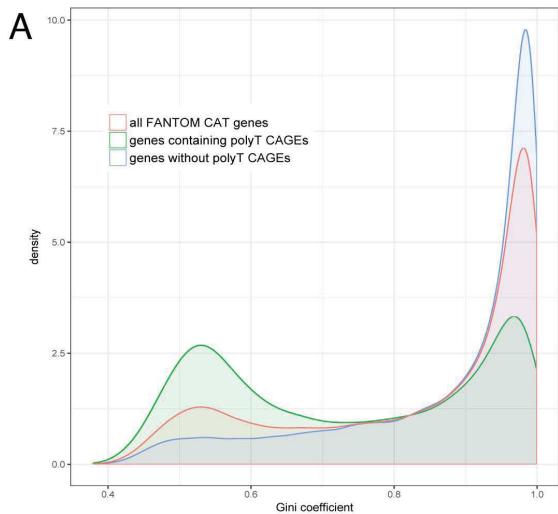












## 2.3 Projet annexe : modélisation de la vitesse d’elongation de l’ARN polymérase II

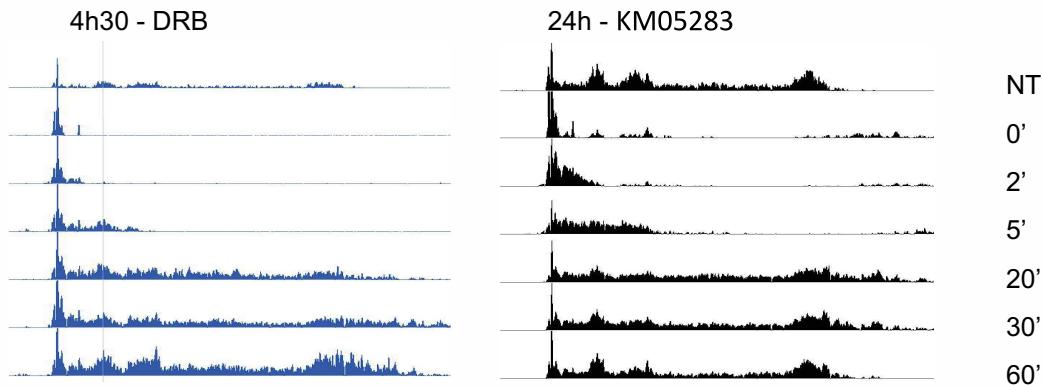
Comme nous l’avons vu dans le premier chapitre, les gènes codants pour des protéines et une grande partie des gènes non-codants sont transcrits grâce à l’ARN Pol II. L’étape d’elongation de la transcription, considérée pendant des années comme une étape triviale [244], est hautement régulée et elle est coordonnée avec d’autres processus co- ou post-transcriptionnels comme l’ajout de la coiffe en 5’ de l’ARN, l’épissage ou l’exportation de l’ARN, via des interactions avec le complexe d’elongation. Cette étape est donc complexe et il est difficile de déterminer les facteurs influençant la vitesse d’elongation et d’évaluer leur contribution. En collaboration avec JC. Andrau (IGMM) et Anne Coleno (IGMM), nous proposons d’établir un modèle pour prédire le taux d’elongation de l’ARN Pol II à partir de différentes variables comme la composition nucléotidique ou les profils de facteurs de transcription et marques épigénétiques. Ce projet est en cours.

### 2.3.1 Méthode expérimentale

Pour mesurer le taux d’elongation de l’ARN Pol II, il faut quantifier la distance parcourue par cette dernière sur un intervalle de temps connu, lorsqu’elle est en cours de transcription. Grâce à la technique d’imagerie de redistribution de fluorescence après photo-blanchiment (*fluorescence recovery after photobleaching*, FRAP), il est aujourd’hui possible de suivre des polymérases en cours d’elongation en *single-cell* (dans une cellule unique) et avec une grande précision. A l’échelle du génome, différentes méthodes peuvent être utilisées : le GRO-seq, le BruDRB-seq qui est proche du GRO-seq mais utilise des nucléotides marqués au bromouridine pour isoler ensuite les ARNs marqués des autres, ou le ChIP-seq (voir partie 1.3.3).

Les données expérimentales analysées ici pour mesurer le taux d’elongation de la polymérase II à l’échelle du génome ont été générées par JC. Andrau et Anne Coleno. La technique utilisée est le ChIP-seq contre la Pol II couplée à l’action d’une drogue qui bloque la Pol II, et donc la transcription, pendant une durée de 4h30 ou 24h selon l’expérience. La drogue utilisée pour le traitement de 24h (DRB) permet de bloquer la protéine Cdk9 et empêche ainsi la phosphorylation de la sérine 2 de l’ARN Pol II qui est nécessaire pour libérer l’ARN Pol II de son état de pause et déclencher l’elongation. Une drogue similaire est utilisée pour le traitement de 4h30 (KM05283). Après 4h30/24h de traitement, les polymérases qui étaient en cours d’elongation se sont décrochées de la matrice ADN et les ”nouvelles” polymérases se sont accumulées

en pause au niveau du promoteur. Les cellules sont ensuite lavées pour retirer l'effet de la drogue et les ARNs Pol II qui s'étaient accumulées commencent la transcription du gène. Elles sont ensuite suivies sur une cinétique d'1h avec 6 points temps pour lesquels les analyses de ChIP-seq sont effectuées. La vitesse de l'ARN Pol II pourra ainsi être déterminée entre ces différents temps, et ce, pour chaque gène.



**FIGURE 2.3 – Profils de ChIP-seq de l’ARN Polymérase II après traitement de 4h30 (gauche) et 24h (droite) sur un gène exemple (gène CD55, 39,5 kb).** Les profils sont représentés pour l’échantillon non traité (profil le plus haut) et aux temps successifs  $t$ .  $t = 0'$  correspond à l’étape d’inhibition de la drogue et donc au relâchement des ARNs Pol II qui étaient bloquées en pause.

Pour chaque temps de la cinétique, le signal de l’ARN Pol II est représenté avec une résolution de 50 pb. Pour le gène exemple représenté sur la Figure 2.3, on observe les profils de ChIP-seq obtenus aux différents temps et on visualise facilement le front de polymérase qui se déplace le long du gène.

### 2.3.2 Méthode computationnelle de détection des fronts

Avec les données de ChIP-seq obtenues, les positions des fronts de migration de l’ARN Pol II aux différents temps et pour chaque gène sont estimés de façon automatique via un algorithme de programmation dynamique développé au sein de l’équipe (L. Bréhélin). La détermination des fronts est vue ici comme un problème d’optimisation où l’on cherche à déterminer la fin de la vague de polymérases avançant sur le gène, à chacun des temps.

Le principe est présenté sur la Figure 2.4 et détaillé ci-dessous :

- Un tableau  $T$  contenant  $K$  lignes et  $N$  colonnes représente le signal d’ARN Pol II extrait des profils de Pol II aux différents temps  $k$  et à chaque position  $n$ .

- Le tableau  $T'$  est construit itérativement de telle sorte que la case  $T'[k, n]$  contienne la somme cumulée maximale que l'on obtient lorsque la position  $n$  du gène fait partie du profil de signal de Pol II au temps  $k$ , soit  $T'[k, n] = \max(T'[k - 1, n - 1], T'[k, n - 1]) + T[k, n]$  avec  $T[k, n]$  le signal d'ARN Pol II mesuré à la position  $n$  et au temps  $k$  (valeur présente dans le tableau initial  $T$ ).
- Dans la colonne 1, seule la cellule  $T'[1, 1]$  est non nulle car cette position ne peut appartenir qu'à la vague de polymérase du premier temps.
- Le tableau  $T'$  est rempli itérativement de haut en bas (du profil au premier temps jusqu'au dernier temps) et de la gauche vers la droite (du premier au dernier bin/position).

Une fois le tableau  $T'$  rempli, il y a une dernière étape de *backpropagation* qui permet d'identifier la segmentation donnant le cumul du signal maximal. A la position  $N$ , la valeur max est  $T[N, K]$ . Ensuite à la position  $N - 1$ , cela peut être  $T[N - 1, K]$  ou  $T[N - 1, K - 1]$ . On remonte ainsi jusqu'à atteindre la position  $[1, 1]$ .

Quand les coordonnées des fronts sont déterminées, nous pouvons calculer simplement le taux d'elongation entre deux temps comme le rapport entre la distance parcourue par les polymérases et le temps écoulé. La vitesse calculée correspond plus précisément à celle des premières polymérases qui ont commencé à transcrire après l'inhibition de l'effet de la drogue et qui sont situées sur le devant de la "vague" de polymérases.

**Définition du site de départ de la transcription.** Pour que l'algorithme de détection des fronts puisse fonctionner, il faut lui fournir le signal de ChIP-seq obtenu aux différents temps mais aussi les coordonnées des gènes d'intérêts qui vont définir les limites au-delà desquelles le signal observé ne sera plus pris en compte. Pour la fin du gène, nous choisissons de prendre les annotations GENCODE. Par contre, pour le début du gène, le signal est généralement très fort et ne se trouve pas systématiquement au niveau du début des gènes annotés dans GENCODE v19. Comme les gènes possèdent généralement plusieurs TSSs, pour déterminer celui étant le plus adapté, nous choisissons, pour chaque gène, de définir une fenêtre d'1 kb autour de chaque TSS annoté dans GENCODE. Nous calculons ensuite le signal total de ChIP-seq d'ARN Pol II à  $t = 0$  sur cette fenêtre. La fenêtre retenue est celle pour laquelle le signal est maximal et le début du gène est donc ici au niveau du pic maximal trouvé dans la fenêtre choisie.

		Positions en bins de 50 bases (n)										
1/ Signal matrix  <i>T</i> table	Profils aux différents temps (k)  position [k,n]	1	2	3	4	5	6	7	8	9	10	11
		t = 0	9	10	5	3	1	0	1	0	0	0
		t = 2	x	5	6	7	6	2	0	1	0	0
		t = 5	x	x	5	4	5	4	3	2	0	0
		t = 20	x	x	x	3	2	2	2	4	4	0
		t = 30	x	x	x	x	2	2	1	2	3	3
2/ Cumulative maximal sum  <i>T'</i> table		1	2	3	4	5	6	7	8	9	10	11
		t = 0	9	19	24	27	28	28	29	29	29	29
		t = 2	x	14	25	32	38	40	40	41	41	41
		t = 5	x	x	19	29	37	42	45	47	47	47
		t = 20	x	x	x	22	31	39	44	49	53	53
		t = 30	x	x	x	x	24	33	40	46	52	55
3/ Backpropagation step		1	2	3	4	5	6	7	8	9	10	11
		t = 0	9	19	24	27	28	28	29	29	29	29
		t = 2	x	14	25	32	38	40	40	41	41	41
		t = 5	x	x	19	29	37	42	45	47	47	47
		t = 20	x	x	x	22	31	39	44	49	53	54
		t = 30	x	x	x	x	24	33	40	46	52	55

FIGURE 2.4 – Principe de l’algorithme de programmation dynamique de détection des fronts d’ARN Pol II optimaux, expliqué sur un exemple fictif. Pour plus d’informations, voir texte.

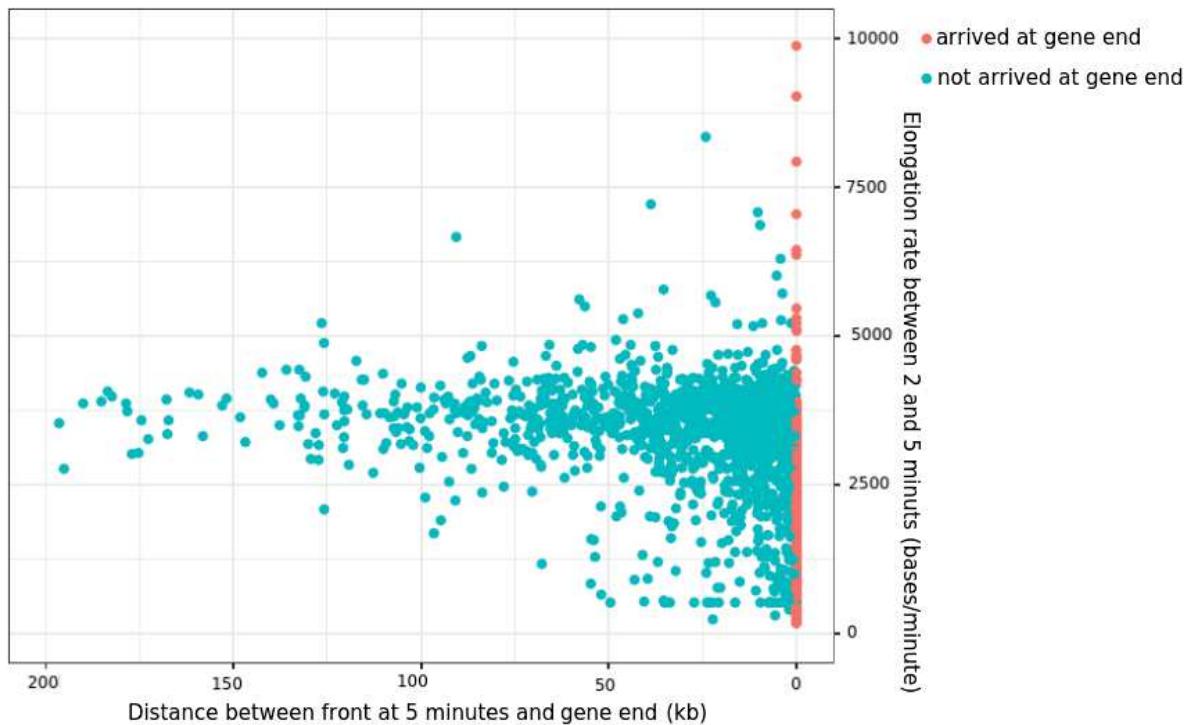
### 2.3.3 Modèle de prédiction du taux d’elongation

Nous nous intéressons ici aux mécanismes de régulation de la vitesse de l’ARN Pol II, et sur le même schéma que le modèle de prédiction de l’expression des gènes présenté dans la partie 2.1.2, nous utilisons un modèle linéaire avec sélection de variables (LASSO) pour sélectionner les facteurs les plus importants dans l’ajustement de la vitesse de l’ARN Pol II. Nous testons l’effet de la composition nucléotidique sur les différentes portions du gènes parcourues par l’ARN Pol II, et notamment entre le cœur et la fin du gène. Nous intégrons également des facteurs comme la proportion intron/exon ou la présence d’îlots CpG qui ont déjà été notés comme ayant une influence sur la vitesse d’elongation [270].

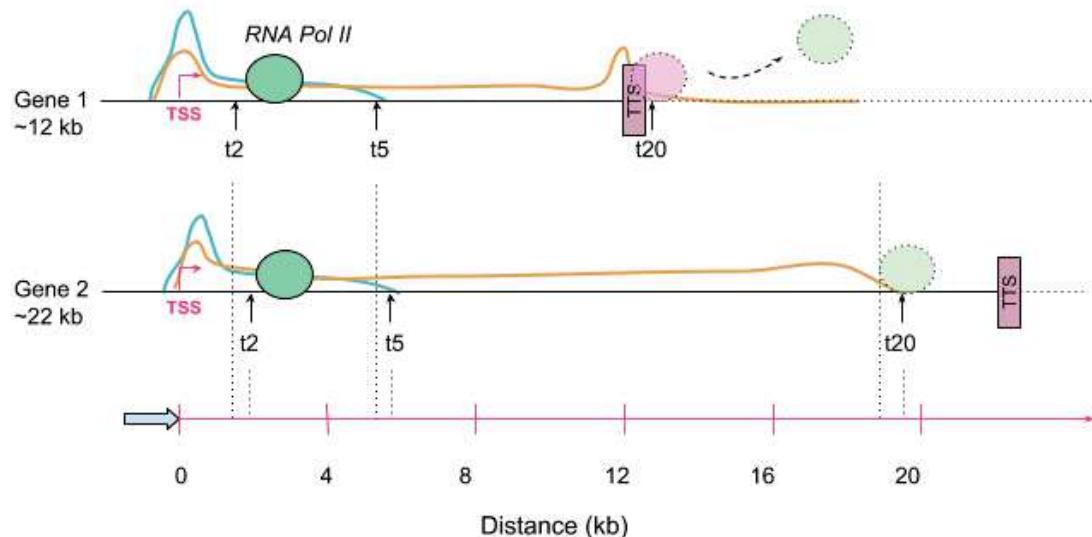
### 2.3.4 Résultats

#### Taux d'elongation de l'ARN Pol II et longueur du gène

La longueur des gènes ayant auparavant été observée comme corrélée à la vitesse d'elongation de la polymérase [270], nous avons tout d'abord étudié le poids de cette variable sur nos données. Au premier abord, le taux d'elongation semble très fortement corrélé à la longueur du gène, avec une corrélation de Spearman entre la longueur des gènes et le taux d'elongation à 5 minutes de 65% et à 20 minutes de 87%. Cependant, en représentant le taux d'elongation de l'ARN Pol II en fonction de la distance à la fin du gène (voir Figure 2.5), nous pouvons directement observer qu'une partie des points (en rose sur la Figure 2.5) correspondent à des polymérases arrivées à la fin de la transcription du gène, pendant l'intervalle de temps considéré. Nous ne pouvons donc pas calculer leur vitesse. L'exemple fictif présenté sur la Figure 2.6 schématise ce phénomène et montre le biais qu'il génère au niveau de l'estimation du taux d'elongation de l'ARN Pol II.



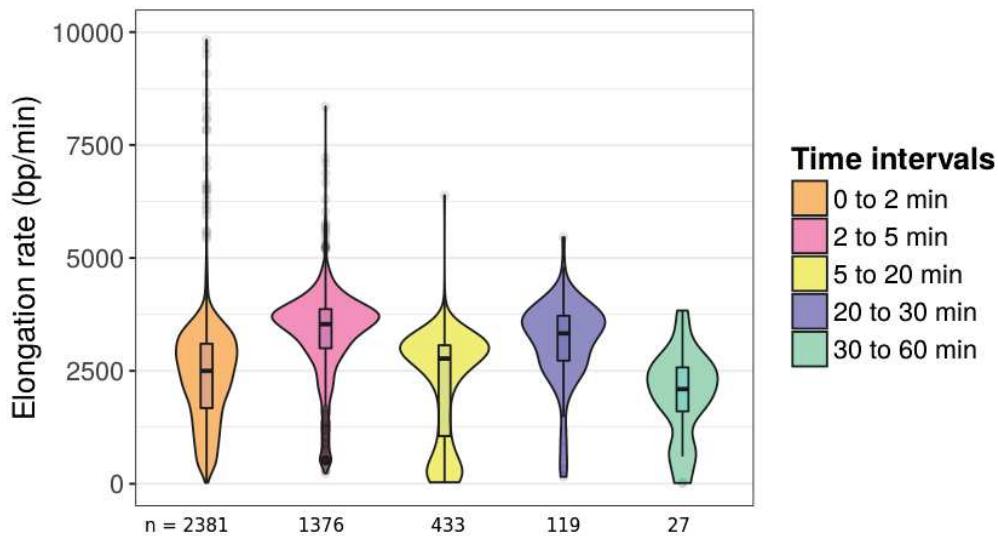
**FIGURE 2.5 – Représentation du taux d'elongation de l'ARN Polymérase II à 5 minutes en fonction de la distance parcourue entre le front à 5 minutes et la fin du gène.**  
Chaque point correspond à un gène, sur un total de 1 918 gènes. Les points roses correspondent à des gènes dont le front à 5 minutes a été détecté dans une fenêtre de 2 bins de 50 pb autour de la fin du gène.



**FIGURE 2.6 – Représentation schématique des fronts de polymérases à 5 et 20 minutes pour deux gènes exemples.** Le premier gène hypothétique est d'une longueur de 12 kb et le deuxième d'une longueur de 22 kb. À 5 minutes, le profil de densité d'ARN Pol II est représenté en bleu et la position du front est notée par "t5". De même, à 20 minutes, le profil de densité est représenté en orange avec la position du front notée "t20". Pour le gène 1, les premières ARNs Pol II ont terminé la transcription du gène à un instant  $t_{rel}$  inconnu, compris dans l'intervalle de temps 5-20 minutes. Le temps utilisé pour le calcul du taux d'élongation est cependant de 15 minutes, entraînant sa sous-estimation. Pour le gène 2, la fin du gène n'est pas encore atteinte à 20 minutes et les premières ARNs Pol II sont, à priori, encore en cours d'élongation. Le taux d'élongation est correctement estimé.

Sur la Figure 2.6, deux gènes sont représentés. Pour le gène 2 (longueur = 22 kb), le front à 20 minutes est détecté à l'intérieur du gène. Les ARNs Pol II en tête sont en cours de transcription et la vitesse, soit le rapport entre la distance parcourue et l'étendue de l'intervalle temps, est correctement estimée. Par contre, pour le gène 1 (longueur = 12 kb), le front à 20 minutes est détecté au niveau de la fin du gène. Dans ce cas de figure, les ARNs Pol II en tête ont terminé de transcrire le gène, mais on ne peut pas savoir à quel moment. Ainsi, dans l'intervalle de 5 à 20 minutes montré en exemple, il s'est écoulé 15 minutes, et il est impossible de savoir à quel temps les polymérases de tête sont arrivées. La vitesse est donc sous-estimée. Pour ne pas biaiser notre jeu de données, nous retirons ces points de l'étude, ainsi que les points situés dans une régions de 2 bins (100 pb) autour de n'importe quel site de terminaison annoté dans GENCODE v19. Les distributions des taux d'élongation des données restantes ainsi que leurs effectifs à chaque intervalle de temps sont représentées sur la Figure 2.7. Avec les gènes restant aux différents temps, la corrélation entre taux d'élongation et longueur des gènes à 5 minutes chute à 35%. Par contre la corrélation à 20 minutes ne diminue que légèrement (corrélation de

62%), effet probablement lié à l'intervalle de temps trop long.



**FIGURE 2.7 – Distribution des taux d’elongation des ARNs Pol II aux différents intervalles de temps.** Les effectifs des gènes que l’on suit encore aux différents temps sont indiqués sous les violins plot.

### Résultats du modèle Lasso pour prédire le taux d’elongation de l’ARN Pol II à partir de différents types de variables

Les modèles de prédiction que nous allons présenter ci-dessous portent sur le jeu de données avec traitement de 24h et sur l’intervalle de temps 2 à 5 minutes, qui comporte un nombre suffisant d’observations ( $n=1376$ ). L’intervalle 0 à 2 minutes a volontairement été exclu des analyses car c’est un temps relativement court pendant lequel l’effet de la drogue est levé et les ARNs Pol II démarrent l’étape de transcription, ce qui peut entraîner plus de fluctuations liées à la méthode expérimentale que sur les autres intervalles.

**Influence de la composition nucléotidique sur le taux d’elongation de l’ARN Pol II.** Comme nous l’avons vu dans la partie 2.1, la composition nucléotidique influence fortement l’expression des gènes. Nous testons ici son effet sur la vitesse d’elongation de l’ARN Pol II. Dans un premier temps, nous prenons comme variables prédictives les taux des 4 nucléotides seulement, calculés sur les régions parcourues pendant l’intervalle de temps considéré, ici 2 à 5 minutes. Un aperçu de la matrice de données utilisée pour le modèle est présenté sur la Figure 2.8.

Le modèle construit à partir des compositions nucléotidiques nous permet d’obtenir une corrélation de Spearman de 47,8% entre le vecteur des taux d’elongation

Gene names	ElongRates_24h	A	C	G	T
ENSG00000001497 2_5	4083.33	0.2320816	0.2429388	0.2589388	0.2659592
ENSG00000003137 2_5	1016.67	0.1872131	0.2783607	0.3229508	0.2111475
ENSG00000004660 2_5	1016.67	0.1993443	0.2865574	0.2898361	0.2239344
ENSG00000004766 2_5	3700.00	0.3071171	0.1509009	0.1862162	0.3556757
ENSG00000005007 2_5	4150.00	0.2054618	0.2375100	0.3043373	0.2526104
ENSG00000005893 2_5	4266.67	0.2956250	0.1825781	0.2114844	0.3102344
:	:	:	:	:	:

Predicted variable

Predictive variables

FIGURE 2.8 – Matrice de données des taux d’elongation de la polymérase II. Aperçu des premières ligne de la matrice de données avec la variable prédictive (taux d’elongation de l’ARN Pol II) et les variables prédictives (taux en nucléotides sur la région parcourue entre 2 et 5 minutes).

observés et le vecteur des taux prédits par le modèle. L’information contenue dans la séquence ADN permet, à elle seule, d’expliquer environ 50% des variations observées lors de l’étape d’elongation de la transcription par l’ARN Pol II. Les compositions en dinucléotides et tri-nucléotides ont également été testées, mais elles ne permettent pas d’améliorer les performances du modèle.

**Variables basées sur des données expérimentales.** De nombreuses variables basées sur des données expérimentales ont été intégrées au modèle de prédiction. Il faut cependant noter que, contrairement aux taux en nucléotides qui peuvent facilement être estimés sur des séquences d’une longueur raisonnable, soit environ 100 pb, il est difficile de le faire pour des variables peu fréquentes au sein du gène, comme par exemple les modifications d’histones qui vont plutôt être présentes à une/des positions spécifiques. Toutes les variables listées ci-dessous sont donc calculées comme des densités sur la totalité du gène, correspondant à la proportion du gène intersectant avec la variable d’intérêt :

1. Segmentation du génome par ChromHMM/Segway (voir partie 1.2). Pour chaque gène on a donc 7 valeurs de variables qui correspondent à la proportion du gène intersectant avec les 7 prédictions d’états chromatiniens.
2. Proportion de régions exoniques en prenant les coordonnées des exons de tous les transcrits annotés dans GENCODE v19.
3. Proportion de petits éléments répétés (SINEs) et longs éléments répétés (LINEs). Pour chacun de ses deux éléments, nous avons deux variables : une correspond à la proportion sens et l’autre à la proportion antisens.
4. Proportion des régions du gène en interaction en utilisant les données de ChIA-PET dans HeLa-S3 du projet ENCODE.
5. Proportion d’îlots CpG en prenant les données de UCSC dans HeLa-S3 (hg19).

6. "Broad domains" correspondant au top 5% des plus larges pics de la marque d'histone caractéristique des promoteurs actifs (H3K4me3).
7. Proportion de pré-microRNA avec les annotations de miRBase 21 (1881 pré-mir prédicts [142]).
8. Enfin, nous avons intégré des variables relatives à la présence de poly-T (Présentés dans le chapitre 2.2), notamment la proportion de poly-T associés à un signal de CAGE et la présence de répétitions de Ts.

L'intégration de ces variables prédictives dans le modèle de base avec les 4 nucléotides n'améliore que très peu les performances. On obtient une corrélation de 50% avec toujours une contribution plus forte des nucléotides. Toutefois, d'autres variables sont sélectionnées en plus des taux en C et T, et on retrouve dans les 5 premières variables sélectionnées, la proportion de régions prédictes comme transcrrites et prédictes comme promoteurs ainsi que la densité exonique.

**Cas des marques épigénétiques.** Les modèles construits à partir des variables présentées dans le paragraphe précédent n'étant pas très performants, nous avons testé un modèle construit sur les données de marques épigénétiques du projet ENCODE [35] pour HeLa-S3. Nous avons ainsi récupéré les fichiers "narrowPeaks" des 11 modifications d'histones disponibles et sommé le signal de ChIP-seq de chacune des marques. Comme pour les compositions nucléotidiques, les variables sont calculées sur les régions parcourues pendant l'intervalle de temps considéré, et sont donc représentées par une densité de signal spécifique à chaque intervalle de temps. Les marques épigénétiques considérées sont : H2AFZ, H3K4me1, H3K4me2, H3K4me3, H4K20me1, H3K36me3, H3K27me3, H3K79me2, H3K9ac, H3K9me3 et H3K27ac. Le modèle intègre donc les 4 nucléotides ainsi que les 11 marques épigénétiques, mais les performances ne sont, encore une fois, pas beaucoup plus élevées qu'avec le modèle intégrant les nucléotides seulement (corrélation de 49%) bien que le modèle sélectionne dans les premières variables, après les nucléotides C et T, des marques activatrices associées à l'activité des gènes : H4K20me1 et H3K79me2 [67].

Un modèle global a également été testé intégrant à la fois les nucléotides, les variables expérimentales, y compris les marques épigénétiques, améliorant sensiblement le modèle (corrélation de 51%).

### 2.3.5 Limites de la méthode expérimentale et du modèle

Les données de ChIP-seq Pol II générées ici sur une cinétique d'une heure nous permettent de suivre une quantité correcte de gènes si l'on se compare aux études publiées ces dernières années [270, 120, 48], soit 1376 jusqu'à 5 minutes et 433 jusqu'à 20 minutes. Cependant, la technique utilisée présente quelques limites qui restreignent nos jeux de données et ne nous permettent pas d'obtenir des résultats complètement satisfaisants :

1. Les données de ChIP-seq ne nous permettent pas d'avoir l'information du brin d'origine (brin - ou brin +), nous devons donc nous restreindre aux gènes qui ne chevauchent aucun autre gène annoté.
2. Les temps de prélèvement des échantillons, sauf pour les deux premiers temps, sont relativement espacés et l'espacement n'est pas régulier. Il est donc difficile de suivre réellement des gènes sur plus de 2 temps (voir les effectifs sur la Figure 2.7 qui chutent rapidement).
3. Les données de ChIP-seq sont connues pour être dépendantes de nombreux facteurs, comme la composition en GC [257] ou l'ouverture de la chromatine. La détection du signal, et donc des fronts, est favorisée dans les régions de chromatine ouverte et pour les gènes fortement exprimés.
4. Les analyses effectuées manquent de réplicats. Les deux conditions étudiées n'ont pas exactement les mêmes paramètres. En effet, deux drogues différentes sont utilisées, bien que leur mode d'action soit similaire, et les temps de traitement ne sont pas les mêmes (4h30 et 24h).

L'algorithme de détection des fronts ayant été vérifié manuellement sur quelques exemples, il semble être adapté à nos données. Toutefois, il faudrait comparer la méthode avec d'autres plus classiques comme l'utilisation de modèles de Markov cachés et les tester en parallèle, et sur différents jeux de données. Cela nous permettrait de nous positionner et de valider l'efficacité de la méthode. Une limite commune aux méthodes existantes est qu'elles ne permettent pas de discriminer le signal global d'ARN Pol II de signaux internes interférant avec l'ARN Pol II, qui pourraient par exemple marquer la présence d'enhancers intragéniques ou de lncRNAs.

Enfin, le modèle lasso utilisé est un modèle linéaire global à tous les gènes. Dans notre cas, il est adapté à la mise en évidence de variables ayant un effet commun à l'ensemble des gènes. Les mécanismes de régulation de la vitesse de l'ARN Pol II spécifiques à seulement certains groupes de gènes ne seront pas capturés ici. De plus, certaines des variables intégrées dans le modèle et testées ont des occurrences

rares et sont représentées dans la matrices par une variable vectorielle contenant beaucoup de valeurs nulles. Leur impact potentiel sur la vitesse de la Pol II est dans ce cas difficile à capturer.

### 2.3.6 Travaux en cours : modélisation du taux d'elongation d'un point de vue biophysique

D'un point de vue biophysique, les polymérases en cours de transcription peuvent être vues comme des particules se déplaçant sur une ligne, la chaîne au temps initial étant vide. La modélisation de ce déplacement est alors possible et en collaboration avec une équipe de physiciens (F. Geniet, L2C-UM), à partir des données de ChIP-seq d'ARN Pol II, nous voulons modéliser et calculer la vitesse de déplacement de la polymérase le long des gènes. Nous utilisons pour cela une approche par une équation de continuité qui décrit le principe de conservation de la masse. Cette approche est soumise à quelques hypothèses permettant de simplifier le problème. La première hypothèse est que la transcription se déroule d'un point A à un point B sans décrochage ou fixation de polymérase, c'est à dire qu'une fois que la polymérase commence le processus de transcription, elle continue jusqu'à la fin du transcrit où elle se désassemble. Une autre hypothèse importante pour pouvoir calculer une vitesse instantanée est celle d'une synchronisation de l'ensemble des cellules qui sont prélevées pour estimer la quantité de polymérases via le ChIP-seq et on assume aussi que tous les échantillons prélevés aux différents temps sont similaires et homogènes. Le modèle est décrit par les équations ci-dessous, avec  $\rho(x, t)$  correspondant au profil de densité de l'ARN Pol II à l'instant  $t$  et en partant de l'hypothèse de départ d'une chaîne vide, soit  $\rho(x, t = 0) = 0$ .

$$\frac{\delta \rho(x, t)}{\delta t} + \frac{\delta j(x, t)}{\delta x} = 0 \quad (2.1)$$

Les particules apparaissant dans l'intervalle  $[x, x + \delta x]$  entre les temps  $t$  et  $t + \delta t$  correspondent à la différence des flux entrant  $j(x, t)$  et sortant  $j(x + \delta x, t)$ . Les particules se déplacent à une vitesse  $v(x, t)$ . Nous avons la relation suivante :

$$j(x, t) = \rho(x, t) * v(x, t) \quad (2.2)$$

En intégrant l'équation 2.1, nous obtenons la relation suivante :

$$j(x, t) - j(L, t) = \int_x^L \frac{\delta \rho(y, t)}{\delta t} dy \quad (2.3)$$

Nous considérons seulement les temps  $t$  pour lesquels la fin de la chaîne, en position  $L$ , n'est pas atteinte et donc  $j(L, t) = 0$  :

$$j(x, t) = \int_x^L \frac{\delta\rho(y, t)}{\delta t} dy \quad (2.4)$$

L'intérêt de ce modèle est de pouvoir calculer des vitesses instantanées, c'est à dire, obtenir un profil de vitesse sur toute la cinétique. Dans un cas idéal, nous avons accès aux données de ChIP-seq de Pol II à tous les temps avec un  $\delta t$  très petit. En réalité, les expériences sont effectuées à des temps finis, le modèle est donc appliqué à des  $\Delta t$  très grands. Cette limite entraîne une estimation faussée de la vitesse avec une décroissance sur l'intervalle  $\Delta t$  qui est en fait un artefact. En l'état, le modèle ne peut donc pas être appliqué et nécessite d'être ajusté. Un stage de Master est prévu pour poursuivre les travaux de modélisation biophysique de l'élongation de l'ARN Pol II.

# Chapitre 3

## Discussion et perspectives

Depuis la publication du premier génome humain en 2001, les différentes annotations de gènes se multiplient et ne cessent d'évoluer. Le nombre de gènes codants pour des protéines estimé autrefois à plus de 30 000 [37], tourne aujourd'hui autour de 20 000 gènes. Avec le développement des technologies de séquençage du transcriptome, il a été montré que la partie fonctionnelle du génome ne s'arrête pas aux gènes codants pour des protéines et que la majorité du génome humain est transcrise [124]. C'est ce qu'on appelle la transcription "pervasive". L'émergence de nouvelles classes de gènes codants pour de petits ou longs ARNs non-codants remet en question la définition de gène autrefois réservée aux portions du génome donnant naissance à des protéines [208]. Bien que cette définition soit encore discutée, elle fait aujourd'hui souvent référence à toutes les portions du génome transcris en une molécule d'ARN fonctionnelle. Ici encore, la notion de "fonctionnelle" peut être discutée [35].

Pendant ma thèse, je me suis intéressée à la caractérisation et l'annotation de la partie non codante du génome humain, afin de montrer son importance dans la régulation de l'expression des gènes et de contribuer à la caractérisation du rôle des transcrits peu exprimés, qui sont souvent filtrés des études actuelles et considérés comme du bruit ou ADN poubelle (*junk DNA*). Mes principales contributions portent sur (i) le développement d'un modèle de prédiction de l'expression des gènes à partir de la séquence, montrant notamment l'implication des régions introniques dans cette régulation, (ii) la caractérisation d'un motif poly-T particulier marquant le départ d'un signal de transcription et qui est associé à une classe de lncRNAs sens-introniques qui semblent fonctionner comme des enhancers agissant sur leurs gènes hôtes.

**La séquence ADN, et notamment la composition des introns, contient de l'information capable d'expliquer l'expression des gènes**

A travers le projet de prédiction de l'expression des gènes uniquement à partir de l'information contenue dans la séquence ADN (voir partie 2.1), nous avons pu tout d'abord montrer que les compositions nucléotidiques basiques (taux en nucléotides et di-nucléotides dans différentes régions) nous permettent d'obtenir une corrélation de Spearman médiane de 60% entre l'expression des gènes observée et prédictive. Étendre ces compositions à des k-mers de taille plus importante comme les tri-nucléotides ne nous permet d'améliorer que modérément les performances du modèle, suggérant l'existence d'un niveau de régulation basé sur la composition de régions relativement larges d'ADN, idée soutenue notamment par la forte contribution des introns dans le modèle. La régulation de l'expression des gènes capturée pourrait ainsi être en partie liée aux TADs actifs/inactifs selon les conditions, et qui sont caractérisés par des compositions nucléotidiques spécifiques. Nous avons également évalué l'importance des motifs de fixation des facteurs de transcription dans notre modèle, à partir de matrices de probabilité de position (PPM). Les scores de motifs attribués à chaque gène et pour chaque PPM sont calculés en scannant la PPM sur le promoteur et le score maximal obtenu, représentant à priori le site de fixation le plus probable, est celui retenu. Cependant, le calcul des scores adopté ici restreint les analyses à un seul site de fixation par gène et par TF, qui n'est pas forcément le site fixé parmi tous les sites potentiels qui avaient un bon score. L'intégration des motifs dans notre modèle ne nous permet pas d'obtenir d'amélioration des performances, leur signal étant masqué par la forte contribution des compositions nucléotidiques. Ces résultats nous confortent dans l'idée que notre méthode de calcul des scores de motifs génère beaucoup de faux positifs et garder le score maximal n'est sûrement pas la meilleure solution pour capturer de l'information. Le modèle TFCoop publié dans l'équipe (travaux de Jimmy Vandel [269]), est une méthode alternative pour la détection des sites de fixation, prenant en compte la coopération entre les TFs. Le modèle s'appuie sur l'identification de l'ensemble des TFs coopérant avec un TF d'intérêt, prédits à partir de données de ChIP-seq (ENCODE) provenant de différents types cellulaires. Il serait donc intéressant de combiner cette méthode avec notre modèle de prédiction des gènes. L'importance des motifs dans la prédiction de l'expression des gènes est toutefois évaluée par May Taha (doctorante dans l'équipe), par intégration des PPM dans un modèle de réseaux de neurones convolutifs.

Notre modèle construit uniquement à partir de la séquence ADN a été comparé à des modèles basés sur des données expérimentales. Les modèles pour prédire

l’expression des gènes à partir de données de ChIP-seq et d’accessibilité de la chromatine (DNaseI-seq) sont nombreux dans la littérature (voir partie 2.1). Les profils de ChIP-seq sont regroupés dans des catalogues de données publiques et reposent sur des algorithmes de recherche de pics (*Peak calling*) pour détecter les sites de l’ADN sur lesquels les protéines d’intérêt sont fixées, dans une condition donnée. En pratique, cela correspond aux fragments d’ADN couverts par plus de *reads* que ce qui est attendu par hasard. Cependant, il faut être prudent avec l’utilisation de ces données qui peuvent présenter plusieurs types de biais : biais lié à la variabilité de la spécificité et de l’efficacité des anticorps utilisés, biais d’amplification par PCR, biais de fragmentation, profondeur de séquençage, influence de la composition nucléotidique des régions étudiées... De plus, il a été observé que les ChIP-seq des différents TFs, les modifications d’histone et les données de sensibilité à la DNase I partagent des informations, ce qui peut conduire à des redondances dans les modèles prédictifs.

Notre modèle possède des performances comparables à celles de modèles construits à partir de données expérimentales (ChIP-seq, DNaseI-seq). Cependant, nous avons montré par une méthode de permutations aléatoires des variables expérimentales associées à chaque gène et utilisées pour prédire leur expression, qu’il existait un biais très fort lié à l’ouverture de la chromatine. En effet, la permutation aléatoire des variables, basées sur les données expérimentales, par gène, n’affecte pas les prédictions du modèle, reflétant une certaine homogénéité des scores de ChIP-seq associés à chaque gène. Les gènes exprimés, se trouvant majoritairement dans des zones de chromatine ouverte, sont souvent fixés par une quantité de TFs importante, ce qui rend les scores, au sein de ces gènes, échangeables. De même pour les gènes non exprimés qui ont des scores de ChIP-seq faibles/nuls. Au contraire, en appliquant cette méthode à nos données basées sur la séquence ADN, nous observons une chute des performances (voir article partie 2.1.2). L’article de Schmidt et al. récemment publié [232] discute cet aspect et reprend notre procédé de permutation aléatoire des variables afin de proposer des pistes pour corriger le biais lié à l’ouverture de la chromatine. Les solutions proposées intègrent une pénalisation des scores par le nombre de pics de données de ChIP-seq/DNaseI associés à chaque gène, ce qui donne plus de poids aux gènes situés dans des régions de chromatine fermée et à l’inverse minimise les scores des gènes situés en zone de chromatine ouverte. Même si l’utilisation d’un modèle corrigé permet de perdre un peu les performances du modèle permuté, et donc de corriger une petite partie du biais d’ouverture de la chromatine, l’effet reste fort. Toutefois, comme indiqué dans le papier de Schmidt et al., les modèles appris sur les données permutes ne peuvent pas être interprétés et les

combinaisons capturées sont fictives, à l'inverse du modèle appris sur la matrice de départ où les combinaisons de TFs capturées ont du sens. Pour le confirmer, nous proposons une étape complémentaire au contrôle effectué avec la permutation aléatoire des variables par gènes, qui consiste à appliquer le modèle appris sur la matrice permutée, à la matrice de variables non-permutée. Cela signifie que les coefficients  $\beta$  sont déterminés sur la matrice permutée puis transposés sur les vraies variables. Ce dernier modèle est appliqué pour prédire l'expression des gènes. Un test préliminaire sur les données de RACER présentées dans le chapitre 2.1.2 nous montre une chute de la corrélation de 52% en médiane (21 conditions) à 18%, confirmant le fait que les combinatoires capturées sur les données de ChIP-seq ont du sens malgré le biais d'ouverture de la chromatine. Ce contrôle peut être par la suite appliqué à tout modèle linéaire de prédiction pour valider la pertinence du modèle et des combinatoires capturées. Dans le cas des modèles appris sur les données de ChIP-seq, on en conclut que le modèle capture l'ouverture de la chromatine soit par une véritable combinaison de TFs importants, soit simplement par le nombre de TFs fixés sur chaque gène (comme montré par le modèle construit sur des variables permutées ou sur un unique vecteur correspondant au maximum des valeurs de chaque variable).

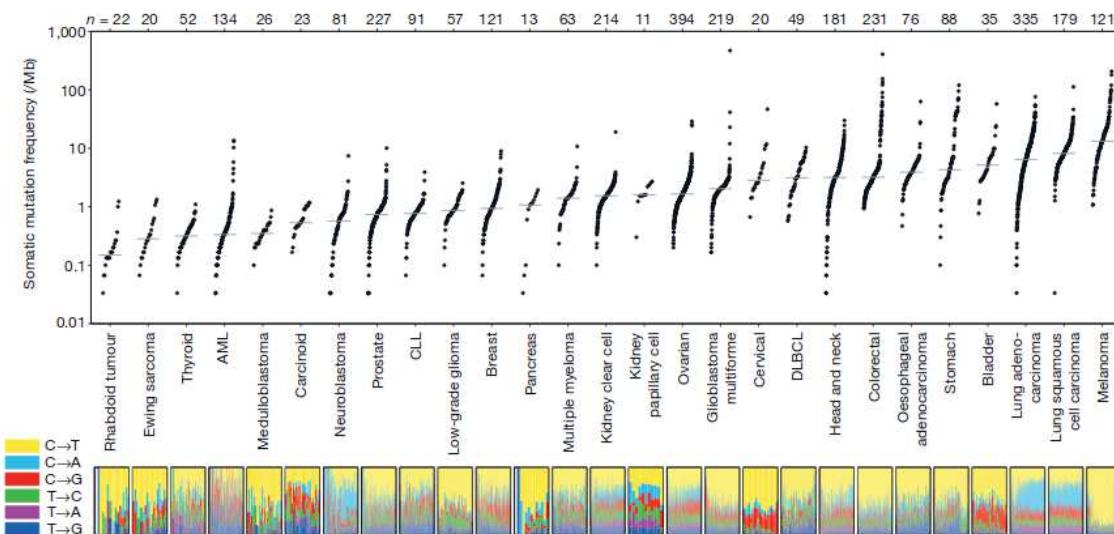
Les données expérimentales ne pouvant pas être générées pour chaque patient ni pour chaque facteur de transcription, la communauté scientifique s'intéresse depuis peu à trouver des combinatoires de variables basées uniquement sur la séquence et permettant d'expliquer l'expression des gènes dans les différents types cellulaires. De manière concomitante à nos travaux et comme nous l'avons vu dans la partie 2.1, plusieurs modèles de réseaux de neurones convolutifs ont été développés pour prédire l'expression des gènes à partir de la séquence. Les résultats présentés nous montrent des performances étonnamment élevées : ExPecto [290] prédit l'expression des gènes transcrits par l'ARN Pol II dans 218 tissus différents avec une corrélation médiane de Spearman entre expression prédite et observée de 0.81, Xpresso [1] explique la variation de l'expression médiane des gènes dans 56 types cellulaires différents avec un coefficient de corrélation  $r^2$  de 0.59 et enfin, Basenji prédit l'expression des gènes quantifiée par CAGE avec une corrélation de Pearson entre expression prédite et observée de 0.86. Ces trois modèles sont basés sur la séquence ADN uniquement, en intégrant des régions autour des TSS des gènes de taille très importante. En effet, les séquences prises en compte dans les modèles sont de l'ordre de 10 kb autour du TSS pour Xpresso à 131 kb pour Basenji, contrairement à notre modèle qui intègre une région promotrice de 2 kb et des régions plus larges (introns) mais spécifiques aux différents gènes. Prendre en compte des régions de taille fixe et importante dans

leur cas permet de capturer l'effet des régions régulatrices distales, comme les enhancers, mais elles peuvent également intersecter avec d'autres gènes. Nous avons vu précédemment que les gènes appartenant à un même TAD pouvaient être co-régulés et que leur expression dépendait de leur environnement nucléotidique, une partie de l'expression des gènes peut donc être ici capturée via l'environnement nucléotidique de ses gènes voisins. Plus généralement, les modèles basés sur des réseaux de neurones convolutifs permettent de capturer des combinaisons de séquences régulatrices complexes, il faut toutefois être prudent sur leur utilisation, comme pour toute méthode de *machine learning*. En effet, il est important de définir les échantillons de test et d'apprentissage pour éviter le sur-apprentissage. De plus, ces modèles de *deep learning* sont limités en terme d'interprétation biologique et extraire les variables ayant un fort impact sur l'expression des gènes n'est pas encore facile, bien que des outils émergent [87].

### **Existe-t-il un lien avec le spectre de mutations observé dans les cancers ?**

Le cancer est souvent nommé comme la maladie du siècle, et sous cette dénomination large, un ensemble hétérogène de maladies se regroupent. Ces maladies diffèrent en fonction du type de cancer, mais également, d'un patient à l'autre. Pour pouvoir adapter les traitements à chaque patient, il est nécessaire d'identifier les particularités de chaque tumeur, c'est à dire les mutations qui sont à l'origine de ces différents cancers. Les chercheurs du Wellcome Trust Sanger Institute (WTSI) ont analysé les spectres de mutations somatiques de plus de 7 000 tumeurs différentes, ce qui leurs a permis de révéler une 20<sup>aine</sup> de signatures différentes reflétant les changements cumulatifs ayant eu lieu pendant le développement de chacun des types de cancers [149, 2] (voir Figure 3.1).

Le développement de notre modèle de prédiction de l'expression des gènes à partir de la séquence et dans les différents cancers (données TCGA) nous permet de mettre en lumière les compositions nucléotidiques importantes, dans les différentes régions définies dans la partie 2.1, qui gouvernent l'expression des gènes. Une des pistes à explorer pour aller plus loin dans l'interprétation biologique de notre modèle de prédiction et dans la compréhension des dérégulations de l'expression des gènes dans les cancers est celle des distributions particulières des mutations observées selon le type de cancer (voir Figure 3.1). Malgré une tendance de notre modèle à mieux prédire les gènes ubiquitaires, il nous a permis d'obtenir des groupes de gènes spécifiques bien prédis et des variables spécifiques à chaque type de cancer et parfois



**FIGURE 3.1 – Mutations somatiques observées dans différents types de cancers.** Les données proviennent d'exomes de 3 083 paires cellules tumorales/cellules normales. Chaque point du graphique correspond à une paire tumeur/normal, avec la position sur l'axe vertical correspondant à la fréquence totale de mutations somatiques dans l'exome en condition tumorale en comparaison à des cellules normales. Les différents types de tumeurs, sur l'axe horizontal, sont ordonnées en fonction de la valeur médiane de fréquence de mutations. La fréquence de mutations varie d'un facteur de plus de 1 000 entre les cancers mais aussi au sein d'un même cancer. Les répartitions des mutations en fonction du type de substitution sont représentées au bas de la figure. [Figure extraite de [149]]

même à chaque patient. A partir de données de RNA-seq et les génotypes associés dans différentes conditions, nous pouvons étudier le lien entre mutations et variables sélectionnées par notre modèle. En effet, de récentes analyses à l'échelle du génome ont permis d'identifier des enrichissements en mutations somatiques dans les promoteurs et sites de fixation des facteurs de transcription fonctionnels [122, 228, 201], renforçant l'intérêt d'évaluer le lien entre les mutations et nos variables sélectionnées. Une piste possible est donc de tester si les variables stables sélectionnées par notre modèle correspondent à celles qui sont les plus mutées dans différentes conditions. Pour répondre à cette question, May Taha a utilisé les données du nombre de mutations décrites dans [149], qui correspondent aux altérations de tri-nucléotides (96 variables) et les expressions de gènes associées (données TCGA pour plus de 1 000 conditions). Les tests d'enrichissement effectués par May ont cependant soulevé un biais : le di-nucléotide CpG est fortement muté comparé aux autres, dans tous les cancers, et il correspond également à une variable stable toujours sélectionnée par notre modèle de prédiction. La méthode nécessite donc un ajustement pour corriger ce biais. De plus, les nombres de mutations décrites dans [149] sont générées à partir des exomes, c'est à dire sur la partie exonique des gènes, et non à l'échelle du génome.

En 2016, Kaiser et al. [122] se sont intéressés aux profils de mutations des TFBSS pour différents types de cancers et à partir de données de *whole genome*. Les régions non-codantes sont ainsi prises en considération. Il serait donc intéressant de regarder les résultats que l'on obtiendrait en prenant en compte les mutations observées sur l'ensemble des régions considérées dans notre modèle, les introns étant les régions ayant la meilleure contribution pour expliquer l'expression des gènes, d'après notre modèle.

### Comparaison avec les différentes classes de gènes non codants

La plupart des contrôles identifiés à ce jour ont été établis sur la catégorie des gènes codants pour des protéines, dont les ARNm ne représentent qu'une petite partie des transcrits générés à partir du génome. Le génome humain est capable de générer une grande variété d'ARNs non-codants dont les régulations demeurent largement inconnues. Une des perspectives au développement du modèle de pré-diction de l'expression des gènes pourrait être son adaptation à d'autres classes de gènes comme les gènes de lncRNA et de microARNs dont des dérégulations peuvent avoir lieu dans les cancers. Cependant, les classes de gènes non-codants n'ont pas la même structure que les gènes codants pour des protéines, il faut donc adapter le modèle en définissant les régions pertinentes à intégrer dans les variables prédictives.

Comprendre les mécanismes de régulation des gènes non-codants est primordial, mais il est aussi important de caractériser leurs fonctions qui sont encore peu connues, et de comprendre comment ils agissent sur la régulation de l'expression des gènes codants pour des protéines. Autrefois considérés comme "junk DNA", ils sont aujourd'hui largement étudiés et leur importance, bien qu'encore remise en cause [194, 155], n'est plus négligée. Cependant, une partie des ARNs non-codants sont moins étudiés que les autres : les ARNs introniques qui représentent pourtant la majeure partie du transcriptome des mammifères [248]. Comme ils sont chevauchants aux ARNs de leurs gènes hôtes et bien moins exprimés que ces derniers, ils sont souvent considérés comme du bruit, provenant d'ARNs immatures qui n'auraient pas été épissés. En 2005, des évidences sont déjà publiées en faveur de la présence de lncRNA introniques antisens et des hypothèses sont faites sur leur rôle dans la régulation fine de l'expression des gènes. L'expression tissus-spécifique des lncRNAs est ensuite étudiée, et il est notamment montré qu'un ensemble de lncRNA introniques sont exprimés dans le pancréas, et leur expression est corrélée au cancer du pancréas, dont les mécanismes d'apparition étaient alors très peu connus [220]. L'environnement de ces lncRNAs est enrichi en marques épigénétiques caractéristiques

des promoteurs soutenant l'idée d'unités transcriptionnelles indépendantes.

Avec l'importance de la séquence intronique mise en évidence dans notre modèle de prédiction de l'expression des gènes et les éléments publiés en faveur d'un rôle régulateur de ces régions, les introns apparaissent comme des réservoirs de nouveaux éléments régulateurs fonctionnels, comme les lncRNA introniques, mais qui sont encore peu étudiés. La découverte d'un motif poly-T associé à un signal de CAGE et localisé dans les introns, à partir des données de CAGE de FANTOM5, vient compléter la découverte des fonctions cachées du génome non-codant. L'association de ces motifs poly-T à des lncRNA et l'hypothèse d'un rôle sur l'expression de leurs gènes hôtes basée sur des évidences présentées dans la partie 1.3.3 nécessite d'être confirmée par des expériences et permettront d'apporter des informations sur la régulation de l'expression des gènes dans différentes conditions, dont l'analyse des exons ne suffit pas à sa compréhension.

La transcription des lncRNA introniques fait partie d'un ensemble de transcription plus large du génome humain dite transcription "pervasive", correspondant majoritairement à une transcription de faible intensité aussi appelée transcription "cachée" [114]. Une des limites importantes à la caractérisation des ARNs non-codants est donc leur faible expression, entraînant des difficultés à distinguer les nouveaux ARNs ayant une vraie fonction biologique des ARNs non-fonctionnels dérivés d'autres transcrits. Ainsi, la découverte des motifs poly-T associés à un signal de CAGE n'exclue pas que l'on détecte de faux positifs et à l'inverse n'exclue pas non plus l'idées que l'on n'a pas détecté tous les vrais signaux, limités par les seuils de détection et la profondeur du séquençage des CAGEs. Les perspectives prochaines de développement de ce projet porteront sur le développement d'un modèle de classification permettant de discriminer les poly-T associés à un signal de CAGE de ceux qui ne le sont pas, à partir de variables pouvant être basées sur la séquence ADN et sur la base d'un modèle de réseaux de neurones convolutifs qui ont déjà fait leurs preuves pour capturer des différences liées à des variations aussi petites que les SNPs [291, 290]. La présence des signaux de CAGEs semble déjà être fortement liée à la longueur de la séquence poly-T et le développement d'un tel modèle nous aidera à annoter et caractériser les motifs poly-T sur l'ensemble du génome et à mieux comprendre l'effet de certains STRs sur l'expression des gènes.



# Bibliographie

- [1] Vikram Agarwal and Jay Shendure. Predicting mrna abundance directly from genomic sequence using deep convolutional neural networks. *bioRxiv*, page 416685, 2018.
- [2] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolo Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463) :415, 2013.
- [3] Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, et al. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493) :455, 2014.
- [4] Christof Angermueller, Heather J Lee, Wolf Reik, and Oliver Stegle. Deepcpg : accurate prediction of single-cell dna methylation states using deep learning. *Genome biology*, 18(1) :67, 2017.
- [5] Francisco Antequera and Adrian Bird. Cpg islands. In *DNA methylation*, pages 169–185. Springer, 1993.
- [6] Jean-Pierre Bachellerie, Jérôme Cavaillé, and Alexander Hüttenhofer. The expanding snorna world. *Biochimie*, 84(8) :775–790, 2002.
- [7] Julian Banerji, Sandro Rusconi, and Walter Schaffner. Expression of a  $\beta$ -globin gene is enhanced by remote sv40 dna sequences. *Cell*, 27(2) :299–308, 1981.
- [8] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo : archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1) :D991–D995, 2012.
- [9] Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4) :823–837, 2007.
- [10] Christine R Beck, José Luis Garcia-Perez, Richard M Badge, and John V Moran. Line-1 elements in structural variation and disease. *Annual review of genomics and human genetics*, 12 :187–215, 2011.
- [11] Michael A Beer and Saeed Tavazoie. Predicting gene expression from sequence. *Cell*, 117(2) :185–198, 2004.
- [12] Victoria P Belancio, Astrid M Roy-Engel, Radhika R Pochampally, and Prescott Deininger. Somatic expression of line-1 elements in human tissues. *Nucleic acids research*, 38(12) :3909–3922, 2010.
- [13] Nicolás Bellora, Domènec Farré, and M Mar Albà. Positional bias of general and tissue-specific regulatory motifs in mouse gene promoters. *BMC genomics*, 8(1) :459, 2007.
- [14] Bradley E Bernstein, Alexander Meissner, and Eric S Lander. The mammalian epigenome. *Cell*, 128(4) :669–681, 2007.
- [15] Andreas Bolzer, Gregor Kreth, Irina Solovei, Daniela Koehler, Kaan Saracoglu, Christine Fauth, Stefan Müller, Roland Eils, Christoph Cremer, Michael R Speicher, et al. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS biology*, 3(5) :e157, 2005.
- [16] Boyan Bonev and Giacomo Cavalli. Organization and function of the 3d genome. *Nature Reviews Genetics*, 17(11) :661, 2016.
- [17] Stefan Bonn, Robert P Zinzen, Charles Girardot, E Hilary Gustafson, Alexis Perez-Gonzalez, Nicolas Delhomme, Yad Ghavi-Helm, Bartek Wilczyński, Andrew Riddell, and Eileen EM Furlong. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature genetics*, 44(2) :148, 2012.
- [18] Miguel R Branco and Ana Pombo. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS biology*, 4(5) :e138, 2006.

- [19] Andrea H Brand, Linda Breeden, Judith Abraham, Rolf Sternganz, and Kim Nasmyth. Characterization of a “silencer” in yeast : a dna sequence with properties opposite to those of a transcriptional enhancer. *Cell*, 41(1) :41–48, 1985.
- [20] Erica M Briggs, Susan Ha, Paolo Mita, Gregory Brittingham, Ilaria Sciamanna, Corrado Spadafora, and Susan K Logan. Long interspersed nuclear element-1 expression and retrotransposition in prostate cancer cells. *Mobile DNA*, 9(1) :1, 2018.
- [21] Spencer W Brown. Heterochromatin. *Science*, 151(3709) :417–425, 1966.
- [22] David W Burt. A comprehensive review on the analysis of qtl in animals. *TRENDS in Genetics*, 18(9) :488, 2002.
- [23] William S Bush and Jason H Moore. Genome-wide association studies. *PLoS computational biology*, 8(12) :e1002822, 2012.
- [24] Sarah E Calvo, David J Pagliarini, and Vamsi K Mootha. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proceedings of the National Academy of Sciences*, 106(18) :7507–7512, 2009.
- [25] Nicolas Carels, Ramon Vidal, and Diego Frías. Universal features for the classification of coding and non-coding dna sequences. *Bioinformatics and biology insights*, 3 :BBI–S2236, 2009.
- [26] Douglas R Cavener. Comparison of the consensus sequence flanking translational start sites in drosophila and vertebrates. *Nucleic acids research*, 15(4) :1353–1361, 1987.
- [27] TR Cech. Ribozymes, the first 20 years, 2002.
- [28] Lingyi Chen and Jonathan Widom. Mechanism of transcriptional silencing in yeast. *Cell*, 120(1) :37–48, 2005.
- [29] Jeanne Chèneby, Marius Gheorghe, Marie Artufel, Anthony Mathelier, and Benoit Ballester. Remap 2018 : an updated atlas of regulatory regions from an integrative analysis of dna-binding chip-seq experiments. *Nucleic acids research*, 46(D1) :D267–D275, 2017.
- [30] Chao Cheng, Roger Alexander, Renqiang Min, Jing Leng, Kevin Y Yip, Joel Rozowsky, Koon-Kiu Yan, Xianjun Dong, Sarah Djebali, Yijun Ruan, et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome research*, 22(9) :1658–1667, 2012.
- [31] Jennifer C Chow, Constance Ciaudo, Melissa J Fazzari, Nathan Mise, Nicolas Servant, Jacob L Glass, Matthew Attreed, Philip Avner, Anton Wutz, Emmanuel Barillot, et al. Line-1 activity in facultative heterochromatin formation during x chromosome inactivation. *Cell*, 141(6) :956–969, 2010.
- [32] L Stirling Churchman and Jonathan S Weissman. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, 469(7330) :368, 2011.
- [33] 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319) :1061, 2010.
- [34] ENCODE Project Consortium et al. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature*, 447(7146) :799, 2007.
- [35] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414) :57–74, 2012.
- [36] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis : Multitissue gene regulation in humans. *Science*, 348(6235) :648–660, 2015.
- [37] International Human Genome Sequencing Consortium et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822) :860, 2001.
- [38] UniProt Consortium. Uniprot : the universal protein knowledgebase. *Nucleic acids research*, 45(D1) :D158–D169, 2016.
- [39] Sara J Cooper, Nathan D Trinklein, Elizabeth D Anton, Loan Nguyen, and Richard M Myers. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome research*, 16(1) :1–10, 2006.
- [40] Richard Cordaux and Mark A Batzer. The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10(10) :691, 2009.
- [41] Leighton J Core, André L Martins, Charles G Danko, Colin T Waters, Adam Siepel, and John T Lis. Analysis of nascent rna identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature genetics*, 46(12) :1311, 2014.

- [42] Leighton J Core, Joshua J Waterfall, Daniel A Gilchrist, David C Fargo, Hojoong Kwak, Karen Adelman, and John T Lis. Defining the status of rna polymerase at promoters. *Cell reports*, 2(4) :1025–1035, 2012.
- [43] Leighton J Core, Joshua J Waterfall, and John T Lis. Nascent rna sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909) :1845–1848, 2008.
- [44] Paula Cramer, C Gustavo Pesce, Francisco E Baralle, and Alberto R Kornblith. Functional association between promoter structure and transcript alternative splicing. *Proceedings of the National Academy of Sciences*, 94(21) :11456–11460, 1997.
- [45] Thomas Cremer, Christoph Cremer, H Baumann, EK Luedtke, K Sperling, V Teuber, and Christian Zorn. Rabl's model of the interphase chromosome arrangement tested in chinise hamster cells by premature chromosome condensation and laser-uv-microbeam experiments. *Human genetics*, 60(1) :46–56, 1982.
- [46] Thomas Cremer, Marion Cremer, Barbara Hübner, Hilmar Strickfaden, Daniel Smeets, Jens Popken, Michael Sterr, Yolanda Markaki, Karsten Rippe, and Christoph Cremer. The 4d nucleome : Evidence for a dynamic nuclear landscape based on co-aligned active and inactive nuclear compartments. *FEBS letters*, 589(20) :2931–2943, 2015.
- [47] Gérard CUNY, Philippe SORIANO, Gabriel MACAYA, and Giorgio BERNARDI. The major components of the mouse and human genomes : 1. preparation, basic properties and compositional heterogeneity. *European Journal of Biochemistry*, 115(2) :227–233, 1981.
- [48] Charles G Danko, Nasun Hah, Xin Luo, André L Martins, Leighton Core, John T Lis, Adam Siepel, and W Lee Kraus. Signaling pathways differentially affect rna polymerase ii initiation, pausing, and elongation rate in cells. *Molecular cell*, 50(2) :212–222, 2013.
- [49] Lan TM Dao, Ariel O Galindo-Albarrán, Jaime A Castro-Mondragon, Charlotte Andrieu-Soler, Alejandra Medina-Rivera, Charbel Souaid, Guillaume Charbonnier, Aurélien Griffon, Laurent Vanhille, Tharshana Stephen, et al. Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nature genetics*, 49(7) :1073, 2017.
- [50] Francesca De Santa, Iros Barozzi, Flore Mietton, Serena Ghisletti, Sara Polletti, Betsabeh Khoramian Tusi, Heiko Muller, Jiannis Ragoussis, Chia-Lin Wei, and Gioacchino Natoli. A large fraction of extragenic rna pol ii transcription sites overlap enhancers. *PLoS biology*, 8(5) :e1000384, 2010.
- [51] Aimée M Deaton and Adrian Bird. Cpg islands and the regulation of transcription. *Genes & development*, 25(10) :1010–1022, 2011.
- [52] Prescott Deininger. Alu elements : know the sines. *Genome biology*, 12(12) :236, 2011.
- [53] Prescott L Deininger, Mark A Batzer, Clyde A Hutchison III, and Marshall H Edgell. Master genes in mammalian repetitive dna amplification. *Trends in Genetics*, 8(9) :307–311, 1992.
- [54] Prescott L Deininger, John V Moran, Mark A Batzer, and Haig H Kazazian. Mobile elements and mammalian genome evolution. *Current opinion in genetics & development*, 13(6) :651–658, 2003.
- [55] Job Dekker and Edith Heard. Structural and functional diversity of topologically associating domains. *FEBS letters*, 589(20PartA) :2877–2884, 2015.
- [56] Job Dekker, Marc A Marti-Renom, and Leonid A Mirny. Exploring the three-dimensional organization of genomes : interpreting chromatin interaction data. *Nature Reviews Genetics*, 14(6) :390, 2013.
- [57] Ilenia D'Errico, Gemma Gadaleta, and Cecilia Saccone. Pseudogenes in metazoa : origin and features. *Briefings in Functional Genomics*, 3(2) :157–167, 2004.
- [58] Nicolas Descotes, Martin Heidemann, Lionel Spinelli, Roland Schüller, Muhammad Ahmad Maqbool, Romain Fenouil, Frederic Koch, Charlène Innocenti, Marta Gut, Ivo Gut, et al. Tyrosine phosphorylation of rna polymerase ii ctd is associated with antisense promoter transcription and active enhancers in mammalian cells. *Elife*, 3 :e02105, 2014.
- [59] Marie Dewannieux, Cécile Esnault, and Thierry Heidmann. Line-mediated retrotransposition of marked alu sequences. *Nature genetics*, 35(1) :41, 2003.
- [60] Marie Dewannieux and Thierry Heidmann. Role of poly (a) tail length in alu retrotransposition. *Genomics*, 86(3) :378–381, 2005.
- [61] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398) :376, 2012.
- [62] Xianjun Dong, Melissa C Greven, Anshul Kundaje, Sarah Djebali, James B Brown, Chao Cheng, Thomas R Gingeras, Mark Gerstein, Roderic Guigó, Ewan Birney, et al. Modeling gene expression using chromatin features in various cellular contexts. *Genome biology*, 13(9) :R53, 2012.

- [63] Shlomi Dvir, Lars Velten, Eilon Sharon, Danny Zeevi, Lucas B Carey, Adina Weinberger, and Eran Segal. Deciphering the rules by which 5'-utr sequences affect protein expression in yeast. *Proceedings of the National Academy of Sciences*, 110(30) :E2792–E2801, 2013.
- [64] Reyad A Elbarbary, Bronwyn A Lucas, and Lynne E Maquat. Retrotransposons as regulators of gene expression. *Science*, 351(6274) :aac7247, 2016.
- [65] Ran Elkon, Alejandro P Ugalde, and Reuven Agami. Alternative cleavage and polyadenylation : extent, regulation and function. *Nature Reviews Genetics*, 14(7) :496, 2013.
- [66] Jason Ernst and Manolis Kellis. Chromhmm : automating chromatin-state discovery and characterization. *Nature methods*, 9(3) :215, 2012.
- [67] Zeenat Farooq, Shahid Banday, Tej K Pandita, and Mohammad Altaf. The many faces of histone h3k79 methylation. *Mutation Research/Reviews in Mutation Research*, 768 :46–52, 2016.
- [68] Elena Fedorova and Daniele Zink. Nuclear architecture and gene regulation. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1783(11) :2174–2184, 2008.
- [69] Andrew P Feinberg. Phenotypic plasticity and the epigenetics of human disease. *Nature*, 447(7143) :433, 2007.
- [70] Cédric Feschotte and Ellen J Pritham. Dna transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.*, 41 :331–368, 2007.
- [71] Lars Feuk, Andrew R Carson, and Stephen W Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2) :85, 2006.
- [72] JT Finch and A Klug. Solenoidal model for superstructure in chromatin. *Proceedings of the National Academy of Sciences*, 73(6) :1897–1901, 1976.
- [73] Matthew A Firpo and Albert E Dahlberg. The role of ribosomal rna in the control of gene expression. In *Post-Transcriptional Control of Gene Expression*, pages 185–195. Springer, 1990.
- [74] Walther Flemming. *Zellsubstanz, kern und zelltheilung*. Vogel, 1882.
- [75] Paul Flicek, Ikhlaq Ahmed, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, et al. Ensembl 2013. *Nucleic acids research*, 41(D1) :D48–D55, 2012.
- [76] Nova Fong, Hyunmin Kim, Yu Zhou, Xiong Ji, Jinsong Qiu, Tassa Saldi, Katrina Diener, Ken Jones, Xiang-Dong Fu, and David L Bentley. Pre-mrna splicing is facilitated by an optimal rna polymerase ii elongation rate. *Genes & development*, 28(23) :2663–2676, 2014.
- [77] Alistair RR Forrest, Hideya Kawaji, Michael Rehli, J Kenneth Baillie, Michiel JL De Hoon, Vanja Haberle, Timo Lassmann, Ivan V Kulakovskiy, Marina Lizio, Masayoshi Itoh, et al. A promoter-level mammalian expression atlas. *Nature*, 507(7493) :462, 2014.
- [78] Geneviève Fourel, Frédérique Magdinier, and Eric Gilson. Insulator dynamics and the setting of chromatin domains. *Bioessays*, 26(5) :523–532, 2004.
- [79] Nicole J Francis and Robert E Kingston. Mechanisms of transcriptional memory. *Nature Reviews Molecular Cell Biology*, 2(6) :409, 2001.
- [80] Melissa J Fullwood, Chia-Lin Wei, Edison T Liu, and Yijun Ruan. Next-generation dna sequencing of paired-end tags (pet) for transcriptome and genome analyses. *Genome research*, 19(4) :521–532, 2009.
- [81] M Gardiner-Garden and M Frommer. Cpg islands in vertebrate genomes. *Journal of molecular biology*, 196(2) :261–282, 1987.
- [82] Marcel Geertz and Sebastian J Maerkli. Experimental strategies for studying transcription factor-dna binding specificities. *Briefings in functional genomics*, 9(5-6) :362–373, 2010.
- [83] Rita Gemayel, Marcelo D Vinces, Matthieu Legendre, and Kevin J Verstrepen. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual review of genetics*, 44 :445–477, 2010.
- [84] Johan H Gibcus and Job Dekker. The hierarchy of the 3d genome. *Molecular cell*, 49(5) :773–782, 2013.
- [85] John L Goodier. Restricting retrotransposons : a review. *Mobile DNA*, 7(1) :16, 2016.
- [86] Johannes Gräff and Isabelle M Mansuy. Epigenetic codes in cognition and behaviour. *Behavioural brain research*, 192(1) :70–87, 2008.
- [87] Peyton G Greenside, Tyler Shimko, Polly Fordyce, and Anshul Kundaje. Discovering epistatic feature interactions from neural network models of regulatory dna sequences. *bioRxiv*, page 302711, 2018.

- [88] Sam Griffiths-Jones, Alex Bateman, Mhairi Marshall, Ajay Khanna, and Sean R Eddy. Rfam : an rna family database. *Nucleic acids research*, 31(1) :439–441, 2003.
- [89] Lars Guelen, Ludo Pagine, Emilie Brasset, Wouter Meuleman, Marius B Faza, Wendy Talhout, Bert H Eussen, Annelies de Klein, Lodewyk Wessels, Wouter de Laat, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197) :948, 2008.
- [90] Mitchell Guttman, Ido Amit, Manuel Garber, Courtney French, Michael F Lin, David Feldser, Maite Huarte, Or Zuk, Bryce W Carey, John P Cassady, et al. Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals. *Nature*, 458(7235) :223, 2009.
- [91] Melissa Gymrek, Thomas Willems, Audrey Guilmartre, Haoyang Zeng, Barak Markus, Stoyan Georgiev, Mark J Daly, Alkes L Price, Jonathan K Pritchard, Andrew J Sharp, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature genetics*, 48(1) :22, 2016.
- [92] Nasun Hah, Shino Murakami, Anusha Nagari, Charles G Danko, and W Lee Kraus. Enhancer transcripts mark active estrogen receptor binding sites. *Genome research*, 2013.
- [93] Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1) :57–70, 2000.
- [94] Matthew P Hare and Stephen R Palumbi. High intron sequence conservation across three mammalian orders suggests functional constraints. *Molecular Biology and Evolution*, 20(6) :969–978, 2003.
- [95] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. Gencode : the reference human genome annotation for the encode project. *Genome research*, 22(9) :1760–1774, 2012.
- [96] Arne Hauenschild, Leonie Ringrose, Christina Altmutter, Renato Paro, and Marc Rehmsmeier. Evolutionary plasticity of polycomb/trithorax response elements in drosophila species. *PLoS biology*, 6(10) :e261, 2008.
- [97] Ericka R Havecker, Xiang Gao, and Daniel F Voytas. The diversity of ltr retrotransposons. *Genome biology*, 5(6) :225, 2004.
- [98] Nathaniel D Heintzman, Gary C Hon, R David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F Harp, Zhen Ye, Leonard K Lee, Rhona K Stuart, Christina W Ching, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243) :108, 2009.
- [99] Nathaniel D Heintzman, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, R David Hawkins, Leah O Barrera, Sara Van Calcar, Chunxu Qu, Keith A Ching, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3) :311, 2007.
- [100] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular cell*, 38(4) :576–589, 2010.
- [101] Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes, and William Stafford Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods*, 9(5) :473, 2012.
- [102] Chung-Chau Hon, Jordan A Ramilowski, Jayson Harshbarger, Nicolas Bertin, Owen JL Rackham, Julian Gough, Elena Denisenko, Sebastian Schmeier, Thomas M Poulsen, Jessica Severin, et al. An atlas of human long non-coding rnas with accurate 5' ends. *Nature*, 543(7644) :199, 2017.
- [103] Rollin D Hotchkiss. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *Journal of Biological Chemistry*, 175(1) :315–332, 1948.
- [104] Chunhui Hou, Li Li, Zhaojun S Qin, and Victor G Corces. Gene density, transcription, and insulators contribute to the partition of the drosophila genome into physical domains. *Molecular cell*, 48(3) :471–484, 2012.
- [105] Kenji Ichiyanagi. Epigenetic regulation of transcription and possible functions of mammalian short interspersed elements, sines. *Genes & genetic systems*, 88(1) :19–29, 2013.
- [106] Kenji Ichiyanagi, Yungfeng Li, Toshiaki Watanabe, Tomoko Ichiyanagi, Kei Fukuda, Junko Kitayama, Yasuhiro Yamamoto, Satomi Kuramochi-Miyagawa, Toru Nakano, Yukihiro Yabuta, et al. Locus-and domain-dependent control of dna methylation at mouse b1 retrotransposons during male germ cell development. *Genome research*, 2011.
- [107] Kohta Ikegami, Thea A Egelhofer, Susan Strome, and Jason D Lieb. Caenorhabditis elegans chromosome arms are anchored to the nuclear membrane via discontinuous association with lem-2. *Genome biology*, 11(12) :R120, 2010.

- [108] Robert S Illingworth and Adrian P Bird. Cpg islands–‘a rough guide’. *FEBS letters*, 583(11) :1713–1720, 2009.
- [109] Nicholas T Ingolia, Sina Ghaemmaghami, John RS Newman, and Jonathan S Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *science*, 324(5924) :218–223, 2009.
- [110] Matthew K Iyer, Yashar S Niknafs, Rohit Malik, Udit Singhal, Anirban Sahu, Yasuyuki Hosono, Terrence R Barrette, John R Prensner, Joseph R Evans, Shuang Zhao, et al. The landscape of long noncoding rnas in the human transcriptome. *Nature genetics*, 47(3) :199, 2015.
- [111] Kamel Jabbari and Giorgio Bernardi. An isochore framework underlies chromatin architecture. *PloS one*, 12(1) :e0168023, 2017.
- [112] Calvin H Jan, Robin C Friedman, J Graham Ruby, and David P Bartel. Formation, regulation and evolution of caenorhabditis elegans 3’utrs. *Nature*, 469(7328) :97, 2011.
- [113] J Yi Jason, Janet Berrios, Jason M Newbern, William D Snider, Benjamin D Philpot, Klaus M Hahn, and Mark J Zylka. An autism-linked mutation disables phosphorylation control of ube3a. *Cell*, 162(4) :795–807, 2015.
- [114] Torben Heick Jensen, Alain Jacquier, and Domenico Libri. Dealing with pervasive transcription. *Molecular cell*, 52(4) :473–484, 2013.
- [115] Thomas Jenuwein and C David Allis. Translating the histone code. *Science*, 293(5532) :1074–1080, 2001.
- [116] Bong-Seok Jo and Sun Shim Choi. Introns : the functional benefits of introns in genomes. *Genomics & informatics*, 13(4) :112–118, 2015.
- [117] Travis S Johnson, Sihong Li, Jonathan R Kho, Kun Huang, and Yan Zhang. Network analysis of pseudogene-gene relationships : From pseudogene evolution to their functional potentials. In *Pac. Symp. Biocomput*, volume 23, pages 536–547. World Scientific, 2018.
- [118] Peter A Jones. Functions of dna methylation : islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, 13(7) :484, 2012.
- [119] Iris Jonkers, Hojoong Kwak, and John T Lis. Genome-wide dynamics of pol ii elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife*, 3 :e02407, 2014.
- [120] Iris Jonkers and John T Lis. Getting up to speed with transcription elongation by rna polymerase ii. *Nature reviews Molecular cell biology*, 16(3) :167, 2015.
- [121] Jerzy Jurka, Vladimir V Kapitonov, A Pavlicek, P Klonowski, O Kohany, and J Walichiewicz. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research*, 110(1-4) :462–467, 2005.
- [122] Vera B Kaiser, Martin S Taylor, and Colin A Semple. Mutational biases drive elevated rates of substitution at regulatory sites across cancer types. *PLoS genetics*, 12(8) :e1006207, 2016.
- [123] Mutsumi Kanamori-Katayama, Masayoshi Itoh, Hideya Kawaji, Timo Lassmann, Shintaro Katayama, Miki Kojima, Nicolas Bertin, Ai Kaiho, Noriko Ninomiya, Carsten O Daub, et al. Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome research*, 2011.
- [124] Philipp Kapranov, Jill Cheng, Sujit Dike, David A Nix, Radharani Duttagupta, Aarron T Willingham, Peter F Stadler, Jana Hertel, Jörg Hackermüller, Ivo L Hofacker, et al. Rna maps reveal new rna classes and a possible function for pervasive transcription. *Science*, 316(5830) :1484–1488, 2007.
- [125] Roshan Karki, Deep Pandya, Robert C Elston, and Cristiano Ferlini. Defining “mutation” and “polymorphism” in the era of personal genomics. *BMC medical genomics*, 8(1) :37, 2015.
- [126] Rosa Karlić, Ho-Ryun Chung, Julia Lasserre, Kristian Vlahoviček, and Martin Vingron. Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences*, 107(7) :2926–2931, 2010.
- [127] Maya Kasowski, Sofia Kyriazopoulou-Panagiotopoulou, Fabian Grubert, Judith B Zaugg, Anshul Kundaje, Yuling Liu, Alan P Boyle, Qiangfeng Cliff Zhang, Fouad Zakharia, Damek V Spacek, et al. Extensive variation in chromatin states across humans. *Science*, 342(6159) :750–752, 2013.
- [128] Haig H Kazazian. Mobile elements : drivers of genome evolution. *science*, 303(5664) :1626–1632, 2004.
- [129] DR Kelley, YA Reshef, D Belanger, C McLean, J Snoek, and M Bileschi. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *biorxiv*, 2018.

- [130] Aziz Khan, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A Castro-Mondragon, Robin van der Lee, Adrien Bessy, Jeanne Cheneby, Shubhada R Kulkarni, Ge Tan, et al. Jaspar 2018 : update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic acids research*, 46(D1) :D260–D266, 2017.
- [131] Saadi Khochbin, André Verdel, Claudie Lemercier, and Daphné Seigneurin-Berny. Functional significance of histone deacetylase diversity. *Current opinion in genetics & development*, 11(2) :162–166, 2001.
- [132] Helena Kilpinen, Sebastian M Waszak, Andreas R Gschwind, Sunil K Raghav, Robert M Witwicki, Andrea Orioli, Eugenia Migliavacca, Michaël Wiederkehr, Maria Gutierrez-Arcelus, Nikolaos I Panousis, et al. Coordinated effects of sequence variation on dna binding, chromatin structure, and transcription. *Science*, 342(6159) :744–747, 2013.
- [133] Songmi Kim, Chun-Sung Cho, Kyudong Han, and Jungnam Lee. Structural variation of alu element and human disease. *Genomics & informatics*, 14(3) :70–77, 2016.
- [134] Tae-Kyung Kim, Martin Hemberg, Jesse M Gray, Allen M Costa, Daniel M Bear, Jing Wu, David A Harmin, Mike Laptevich, Kellie Barbara-Haley, Scott Kuersten, et al. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295) :182, 2010.
- [135] Adrienne Kitts and Stephen Sherry. The single nucleotide polymorphism database (dbSNP) of nucleotide sequence variation. *The NCBI Handbook. McEntyre J, Ostell J, eds. Bethesda, MD : US National Center for Biotechnology Information*, 2002.
- [136] Wigard P Kloosterman, Laurent C Francioli, Fereydoun Hormozdiari, Tobias Marschall, Jayne Y Hehir-Kwa, Abdel Abdellaoui, Eric-Wubbo Lameijer, Matthijs H Moed, Vyacheslav Koval, Ivo Renkens, et al. Characteristics of de novo structural changes in the human genome. *Genome research*, 2015.
- [137] Rimantas Kodzius, Miki Kojima, Hiromi Nishiyori, Mari Nakamura, Shiro Fukuda, Michihira Tagami, Daisuke Sasaki, Kengo Imamura, Chikatoshi Kai, Matthias Harbers, et al. Cage : cap analysis of gene expression. *Nature methods*, 3(3) :211–222, 2006.
- [138] Nikolay Kolesnikov, Emma Hastings, Maria Keays, Olga Melnichuk, Y Amy Tang, Eleanor Williams, Miroslaw Dylag, Natalja Kurbatova, Marco Brandizi, Tony Burdett, et al. Arrayexpress update—simplifying data submissions. *Nucleic acids research*, 43(D1) :D1113–D1116, 2014.
- [139] Petros Kolovos, Tobias A Knoch, Frank G Grosveld, Peter R Cook, and Argyris Papantonis. Enhancers and silencers : an integrated and simple model for their function. *Epigenetics & chromatin*, 5(1) :1, 2012.
- [140] Eugene V Koonin. The origin of introns and their role in eukaryogenesis : a compromise solution to the introns-early versus introns-late debate? *Biology direct*, 1(1) :22, 2006.
- [141] Roger D Kornberg. Chromatin structure : a repeating unit of histones and dna. *Science*, 184(4139) :868–871, 1974.
- [142] Ana Kozomara and Sam Griffiths-Jones. mirbase : annotating high confidence micrornas using deep sequencing data. *Nucleic acids research*, 42(D1) :D68–D73, 2013.
- [143] Emily N Kroutter, Victoria P Belancio, Bradley J Wagstaff, and Astrid M Roy-Engel. The rna polymerase dictates orf1 requirement and timing of line and sine retrotransposition. *PloS genetics*, 5(4) :e1000458, 2009.
- [144] Maarten Kruithof, Fan-Tso Chien, Andrew Routh, Colin Logie, Daniela Rhodes, and John Van Noort. Single-molecule force spectroscopy reveals a highly compliant helical folding for the 30-nm chromatin fiber. *Nature structural & molecular biology*, 16(5) :534, 2009.
- [145] Jack Kuipers, Katharina Jahn, Benjamin J Raphael, and Niko Beerenwinkel. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome research*, 2017.
- [146] Ivan V Kulakovskiy, Ilya E Vorontsov, Ivan S Yevshin, Ruslan N Sharipov, Alla D Fedorova, Eugene I Rumynskiy, Yulia A Medvedeva, Arturo Magana-Mora, Vladimir B Bajic, Dmitry A Papatsenko, et al. Hocomoco : towards a complete collection of transcription factor binding models for human and mouse via large-scale chip-seq analysis. *Nucleic acids research*, 46(D1) :D252–D259, 2017.
- [147] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539) :317, 2015.
- [148] Laimonis Laimins, Monika Holmgren-Koenig, and George Khoury. Transcriptional "silencer" element in rat repetitive sequences associated with the rat insulin 1 gene locus. *Proceedings of the National Academy of Sciences*, 83(10) :3151–3155, 1986.

- [149] Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457) :214, 2013.
- [150] Rosalind C Lee, Rhonda L Feinbaum, and Victor Ambros. The *c. elegans* heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *cell*, 75(5) :843–854, 1993.
- [151] William Lee, Desiree Tillo, Nicolas Bray, Randall H Morse, Ronald W Davis, Timothy R Hughes, and Corey Nislow. A high-resolution atlas of nucleosome occupancy in yeast. *Nature genetics*, 39(10) :1235, 2007.
- [152] Yeon Lee and Donald C Rio. Mechanisms and regulation of alternative pre-mrna splicing. *Annual review of biochemistry*, 84 :291–323, 2015.
- [153] Benjamin P Lewis, Richard E Green, and Steven E Brenner. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mrna decay in humans. *Proceedings of the National Academy of Sciences*, 100(1) :189–192, 2003.
- [154] Bo Li and Colin N Dewey. Rsem : accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1) :323, 2011.
- [155] Cai Li, Boris Lenhard, and Nicholas M Luscombe. Integrated analysis sheds light on evolutionary trajectories of young transcription start sites in the human genome. *Genome research*, 2018.
- [156] Guoliang Li, Liuyang Cai, Huidan Chang, Ping Hong, Qiangwei Zhou, Ekaterina V Kulakova, Nikolay A Kolanov, and Yijun Ruan. Chromatin interaction analysis with paired-end tag (chia-pet) sequencing technology and application. *BMC genomics*, 15(12) :S11, 2014.
- [157] Wenbo Li, Dimple Notani, and Michael G Rosenfeld. Enhancers as non-coding rna transcription units : recent insights and future perspectives. *Nature Reviews Genetics*, 17(4) :207, 2016.
- [158] Yue Li, Minggao Liang, and Zhaolei Zhang. Regression analysis of combined gene expression regulation in acute myeloid leukemia. *PLoS Comput Biol*, 10(10) :e1003908, 2014.
- [159] Peter Lichter, Thomas Cremer, J Borden, L Manvelidis, and DC Ward. Delineation of individual human chromosomes in metaphase and interphase cells by in situ suppression hybridization using recombinant dna libraries. *Human genetics*, 80(3) :224–234, 1988.
- [160] Erez Lieberman-Aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950) :289–293, 2009.
- [161] Shuibin Lin and Richard I Gregory. Microrna biogenesis pathways in cancer. *Nature reviews cancer*, 15(6) :321, 2015.
- [162] Colton Linnertz, Lauren Anderson, William Gottschalk, Donna Crenshaw, Michael W Lutz, Jawara Allen, Sunita Saith, Mirta Mihovilovic, James R Burke, Kathleen A Welsh-Bohmer, et al. The cis-regulatory effect of an alzheimer’s disease-associated poly-t locus on expression of tomm40 and apolipoprotein e genes. *Alzheimer’s & Dementia*, 10(5) :541–551, 2014.
- [163] Stavros Lomvardas, Gilad Barnea, David J Pisapia, Monica Mendelsohn, Jennifer Kirkland, and Richard Axel. Interchromosomal interactions and olfactory receptor choice. *Cell*, 126(2) :403–413, 2006.
- [164] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6) :580, 2013.
- [165] Michael Lynch. Intron evolution as a population-genetic process. *Proceedings of the National Academy of Sciences*, 99(9) :6118–6123, 2002.
- [166] Gabriel Macaya, Jean-Paul Thiery, and Giorgio Bernardi. An approach to the organization of eukaryotic genomes at a macromolecular level. *Journal of molecular biology*, 108(1) :237–254, 1976.
- [167] Trudy FC Mackay, Eric A Stone, and Julien F Ayroles. The genetics of quantitative traits : challenges and prospects. *Nature Reviews Genetics*, 10(8) :565, 2009.
- [168] Glenn A Maston, Sara K Evans, and Michael R Green. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, 7 :29–59, 2006.
- [169] A Gregory Matera, Rebecca M Terns, and Michael P Terns. Non-coding rnas : lessons from the small nuclear and small nucleolar rnas. *Nature reviews Molecular cell biology*, 8(3) :209, 2007.
- [170] Anthony Mathelier and Wyeth W Wasserman. The next generation of transcription factor binding site prediction. *PLoS computational biology*, 9(9) :e1003214, 2013.

- [171] A Mattout-Drubetzki and Y Gruenbaum. Dynamic interactions of nuclear lamina proteins with chromatin and transcriptional machinery. *Cellular and Molecular Life Sciences CMLS*, 60(10) :2053–2063, 2003.
- [172] Christine Mayr. Regulation by 3'-untranslated regions. *Annual review of genetics*, 51 :171–194, 2017.
- [173] Barbara McClintock. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, 36(6) :344–355, 1950.
- [174] Robert C McLeay, Tom Lesluyes, Gabriel Cuellar Partida, and Timothy L Bailey. Genome-wide in silico prediction of gene expression. *Bioinformatics*, 28(21) :2789–2796, 2012.
- [175] Graham McVicker, Bryce van de Geijn, Jacob F Degner, Carolyn E Cain, Nicholas E Banovich, Anil Raj, Noah Lewellen, Marsha Myrthil, Yoav Gilad, and Jonathan K Pritchard. Identification of genetic variants that affect histone modifications in human cells. *Science*, 342(6159) :747–749, 2013.
- [176] Karen J Meaburn and Tom Misteli. Cell biology : chromosome territories. *Nature*, 445(7126) :379, 2007.
- [177] Pankaj Mehta, David J Schwab, and Anirvan M Sengupta. Statistical mechanics of transcription-factor binding site discovery using hidden markov models. *Journal of statistical physics*, 142(6) :1187–1205, 2011.
- [178] Matthew Meselson, Franklin W Stahl, and Jerome Vinograd. Equilibrium sedimentation of macromolecules in density gradients. *Proceedings of the National Academy of Sciences*, 43(7) :581–588, 1957.
- [179] Wouter Meuleman, Daan Peric-Hupkes, Jop Kind, Jean-Bernard Beaudry, Ludo Pagie, Manolis Kellis, Marcel Reinders, Lodewyk Wessels, and Bas van Steensel. Constitutive nuclear lamina-genome interactions are highly conserved and associated with a/t-rich sequence. *Genome research*, pages gr–141028, 2012.
- [180] Laurence R Meyer, Ann S Zweig, Angie S Hinrichs, Donna Karolchik, Robert M Kuhn, Matthew Wong, Cricket A Sloan, Kate R Rosenbloom, Greg Roe, Brooke Rhead, et al. The ucsc genome browser database : extensions and updates 2013. *Nucleic acids research*, 41(D1) :D64–D69, 2012.
- [181] Yoshio Miki, Jeff Swensen, Donna Shattuck-Eidens, P Andrew Futreal, Keith Harshman, Sean Tavtigian, Qingyun Liu, Charles Cochran, L Michelle Bennett, Wei Ding, et al. A strong candidate for the breast and ovarian cancer susceptibility gene brca1. *Science*, 266(5182) :66–71, 1994.
- [182] Sergei M Mirkin. Expandable dna repeats and human disease. *Nature*, 447(7147) :932, 2007.
- [183] Csaba Miskey, Zsuzsanna Izsvák, Koichi Kawakami, and Zoltán Ivics. Dna transposons in vertebrate functional genomics. *Cellular and molecular life sciences*, 62(6) :629, 2005.
- [184] Tom Misteli. Higher-order genome organization in human disease. *Cold Spring Harbor perspectives in biology*, page a000794, 2010.
- [185] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7) :621, 2008.
- [186] Raphaël Mourad and Olivier Cuvier. Tad-free analysis of architectural proteins and insulators. *Nucleic acids research*, 46(5) :e27–e27, 2017.
- [187] Gioacchino Natoli and Jean-Christophe Andrau. Noncoding transcription at enhancers : general principles and functional models. *Annual review of genetics*, 46 :1–19, 2012.
- [188] Natalia Naumova, Emily M Smith, Ye Zhan, and Job Dekker. Analysis of long-range chromatin interactions using chromosome conformation capture. *Methods*, 58(3) :192–203, 2012.
- [189] Elphège P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L van Berkum, Johannes Meisig, John Sedat, et al. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398) :381, 2012.
- [190] Steven Ogbourne and Toni M Antalis. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochemical Journal*, 331(1) :1–14, 1998.
- [191] Ada L Olins and Donald E Olins. Spheroid chromatin units ( $\nu$  bodies). *Science*, 183(4122) :330–332, 1974.
- [192] Eric M Ostertag and Haig H Kazazian Jr. Biology of mammalian l1 retrotransposons. *Annual review of genetics*, 35(1) :501–538, 2001.
- [193] Zhengqing Ouyang, Qing Zhou, and Wing Hung Wong. Chip-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences*, pages pnas–0904863106, 2009.
- [194] Alexander F Palazzo and Eliza S Lee. Non-coding rna : what is functional and what is junk ? *Frontiers in genetics*, 6 :2, 2015.
- [195] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12) :1413, 2008.

- [196] Peter J Park. Chip-seq : advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10) :669, 2009.
- [197] Sung-Joon Park and Kenta Nakai. A regression analysis of gene expression in es cells reveals two gene classes that are significantly different in epigenetic patterns. *BMC bioinformatics*, 12(1) :S50, 2011.
- [198] EBERHARD Passarge. Emil heitz and the concept of heterochromatin : longitudinal chromosome differentiation was recognized fifty years ago. *American journal of human genetics*, 31(2) :106, 1979.
- [199] A Payton, P Sindrewicz, V Pessoa, H Platt, M Horan, W Ollier, VJ Bubb, N Pendleton, and JP Quinn. A tomm40 poly-t variant modulates gene expression and is associated with vocabulary ability and decline in nonpathologic aging. *Neurobiology of aging*, 39 :217–e1, 2016.
- [200] Christopher E Pearson, Kerrie Nichol Edamura, and John D Cleary. Repeat instability : mechanisms of dynamic mutations. *Nature Reviews Genetics*, 6(10) :729, 2005.
- [201] Dilmi Perera, Rebecca C Poulos, Anushi Shah, Dominik Beck, John E Pimanda, and Jason WH Wong. Differential dna repair underlies mutation hotspots at active promoters in cancer genomes. *Nature*, 532(7598) :259, 2016.
- [202] Daan Peric-Hupkes, Wouter Meuleman, Ludo Pagie, Sophia WM Bruggeman, Irina Solovei, Wim Brugman, Stefan Gräf, Paul Flicek, Ron M Kerkhoven, Maarten van Lohuizen, et al. Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Molecular cell*, 38(4) :603–613, 2010.
- [203] Becky M Pickering and Anne E Willis. The implications of structured 5' untranslated regions on translation and disease. In *Seminars in cell & developmental biology*, volume 16, pages 39–47. Elsevier, 2005.
- [204] Helen Pickersgill, Bernike Kalverda, Elzo de Wit, Wendy Talhout, Maarten Fornerod, and Bas van Steensel. Characterization of the drosophila melanogaster genome at the nuclear lamina. *Nature genetics*, 38(9) :1005, 2006.
- [205] D Pinkel, J Landegent, C Collins, J Fuscoe, R Segraves, J Lucas, and Joe Gray. Fluorescence in situ hybridization with human chromosome-specific libraries : detection of trisomy 21 and translocations of chromosome 4. *Proceedings of the National Academy of Sciences*, 85(23) :9138–9142, 1988.
- [206] Hara Polioudaki, Niki Kourmouli, Victoria Drosou, Alexandra Bakou, Panayiotis A Theodoropoulos, Prim B Singh, Thomas Giannakouros, and Spyros D Georgatos. Histones h3/h4 form a tight complex with the inner nuclear membrane protein lbr and heterochromatin protein 1. *EMBO reports*, 2(10) :920–925, 2001.
- [207] Roy Pollock and Richard Treisman. A sensitive method for the determination of protein-dna binding specificities. *Nucleic Acids Research*, 18(21) :6197–6204, 1990.
- [208] Petter Portin and Adam Wilkins. The evolving definition of the term “gene”. *Genetics*, 205(4) :1353–1364, 2017.
- [209] Sebastian Pott and Jason D Lieb. What are super-enhancers? *Nature genetics*, 47(1) :8, 2015.
- [210] Nick J Proudfoot. Transcriptional termination in mammals : Stopping the rna polymerase ii juggernaut. *Science*, 352(6291) :aad9926, 2016.
- [211] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. Ncbi reference sequence (refseq) : a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 33(suppl\_1) :D501–D504, 2005.
- [212] Daniel Quang and Xiaohui Xie. Danq : a hybrid convolutional and recurrent deep neural network for quantifying the function of dna sequences. *Nucleic acids research*, 44(11) :e107–e107, 2016.
- [213] T. Quante and A. Bird. Do short, frequent DNA sequence motifs mould the epigenome? *Nat. Rev. Mol. Cell Biol.*, 17(4) :257–262, Apr 2016.
- [214] Jesse R Raab and Rohinton T Kamakaka. Insulators and promoters : closer than we think. *Nature Reviews Genetics*, 11(6) :439, 2010.
- [215] Alvaro Rada-Iglesias, Frank G Grosveld, and Argyris Papantonis. Forces driving the three-dimensional folding of eukaryotic genomes. *Molecular systems biology*, 14(6) :e8214, 2018.
- [216] Gajendra PS Raghava and Joon H Han. Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein. *BMC bioinformatics*, 6(1) :59, 2005.
- [217] Julia D Ransohoff, Yuning Wei, and Paul A Khavari. The functions and unique features of long intergenic non-coding rna. *Nature reviews Molecular cell biology*, 19(3) :143, 2018.
- [218] Suhas SP Rao, Su-Chen Huang, Brian Glenn St Hilaire, Jesse M Engreitz, Elizabeth M Perez, Kyong-Rim Kieffer-Kwon, Adrian L Sanborn, Sarah E Johnstone, Gavin D Bascom, Ivan D Bochkov, et al. Cohesin loss eliminates all loop domains. *Cell*, 171(2) :305–320, 2017.

- [219] Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7) :1665–1680, 2014.
- [220] Eduardo M Reis, Rodrigo Louro, Helder I Nakaya, and Sergio Verjovski-Almeida. As antisense rna gets intronic. *Omics : a journal of integrative biology*, 9(1) :2–12, 2005.
- [221] Ruth-Ariane Rober, Klaus Weber, and Mary Osborn. Differential timing of nuclear lamin a/c expression in the various organs of the mouse embryo and the young animal : a developmental study. *Development*, 105(2) :365–378, 1989.
- [222] Nemanja Rodić, Reema Sharma, Rajni Sharma, John Zampella, Lixin Dai, Martin S Taylor, Ralph H Hruban, Christine A Iacobuzio-Donahue, Anirban Maitra, Michael S Torbenson, et al. Long interspersed element-1 protein expression is a hallmark of many human cancers. *The American journal of pathology*, 184(5) :1280–1286, 2014.
- [223] Igor B Rogozin, Liran Carmel, Miklos Csuros, and Eugene V Koonin. Origin and evolution of spliceosomal introns. *Biology direct*, 7(1) :11, 2012.
- [224] Casey E Romanoski, Christopher K Glass, Hendrik G Stunnenberg, Laurence Wilson, and Genevieve Almouzni. Epigenomics : Roadmap for regulation. *Nature*, 518(7539) :314, 2015.
- [225] Gil Ron, Yuval Globerson, Dror Moran, and Tommy Kaplan. Promoter-enhancer interactions identified from hi-c data using probabilistic models and hierarchical topological domains. *Nature communications*, 8(1) :2237, 2017.
- [226] Scott William Roy and Walter Gilbert. The evolution of spliceosomal introns : patterns, puzzles and progress. *Nature Reviews Genetics*, 7(3) :211, 2006.
- [227] Astrid M Roy-Engel, Abdel-Halim Salem, Oluwatosin O Oyeniran, Lisa Deininger, Dale J Hedges, Gail E Kilroy, Mark A Batzer, and Prescott L Deininger. Active alu element “a-tails” : size does matter. *Genome research*, 12(9) :1333–1344, 2002.
- [228] Radhakrishnan Sabarinathan, Loris Mularoni, Jordi Deu-Pons, Abel Gonzalez-Perez, and Núria López-Bigas. Nucleotide excision repair is impaired by binding of transcription factors to dna. *Nature*, 532(7598) :264, 2016.
- [229] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. Jaspar : an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(suppl\_1) :D91–D94, 2004.
- [230] Noriko Sasaki-Haraguchi, Makoto K Shimada, Ichiro Taniguchi, Mutsuhito Ohno, and Akila Mayeda. Mechanistic insights into human pre-mrna splicing of human ultra-short introns : potential unusual mechanism identifies g-rich introns. *Biochemical and biophysical research communications*, 423(2) :289–294, 2012.
- [231] Florian Schmidt, Nina Gasparoni, Gilles Gasparoni, Kathrin Gianmoena, Cristina Cadenas, Julia K Polansky, Peter Ebert, Karl Nordstroem, Matthias Barann, Anupam Sinha, et al. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic acids research*, 45(1) :54–66, 2016.
- [232] Florian Schmidt, Marcel H Schulz, and Inanc Birol. On the problem of confounders in modeling gene expression. *Bioinformatics*, 1 :9, 2018.
- [233] Bernd Schuettengruber, Daniel Chourrout, Michel Vervoort, Benjamin Leblanc, and Giacomo Cavalli. Genome regulation by polycomb and trithorax proteins. *Cell*, 128(4) :735–745, 2007.
- [234] Björn Schwalb, Margaux Michel, Benedikt Zacher, Katja Fröhlauf, Carina Demel, Achim Tresch, Julien Gagneur, and Patrick Cramer. Tt-seq maps the human transient transcriptome. *Science*, 352(6290) :1225–1228, 2016.
- [235] Wibke Schwarzer, Nezar Abdennur, Anton Goloborodko, Aleksandra Pekowska, Geoffrey Fudenberg, Yann Loe-Mie, Nuno A Fonseca, Wolfgang Huber, Christian H Haering, Leonid Mirny, et al. Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, 551(7678) :51, 2017.
- [236] Eran Segal, Yvonne Fondufe-Mittendorf, Lingyi Chen, AnnChristine Thåström, Yair Field, Irene K Moore, Ji-Ping Z Wang, and Jonathan Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104) :772, 2006.
- [237] Luke A Selth, Stefan Sigurdsson, and Jesper Q Svejstrup. Transcript elongation by rna polymerase ii. *Annual review of biochemistry*, 79 :271–293, 2010.

- [238] Tom Sexton, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, 148(3) :458–472, 2012.
- [239] PC Sham and SS Cherny. Genetic architecture of complex diseases. In *Analysis of Complex Disease Association Studies*, pages 1–13. Elsevier, 2011.
- [240] Shikhar Sharma, Theresa K Kelly, and Peter A Jones. Epigenetics in cancer. *Carcinogenesis*, 31(1) :27–36, 2010.
- [241] Stephen Jefferson Sharp, Jerone Schaack, Lyan Cooley, Debroh Johnson Burke, and Dieter Soil. Structure and transcription of eukaryotic trna gene. *Critical Reviews In Biochemistry*, 19(2) :107–144, 1985.
- [242] Orit Shaul. How introns enhance gene expression. *The international journal of biochemistry & cell biology*, 91 :145–155, 2017.
- [243] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8) :1034–1050, 2005.
- [244] Robert J Sims, Rimma Belotserkovskaya, and Danny Reinberg. Elongation by rna polymerase ii : the short and long of it. *Genes & development*, 18(20) :2437–2468, 2004.
- [245] Raz Somech, Sigal Shaklai, Orit Geller, Ninette Amariglio, Amos J Simon, Gideon Rechavi, and Einav Nili Gal-Yam. The nuclear-envelope protein and transcriptional repressor lap2 $\beta$  interacts with hdac3 at the nuclear periphery, and induces histone h4 deacetylation. *Journal of cell science*, 118(17) :4017–4025, 2005.
- [246] Wendy Weijia Soon, Manoj Hariharan, and Michael P Snyder. High-throughput sequencing for biology and medicine. *Molecular systems biology*, 9(1) :640, 2013.
- [247] Rotem Sorek and Gil Ast. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Research*, 13(7) :1631–1637, 2003.
- [248] Georges St Laurent, Dmitry Shtokalo, Michael R Tackett, Zhaoqing Yang, Tatyana Eremina, Claes Wahlestedt, Silvio Urcuqui-Inchima, Bernd Seilheimer, Timothy A McCaffrey, and Philipp Kapranov. Intronic rnas constitute the major fraction of the non-coding rna in mammalian cells. *BMC genomics*, 13(1) :504, 2012.
- [249] Tim J Stevens, David Lando, Srinjan Basu, Liam P Atkinson, Yang Cao, Steven F Lee, Martin Leeb, Kai J Wohlfahrt, Wayne Boucher, Aoife O'Shaughnessy-Kirwan, et al. 3d structures of individual mammalian genomes studied by single-cell hi-c. *Nature*, 544(7648) :59, 2017.
- [250] Gary D Stormo. [13] consensus patterns in dna. 1990.
- [251] H Su, T Xu, S Ganapathy, M Shadfan, M Long, T HM Huang, I Thompson, and ZM Yuan. Elevated snorna biogenesis is essential in breast cancer. *Oncogene*, 33(11) :1348, 2014.
- [252] Ming Su, Dali Han, Jerome Boyd-Kirkup, Xiaoming Yu, and Jing-Dong J Han. Evolution of alu elements toward enhancers. *Cell reports*, 7(2) :376–385, 2014.
- [253] Ye Su, Haijiang Wu, Alexander Pavlosky, Ling-Lin Zou, Xinna Deng, Zhu-Xu Zhang, and Anthony M Jevnikar. Regulatory non-coding rna : new instruments in the orchestration of cell death. *Cell death & disease*, 7(8) :e2333, 2016.
- [254] Miho M Suzuki and Adrian Bird. Dna methylation landscapes : provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6) :465, 2008.
- [255] Orsolya Symmons, Leslie Pan, Silvia Remeseiro, Tugce Aktas, Felix Klein, Wolfgang Huber, and François Spitz. The shh topological domain facilitates the action of remote enhancers by reducing the effects of genomic distances. *Developmental cell*, 39(5) :529–543, 2016.
- [256] Orsolya Symmons, Veli Vural Uslu, Taro Tsujimura, Sandra Ruf, Sonya Nassari, Wibke Schwarzer, Laurence Ettwiller, and François Spitz. Functional and topological characteristics of mammalian regulatory domains. *Genome research*, 2014.
- [257] Mingxiang Teng and Rafael A Irizarry. Accounting for gc-content bias reduces systematic errors and batch effects in chip-seq data. *Genome research*, 2017.
- [258] F Thoma, Th Koller, and A Klug. Involvement of histone h1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. *The Journal of cell biology*, 83(2) :403–427, 1979.
- [259] Bin Tian, Jun Hu, Haibo Zhang, and Carol S Lutz. A large-scale analysis of mrna polyadenylation of human and mouse genes. *Nucleic acids research*, 33(1) :201–212, 2005.
- [260] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

- [261] Anne-Laure Todeschini, Adrien Georges, and Reiner A Veitia. Transcription factors : specific dna binding and specific gene regulation. *Trends in genetics*, 30(6) :211–219, 2014.
- [262] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. The cancer genome atlas (tcga) : an immeasurable source of knowledge. *Contemporary oncology*, 19(1A) :A68, 2015.
- [263] Yusuf Tutar. Pseudogenes. *Comparative and functional genomics*, 2012, 2012.
- [264] Barbara Uszczynska-Ratajczak, Julien Lagarde, Adam Frankish, Roderic Guigó, and Rory Johnson. Towards a complete map of the human long non-coding rna transcriptome. *resource*, 8(67) :276, 2018.
- [265] Patricia Valencia, Anusha P Dias, and Robin Reed. Splicing promotes rapid and efficient mrna export in mammalian cells. *Proceedings of the National Academy of Sciences*, 105(9) :3386–3391, 2008.
- [266] Lourdes Valenzuela and Rohinton T Kamakaka. Chromatin insulators. *Annu. Rev. Genet.*, 40 :107–138, 2006.
- [267] Bas van Steensel and Andrew S Belmont. Lamina-associated domains : links with chromosome architecture, heterochromatin, and gene repression. *Cell*, 169(5) :780–791, 2017.
- [268] Bas van Steensel and Steven Henikoff. Identification of in vivo dna targets of chromatin proteins using tethered dam methyltransferase. *Nature biotechnology*, 18(4) :424, 2000.
- [269] Jimmy Vandel, Oceane Cassan, Sophie Lebre, Charles-Henri Lecellier, and Laurent Brehelin. Modeling transcription factor combinatorics in promoters and enhancers. *bioRxiv*, page 197418, 2017.
- [270] Artur Veloso, Killeen S Kirkconnell, Brian Magnuson, Benjamin Biewen, Michelle T Paulsen, Thomas E Wilson, and Mats Ljungman. Rate of elongation by rna polymerase ii is associated with specific gene features and epigenetic modifications. *Genome research*, 2014.
- [271] Ciira wa Maina, Antti Honkela, Filomena Matarese, Korbinian Grote, Hendrik G Stunnenberg, George Reid, Neil D Lawrence, and Magnus Rattray. Inference of rna polymerase ii transcription dynamics from chromatin immunoprecipitation time course data. *PLoS computational biology*, 10(5) :e1003598, 2014.
- [272] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221) :470, 2008.
- [273] YX Rachel Wang and Haiyan Huang. Review on statistical methods for gene network reconstruction using expression data. *Journal of theoretical biology*, 362 :53–61, 2014.
- [274] Wyeth W Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4) :276, 2004.
- [275] Matthew T Weirauch, Ally Yang, Mihai Albu, Atina G Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S Najafabadi, Samuel A Lambert, Ishminder Mann, Kate Cook, et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, 158(6) :1431–1443, 2014.
- [276] John W Whitaker, Zhao Chen, and Wei Wang. Predicting the human epigenome from dna motifs. *Nature methods*, 12(3) :265–272, 2015.
- [277] Laurens G Wlming, James GR Gilbert, Kerstin Howe, S Trevanion, T Hubbard, and Jennifer L Harrow. The vertebrate genome annotation (vega) database. *Nucleic acids research*, 36(suppl\_1) :D753–D760, 2007.
- [278] Edgar Wingender, Xin Chen, Reinhard Hehl, Holger Karas, Ines Liebich, V Matys, T Meinhardt, M Prüß, Ingmar Reuter, and Frank Schacherer. Transfac : an integrated system for gene expression regulation. *Nucleic acids research*, 28(1) :316–319, 2000.
- [279] CL Woodcock, L-LY Frado, and JB Rattner. The higher-order structure of chromatin : evidence for a helical ribbon arrangement. *The Journal of cell biology*, 99(1) :42–52, 1984.
- [280] Gordana Wutz, Csilla Várnai, Kota Nagasaka, David A Cisneros, Roman R Stocsits, Wen Tang, Stefan Schoenfelder, Gregor Jessberger, Matthias Muhar, M Julius Hossain, et al. Topologically associating domains and chromatin loops depend on cohesin and are regulated by ctcf, wapl, and pds5 proteins. *The EMBO journal*, 36(24) :3573–3599, 2017.
- [281] Yao Xue and Nilanjan Ray. Cell detection with deep convolutional neural network and compressed sensing. *arXiv preprint arXiv :1708.03307*, 2017.
- [282] J Omar Yáñez-Cuna and Bas van Steensel. Genome–nuclear lamina interactions : from cell populations to single cells. *Current opinion in genetics & development*, 43 :67–72, 2017.
- [283] Chuhu Yang, Eugene Bolotin, Tao Jiang, Frances M Sladek, and Ernest Martinez. Prevalence of the initiator over the tata box in human and yeast genes and identification of dna motifs enriched in human tata-less core promoters. *Gene*, 389(1) :52–65, 2007.

- [284] Guo-Cheng Yuan. Targeted recruitment of histone modifications in humans predicted by genomic sequences. *Journal of Computational Biology*, 16(2) :341–355, 2009.
- [285] Yuan Yuan, Lei Guo, Lei Shen, and Jun S Liu. Predicting gene expression from sequence : a reexamination. *PLoS computational biology*, 3(11) :e243, 2007.
- [286] Kenneth S Zaret and Jason S Carroll. Pioneer transcription factors : establishing competence for gene expression. *Genes & development*, 25(21) :2227–2241, 2011.
- [287] Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, et al. Ensembl 2018. *Nucleic acids research*, 46(D1) :D754–D761, 2017.
- [288] MQ Zhang. Statistical features of human exons and their flanking regions. *Human molecular genetics*, 7(5) :919–932, 1998.
- [289] Tianyi Zhang, Sarah Cooper, and Neil Brockdorff. The interplay of histone modifications—writers that read. *EMBO reports*, 16(11) :1467–1481, 2015.
- [290] Jian Zhou, Chandra L Theesfeld, Kevin Yao, Kathleen M Chen, Aaron K Wong, and Olga G Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics*, 50(8) :1171, 2018.
- [291] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods*, 12(10) :931, 2015.
- [292] Tianyin Zhou, Lin Yang, Yan Lu, Iris Dror, Ana Carolina Dantas Machado, Tahereh Ghane, Rosa Di Felice, and Remo Rohs. Dnashape : a method for the high-throughput prediction of dna structural features on a genomic scale. *Nucleic acids research*, 41(W1) :W56–W62, 2013.
- [293] Christian Zorn, Christoph Cremer, Thomas Cremer, and Jürgen Zimmer. Unscheduled dna synthesis after partial uv irradiation of the cell nucleus : distribution in interphase and metaphase. *Experimental cell research*, 124(1) :111–119, 1979.
- [294] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 67(2) :301–320, 2005.