Structure de la chromatine et éléments transposables : une approche évolutive

Jérémy Barbier

Numéro National de Thèse : 2022LYSEN010

THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE LYON

opérée par

l'École Normale Supérieure de Lyon

École doctorale N° 52 :
Physique et Astrophysique de Lyon (PHAST)

Discipline : Physique
Spécialité : Interface Physique-Biologie

Soutenue publiquement le 22 Avril 2022, par :

Jérémy BARBIER

Structure de la chromatine et éléments transposables : une approche évolutive

---

Devant le jury composé de :

BARBI,	Maria	PR	LPTMC, Sorbonne Univ.	Rapporteure
CRISTOFARI,	Gaël	Directeur de recherche	IRCAN - Nice	Rapporteur
CHAMBEYRON,	Séverine	Directrice de recherche	IGH - Montpellier	Examinatrice
RADMAN-LIVAJA,	Marta	Chargée de recherche	IGMM - Montpellier	Examinatrice
VIEIRA,	Cristina	PR	Univ. Lyon 1	Examinatrice
AUDIT,	Benjamin	Directeur de recherche	ENS de Lyon	Directeur de thèse
BRUNET,	Frédéric	Ingénieur de recherche	ENS de Lyon	Co-encadrant

Résumé de la thèse

Un modèle physique de formation des nucléosomes a révélé que sur un tiers du génome humain, la séquence d'ADN code pour des barrières inhibitrices de la formation du nucléosome (NIEBs) bordées de chaque côté par ~2-3 nucléosomes.

Nos résultats corroborent que les NIEBs seraient présent chez tous les eucaryotes. En effet, (i) le modèle prédit des NIEBs aux propriétés identiques aux NIEBs humains pour une large gamme de génomes, (ii) ces NIEBs sont confirmés par des données expérimentales de positionnement du nucléosome chez la souris, la drosophile, le poisson-zèbre et l'arabette et (iii) le profil des mutations humain-chimpanzé associées à la méthylation des dinucléotides CpG est en phase avec l'accessibilité de l'ADN déduite du positionnement des nucléosomes aux bords des NIEBs. Les données expérimentales les plus récentes révèlent des nucléosomes très instables dans les NIEBs, qui pourraient donc agir comme des "portes d'entrée" pour les modifications épigénétiques, ce que nous observons pour le variant d'histone H3.3.

Nous proposons un scénario évolutif chez l'humain et le chimpanzé où l'insertion des éléments transposables (ET) de type Alu est à l'origine de nouveaux NIEBs car (i) 70 % des éléments Alu sont insérés à un bord de NIEB, (ii) les éléments Alu les plus récemment insérés sont les plus proches des NIEBs et (iii) les éléments Alu spécifiques à une espèce sont principalement positionnés aux bords des NIEBs eux aussi spécifiques à cette espèce. L'identification chez la souris et le porc d'ET bons candidats pour suivre le même modèle de création de NIEB par insertion suggère un mécanisme général qui pourrait expliquer le succès évolutif de certaines familles d'ETs. Cette thèse permet donc de mieux appréhender les mécanismes évolutifs responsables de l'organisation nucléosomale intrinsèque à l'échelle des génomes.

Anything that can go wrong will go wrong.

Edward J. Murphy

REMERCIEMENTS

Pour commencer, je ne peux que me tourner vers Benjamin. Merci pour ces quatre ans. Merci pour ton soutien, ton encadrement quotidien, tes conseils précieux, et ton enthousiasme même dans les moments plus difficiles. Merci pour toutes ces discussion, scientifiques ou non-scientifiques, qui m'ont fait progresser en tant que chercheur et en tant que personne. Et enfin merci pour les (quelques) passes décisives pendant nos matchs du vendredi !

Je poursuivrai avec Frédéric. D'abord merci Fred de m'avoir accueilli dès 2018 en stage, et fait confiance du premier au dernier jour. Merci pour ton engouement pour mes résultats et la mise en avant dont tu m'as fait profiter. Enfin, merci pour ta bonne humeur et ton immuable optimisme.

Merci au coach Jean-Nicolas, rigoureux entraîneur plus Bielsa que Tuchel, plus Mourinho qu'Ancelotti. Merci pour la liberté scientifique que tu m'as laissée durant ces quatre ans. Ton exigence est ce qui a amené à cette thèse, qui je l'espère aura fait de moi un meilleur biologiste.

Merci également aux autres membres du projet Chromagnon. Merci à toi Cédric pour ta bonne humeur contagieuse et ton enthousiasme débordant. Merci à Kiran et Kévin, pour les données auxquelles vous m'avez permis d'accéder ainsi que pour les discussions scientifiques constructives auxquelles ce projet a donné lieu. Et surtout bon courage à Fabien pour prendre la suite !

La recherche est un travail collaboratif, et cette thèse à cheval sur deux laboratoires en est une excellente illustration. Je remercie donc à la fois les équipes Volff et SiSyPhe. Au LP, je souhaite remercier chaleureusement Hadi et Jean-Michel, qui m'ont accueilli dans leur bureau à mon arrivée en thèse. Merci également à mes co-bureaux actuels, Salambô, Eric et Stéphane, pour l'ambiance agréable qui règne au M7.107. A l'IGFL, merci à Coco, Mag, Fabien et Laure pour la bonne ambiance quotidienne au sein de l'équipe. Merci également aux membres plus éphémères : Sho, Emilie, Candice et Théo, ainsi qu'à un illustre ancien : Thibault. Enfin, merci à toi Sara, j'ai au moins autant appris que toi pendant l'encadrement de tes stages.

Plus généralement, je remercie l'ensemble de l'IGFL et du Laboratoire de Physique, pour le cadre et l'ambiance de travail, ainsi que pour les retours constructifs dont j'ai pu bénéficier. Un merci tout particulier aux équipes administratives pour votre gentillesse et surtout votre patience à mon égard. Merci également à Camille, Béryl et Jessika, avec qui j'ai eu la (presque) lourde tâche de représenter les docs/post-docs au conseil d'unité.

Merci aux camarades doctorants (ou non) : Augustin, Théodore, Amélie, Jonathan, Cindy, Camille, Houssam, Lies, Juliette, Jessika et j'en passe, pour le soutien mutuel et les happy hours. Merci également aux footeux, biologistes comme physiciens, Vincent, Sylvain, Abdou, Samir, David, Amélie, Stéphane et tous les autres, et évidemment un merci spécial aux organisateurs Angel et Lies !

Je souhaite également remercier Jean-Baptiste et Fabien, pour leurs encouragements et précieux conseils dispensés durant mon comité de suivi. Et évidemment l'ensemble des membres de mon jury de thèse.

Pour s'éloigner (mais pas trop encore) du laboratoire, j'aimerai remercier Ivan Gentil, qui m'a donné l'opportunité d'enseigner à l'université. Merci à tous les responsables d'UE qui m'ont épaulé, avec une mention spéciale à Elodie et Yoann.

Pour s'éloigner (vraiment, cette fois) du laboratoire, je tiens à remercier certaines personnes sans qui je ne serai jamais arrivé jusqu'ici. D'abord les fabuleux Coco et Axel, ainsi évidemment que Popo et Max. Personne n'arrive à la cheville des amis de Ioverth. Merci aux gens imaginaires, Adelme, Cécile et Maëlle, tabouret tabouret ! Aux plus anciens : Géraud, Lilian, Régis, Sarah, Manon, sans oublier l'immémorial Jerem. Pour finir, merci aux bonobos mangeurs de cartes graphiques pour les soirées loliennes, valorantes ou autres jeux potamochère-compatibles (presque) civilisées.

Enfin, je souhaite remercier ma famille, tout particulièrement mes parents pour leur soutien indéfectible depuis le début de ces (longues) années d'études.

Une dernière ligne à part pour une personne qui l'est tout autant. Je n'en serai pas là si tu n'étais pas là.
Merci Ema.

Table des matières

Liste des figures

Liste des tableaux

1 Introduction

Sommaire

1.1 Le nucléosome

Totalement dépliées et mises bout-à-bout, les trois milliards de paires de bases contenues dans le génome humain formeraient une molécule de plus de deux mètres de long (McGinty & Tan, 2015). Replier cette molécule pour qu'elle entre dans le noyau de chacune de nos cellules revient à disposer un fil d'environ cinquante kilomètres dans un ballon de basket. Ce formidable repliement est réalisé au moyen de plusieurs étapes d'enroulement de l'ADN aboutissant en métaphase à la formation des chromosomes, ces structures formant plus ou moins un X que l'on peut observer en réalisant le caryotype d'un individu. La première étape de repliement de l'ADN consiste en l'enroulement de ce dernier autour de protéines histones pour former des nucléosomes (Kornberg, 1974). Le lien entre la séquence enroulée autour des histones et le positionnement des nucléosomes dans les génomes est la question centrale de cette thèse.

1.1.1 Qu'est-ce qu'un nucléosome ?

La structure du nucléosome

Les nucléosomes constituent l'unité de base de la chromatine (Kornberg, 1974, 1977). Chaque nucléosome est composé d'un noyau, ou coeur constitué d'un octamère des protéines histones H2A, H2B, H3 et H4, chacune présente en deux exemplaires, et formant un cœur d'histones de forme cylindrique (Luger et al., 1997; McGhee & Felsenfeld, 1980). Autour de ce cœur d'histone s'enroulent 145 à 147 paires de bases (pb) d'ADN (McGinty & Tan, 2015), faisant ~ 1.75 fois le tour du cœur d'histones, ce qui forme le coeur du nucléosome (McGhee & Felsenfeld, 1980). Les nucléosomes sont reliés entre eux par de l'ADN non-enroulé autour des histones, que l'on appelle le "linker ADN", ou simplement "linker". Souvent, ce linker est occupé par une protéine appelée "linker histone". Il en existe plusieurs, notamment H1 et H5 (McGinty & Tan, 2015). L'ensemble constitué du cœur d'histones, d'environ 165 pb d'ADN et d'un linker histone forme ce que l'on appelle le chromatosome (McGhee & Felsenfeld, 1980). Le chromatosome associé au linker ADN forment ce que l'on appelle le nucléosome. Cependant, malgré ces définitions techniques, le coeur du nucléosome (à savoir simplement l'octamère d'histones et les ~ 146 pb d'ADN enroulées autour) est souvent désigné comme le nucléosome. Il en sera de même dans ce manuscrit de thèse.

Un objet commun aux eucaryotes

La structure des nucléosomes composée de deux copies de quatre protéines histones autour desquelles s'enroulent 147 pb d'ADN (**Figure 1.1**) est très conservée chez les eucaryotes, de la levure aux métazoaires, notamment via la conservation des protéines histones entre les différentes espèces (Cutter & Hayes, 2015; Luger et al., 1997). Des structures nucléosomales particulières ont également été découvertes chez les archées, où l'ADN est enroulé autour d'un tétramère de protéines homologues aux histones H3 et H4 (Ammar et al., 2012). Le repliement des génomes en nucléosomes semble donc être un mécanisme largement utilisé. Ce succès évolutif des nucléosomes est également du au nombre important de fonctions qu'ils peuvent avoir en plus de celle de replier les génomes.

FIGURE 1.1 – **Structure du cœur du nucléosome**. L'ADN est représenté en marron et turquoise. Les 8 protéines histones sont représentées en bleu (H3), vert (H4), jaune (H2A) et rouge (H2B). La partie gauche représente une vue "du dessus". La partie droite est une rotation à 90° de la gauche selon un axe vertical. © Luger et al. (1997)

1.1.2 Un objet aux fonctions multiples, élément clé de la régulation des processus génomiques

Plusieurs fonctions sont associées au nucléosome, ce qui en fait un élément capital de la vie de la cellule. Tout d'abord, et comme expliqué précédemment, le nucléosome est le premier niveau de repliement de l'ADN, permettant à 146 pb de former autour de l'octamère d'histone une unité génomique de 2.8 Å (Luger et al., 1997). Le nucléosome est également associé à des fonctions régulatrices des différents mécanismes génomiques tels que la transcription ou la réplication, à la fois en contrôlant l'accès à la séquence d'ADN et en servant de substrat pour un ensemble d'enzymes apportant des modifications post-traductionnelles (MPT) aux protéines histones. Cela permet à la fois d'ajuster l'accessibilité aux séquences et le niveau de la compaction de la chromatine (McGinty & Tan, 2015). En effet, le troisième aspect fonctionnel du nucléosome est sa capacité à s'assembler à d'autres nucléosomes pour former des fibres de chromatine plus épaisses telle que la fibre 30 nm (Barbi et al., 2014; Robinson & Rhodes, 2006), bien que l'existence de cette fibre *in vivo* soit encore à ce jour sujette à débat.

L'accessibilité au génome peut être régulée à grande échelle en modifiant le degré de compaction de la chromatine. Localement, l'accès à la séquence par des facteurs se fixant à l'ADN est dépendant du positionnement des nucléosomes. En effet, beaucoup de facteurs de transcription ne peuvent se fixer à l'ADN qu'en l'absence de nucléosome (McGinty & Tan, 2015; Onufriev & Schiessel, 2019). L'occurrence ou non de nucléosomes peut donc être d'une importance capitale à certains loci comme les séquences régulatrices de gènes telles que les promoteurs et enhancers. Par exemple,

chez la levure, il a été observé que le passage d'un métabolisme majoritairement aérobie à un métabolisme majoritairement anaérobie est associé à des patrons spécifiques de présence/absence de nucléosomes au promoteur des gènes associés à la respiration et la fermentation (Field et al., 2009; Tsankov et al., 2010). Dans ces articles, les auteurs montrent que le nucléosome, en diminuant l'accessibilité à la séquence, est un élément régulateur de l'expression des gènes.

La diminution d'accessibilité à l'ADN lorsqu'il est enroulé autour des protéines histones permet également de protéger la séquence de mutagènes potentiels (Barbier et al., 2021). En effet, certains mécanismes mutationnels sont enrayés par la présence d'un nucléosome. Par exemple, les séquences inter-nucléosomales sont plus sujettes aux substitutions nucléotidiques de C vers T que les séquences nucléosomales (Chen et al., 2012). Ce type de mutations résulte souvent d'une désamination de la cytosine, or ce mécanisme mutationnel est plus efficace si la double hélice d'ADN est localement ouverte. La formation d'un nucléosome impose des contraintes structurelles à l'ADN qui inhibent fortement cette ouverture, inhibant dans le même temps le mécanisme mutationnel dans l'ADN nucléosomal (Makova & Hardison, 2015). Le nucléosome a donc également des vertus protectrices de l'ADN. Cependant, d'autres types de mutations peuvent également être favorisées dans l'ADN nucléosomal par rapport aux linkers, notamment car la présence d'un nucléosome diminue l'accessibilité à l'ADN pour les mécanismes de réparation des mutations (Wu et al., 2018). L'interaction complexe entre présence d'un nucléosome et patrons de mutations est détaillée dans la revue bibliographique publiée pendant cette thèse et présentée en **partie 1.1.3**.

La présence d'un nucléosome peut également être nécessaire à la fixation de facteurs chromatiniens, particulièrement ceux entraînant des modifications post-traductionnelles (MPT) des protéines histones (McGinty & Tan, 2015). Ces modifications, selon leur type, peuvent être associées à des structures chromatiniennes fermées, inaccessibles et répressives, ou également à des structures ouvertes et accessibles favorisant l'expression des gènes. Il existe une très grande variété de MPT, dont les diverses combinaisons permettent d'ajuster au besoin l'état chromatinien des génomes (Taylor & Young, 2021). Un nucléosome peut également être modifié par le remplacement de protéines histones canoniques par des variants de ces protéines, pour former des nucléosomes ayant des propriétés différentes. Un exemple observé dans plusieurs espèces comme la drosophile, la souris ou encore l'humain est celui du variant H3.3, qui a d'abord été associé à des régions de chromatine actives favorisant l'expression des gènes environnants, avant d'être également retrouvé dans des régions où l'expression est réprimée (Szenker et al., 2011). Le nucléosome est donc un élément de base de l'épigénétique, à savoir les changements d'activité génomique qui ne sont pas directement encodés dans la séquence d'ADN. Ces changements ne sont généralement pas transmis à la descendance, cependant certains nucléosomes ont été observés comme retenus à certains loci dans les cellules germinales. Cette rétention pourrait permettre la transmission de marques épigénétiques à la descendance (Ben Maamar et al., 2020).

Le nucléosome est donc une structure dynamique, dont les protéines histones peuvent être modifiées ou remplacées, mais qui peut aussi être déplacé le long du génome par des complexes de remodelage de la chromatine, ou bien désassemblé, partiellement ou complètement, pour permettre l'accès à l'ADN nucléosomal (Onufriev & Schiessel, 2019). Cette dynamique du nucléosome est cruciale pour les mécanismes cellulaires, car elle permet de compacter le génome et de le décompacter localement ou globalement. Cela afin de permettre la mise en place de mécanismes

cellulaires locaux, comme la transcription, ou globaux, comme la réplication. Le positionnement des nucléosomes, en régulant l'accès à l'ADN et en servant de base aux modifications épigénétiques, a donc une importance cruciale pour la vie de la cellule. Plusieurs facteurs peuvent influencer le positionnement nucléosomal. C'est le cas de la compétition entre les histones et les autres protéines liées à l'ADN (DNA binding protein) telles que les facteurs de transcription, et du remodelage pouvant déplacer ou désassembler des nucléosomes. Le positionnement nucléosomal est également influencé par des effets liés à la séquence (Segal & Widom, 2009b; Struhl & Segal, 2013). Les effets de séquence sont étudiés en détail dans ce manuscrit, avec pour objectif de comprendre dans quelle mesure ils sont liées à l'évolution de la séquence d'ADN et l'évolution de la structure chromatinienne.

1.1.3 Une relation étroite entre nucléosome et séquence génomique

La séquence est un facteur déterminant du positionnement nucléosomal, dans le sens où elle peut, selon sa composition, favoriser ou inhiber la formation du nucléosome. Dans le cadre de cette thèse, une revue de la littérature sur ce sujet a été effectuée et publiée dans le journal Genes (Barbier et al., 2021). Dans cette revue sont détaillés les mécanismes de positionnement nucléosomal liés à la séquence d'ADN. Y sont étudiés également les effets de l'évolution des séquences sur la position des nucléosomes. Enfin, la dernière partie de cette revue est consacrée à l'influence des nucléosomes sur l'évolution des séquences, particulièrement en terme de patrons de mutations.

genes

MDPI

Review

Coupling between Sequence-Mediated Nucleosome Organization and Genome Evolution

Jérémy Barbier [1,2], Cédric Vaillant [2], Jean-Nicolas Volff [1,*], Frédéric G. Brunet [1] and Benjamin Audit [2,*]

[1] Institut de Génomique Fonctionnelle de Lyon, Univ Lyon, CNRS UMR 5242, Ecole Normale Supérieure de Lyon, Univ Claude Bernard Lyon 1, F-69364 Lyon, France; jeremy.barbier@ens-lyon.fr (J.B.); frederic.brunet@ens-lyon.fr (F.G.B.)

[2] Laboratoire de Physique, Univ Lyon, ENS de Lyon, CNRS, F-69342 Lyon, France; cedric.vaillant@ens-lyon.fr

* Correspondence: jean-nicolas.volff@ens-lyon.fr (J.-N.V.); benjamin.audit@ens-lyon.fr (B.A.)

Abstract: The nucleosome is a major modulator of DNA accessibility to other cellular factors. Nucleosome positioning has a critical importance in regulating cell processes such as transcription, replication, recombination or DNA repair. The DNA sequence has an influence on the position of nucleosomes on genomes, although other factors are also implicated, such as ATP-dependent remodelers or competition of the nucleosome with DNA binding proteins. Different sequence motifs can promote or inhibit the nucleosome formation, thus influencing the accessibility to the DNA. Sequence-encoded nucleosome positioning having functional consequences on cell processes can then be selected or counter-selected during evolution. We review the interplay between sequence evolution and nucleosome positioning evolution. We first focus on the different ways to encode nucleosome positions in the DNA sequence, and to which extent these mechanisms are responsible of genome-wide nucleosome positioning in vivo. Then, we discuss the findings about selection of sequences for their nucleosomal properties. Finally, we illustrate how the nucleosome can directly influence sequence evolution through its interactions with DNA damage and repair mechanisms. This review aims to provide an overview of the mutual influence of sequence evolution and nucleosome positioning evolution, possibly leading to complex evolutionary dynamics.

Keywords: DNA sequence-encoded nucleosome ordering; nucleosome depleted regions; DNA sequence mutation; chromatin evolution

Citation: Barbier, J.; Vaillant, C.; Volff, J.-N.; Brunet, F.G.; Audit, B. Coupling between Sequence-Mediated Nucleosome Organization and Genome Evolution. *Genes* **2021**, *12*, 851. https://doi.org/10.3390/genes12060851

Academic Editor: Manuel A. Garrido-Ramos

Received: 1 May 2021
Accepted: 27 May 2021
Published: 1 June 2021

1. Introduction

To fit in the nucleus of each cell, eukaryotic DNA needs to be highly compacted. This compaction is achieved by the formation of a protein-DNA complex called chromatin [1]. The first level of compaction consists of the wrapping of ∼146 bp of DNA around an octamer of four core histone proteins (H2A, H2B, H3 and H4), forming a nucleosome [2]. In the nucleosome, the DNA is wrapped almost twice around the core histone octamer (a tetramer of $(H3\text{-}H4)_2$ flanked by two dimers of H2A-H2B), with contact points between DNA and the histone proteins every ∼10 bp [3,4]. The mid-point of the complexed DNA is called the dyad, and serves as a reference to specify nucleosome positions. The nucleosome repeat length (NRL), that represents the distance between two consecutive nucleosome dyads, ranges from 155 bp in fission yeast [5] to about 240 bp in echinoderm sperm [6]. Taking into account the length of DNA wrapped in each nucleosomes, there is thus a high density of nucleosome in living cells regardless of the cell type or organism, with at least two third of the genome participating in a nucleosome. Nucleosomes come in several forms. Core histones may carry post-translational modifications (PTMs), such as methylation, acetylation or phosphorylation occurring mostly in the N-terminal tail of histones (e.g., tri-methylation of histone H3 lysine 9, also known as H3K9me3). Histone cores may also contain histone variants, which are alternative histone proteins encoded by genes that appeared throughout the evolution of Eukaryotes [4,7,8]. PTMs and histone variants are

associated with different chromatin states of genome compaction and genome regulation and have thus received most of the attention in chromatin biology studies. Nevertheless, the precise position of nucleosomes on the DNA is also of great importance [1]. Indeed, the accessibility of DNA to non-histone chromatin factors like transcription and replication factors is modulated by nucleosome occupancy, with nucleosomal DNA being considerably less accessible to theses factors than the naked "linker" DNA between nucleosomes. From a collective perspective, the position of nucleosomes relative to each other is also associated to chromatin state, probably in relation to higher order chromatin compaction. Indeed, actively transcribed genomes where chromatin needs to be open and accessible tend to have shorter NRL (ranging from 160 to 189 bp in yeast, embryonic stem cells and tumour cells for example) than transcriptionaly inactive genomes (NRL ranging from 190 to 240 in chicken erythrocytes and echinoderm sperm for example) [9]. This distinction has also been made within the human genome, where the NRL of active genes is way shorter (178 bp) than the NRL of repressed or heterochromatic non-coding sequences (206 bp) [10]. However, there are exceptions to this rule. For example, in higher eukaryotes, telomeric DNA is packaged in nucleosomes with a NRL 20–40 bp shorter than the NRL of bulk nucleosome [11]. This has been observed in vertebrates [12–16] but also in sea urchin [16], and several plant species [17–19]. The position of nucleosomes on the DNA and relative to each other is thus crucial for genetic functions, because it modulates the efficiency of trans-acting factors such as the transcription machinery [1,2,20]. Nucleosomal positioning on DNA depends on various factors, including DNA sequence effects, competition for DNA such as with transcription factors, and remodeling by ATP-dependent enzyme [21]. Notably, the DNA sequence has an important contribution to nucleosomal positioning at the genome scale [1,10,21,22]. Nucleosome positions are thus to some significant extend a sequence-encoded feature that have a functional role in genomes (as modulator of the accessibility to DNA). As other sequence-encoded functional features (such as genes), nucleosome positions can then be selected during evolution. In other words, sequences could be selected not for their direct coding properties as genes, but for their abilities to favor or impair nucleosome formation at specific loci, directly impacting their accessibility to external regulatory factors. Selection of sequences for their nucleosomal affinity has been described in several species such as yeasts [23,24] but also in more complex organisms like maize [25,26] or human [27–29]. Note that the repositioning of nucleosomes according to the evolution of sequences can also occur in a neutral scenario leading to possible drifts of nucleosome positions [30]. Interestingly, the nucleosome itself also shapes the evolution of sequences by interacting with DNA damage and repair mechanisms, leading to biased mutational patterns inside and around nucleosomes [31]. Here, we will review some of the findings about these mechanisms, focusing first on how nucleosome positions are encoded in the DNA sequence, then on how sequence nucleosomal properties have been selected during evolution, and finally, on how the nucleosome directly modulates mutational patterns. This provides an opportunity to discuss the mutual feedback between the evolution of DNA sequence and chromatin organization at a genomic scale.

2. How Is Nucleosome Positioning Encoded in the DNA Sequence?

2.1. DNA Sequence Does Influence Nucleosome Positioning

Using SELEX (Systematic Evolution of Ligands by EXponential enrichment) experiments on synthetic and genomic DNA with the core histone proteins as ligands, it was shown that the DNA sequence does influence the affinity of a DNA fragment for histones up to a 5000-fold range [32–35]. In such experiments, an excess of DNA fragments of variable sequence compete for a ligand. The DNA-ligand complexes are then extracted, DNA fragments are purified, amplified and brought back into competition with the same ligand, a process repeated several times to purify sequences with the highest affinities for the ligand of interest. Lowary and Widom used this approach with synthetic DNA fragments and core histone proteins as ligands to select from a random set of sequences the ones with the highest affinities for the nucleosome [33]. It revealed the existence of sequences with un-

expectedly high affinity for the histone octamer. Similar experiments were also performed with fragments extracted from genomic DNA. It showed that their affinity for histones had a much narrower range than random DNA fragments [32,35]. These experiments clearly indicate that the DNA sequence matters on how easily a nucleosome can be formed and so where nucleosome are intrinsically positioned along chromosomes. The sequence-encoded nucleosome positioning can therefore be seen as a basal "ground state" that can be "remodeled" in vivo by the site-specific recruitment and (energy consuming) action of trans-acting factors to establish at proper times and positions an "epigenetic" reversible nucleosome positioning pattern, either permissive or repressive for genome activity. As demonstrated by Parmar et al. [36] when considering a composite model of nucleosome positioning that accounts for both sequence effects and ATP-dependent remodelers and as evidenced by experiments [37], sequence effects are indeed sufficiently strong to control the first steps of the relaxation dynamics of the nucleosomal array after strong perturbation, i.e., in a transient phase of non or weak activity of remodelers. Strikingly, nucleosomal pattern in germ cells where remodelers activity is reduced has been shown to be mostly controlled by the DNA sequence [38]. in vitro nucleosome reconstitution experiments on the yeast genome further demonstrated that ATP was required to obtain a nucleosome positioning pattern that deviate from the sequence encoded pattern and resemble the native pattern [39]. All these results suggest that the primary sequence is a parameter that needs to be taken into account in nucleosome positioning studies, even if sequence effects can be refined or even overridden in vivo by other factors such as ATP-dependent remodelers.

Technical progresses made it possible to decipher DNA sequence-mediated effects genome-wide, mainly with experiments such as MNase-seq, in which the chromatin is digested with an enzyme (the micrococcal nuclease, MNase) that cuts and digests the naked linker DNA between nucleosomes [40–42]. After histone removal, the remaining DNA can be sequenced with high-throughput sequencing techniques, and the alignment of the reads on the reference genome provides information about the genome-wide positioning of nucleosomes [10,42–47]. Such genome-wide mapping of nucleosomes has been established in vivo in various species, including yeast [43,47,48], human [10,44,45], fly [49], plants [25,50], mouse [51], and the nematode *Caernorhabditis elegans* [52], but also in vitro [10,53]. The availability of such experimental data has been reviewed by Teif [54]. Comparison of in vivo and in vitro nucleosome maps revealed a high consistency between in vitro and in vivo genome-wide positioning of nucleosomes [10,53,55]. These results showed that the sequence effects are relevant even in vivo in the presence of external factors influencing nucleosomal positioning . Indeed, the sequence-directed nucleosome positioning is directly observed from in vitro data, because chromatin is reconstituted from DNA and histones only, without any other external factors such as remodelers found in vivo. Accordingly, models established from in vitro genome-wide reconstitution of chromatin predict rather well in vivo nucleosome positioning [22,53,55–60], corroborating the hypothesis that the DNA sequence plays a major role among the different factors influencing the position of nucleosomes [61]. During the past 40 years, attempts to describe the sequence-directed nucleosomal positioning showed that one needs to consider two types of mechanisms (Figure 1): (i) positioning mechanisms where DNA motifs at specific location accommodate DNA wrapping in the nucleosome, for example by favoring certain dinucleotides at contact points between DNA and histones; and (ii) inhibiting mechanisms, with sequences such as poly(dA:dT) preventing nucleosome formation [1].

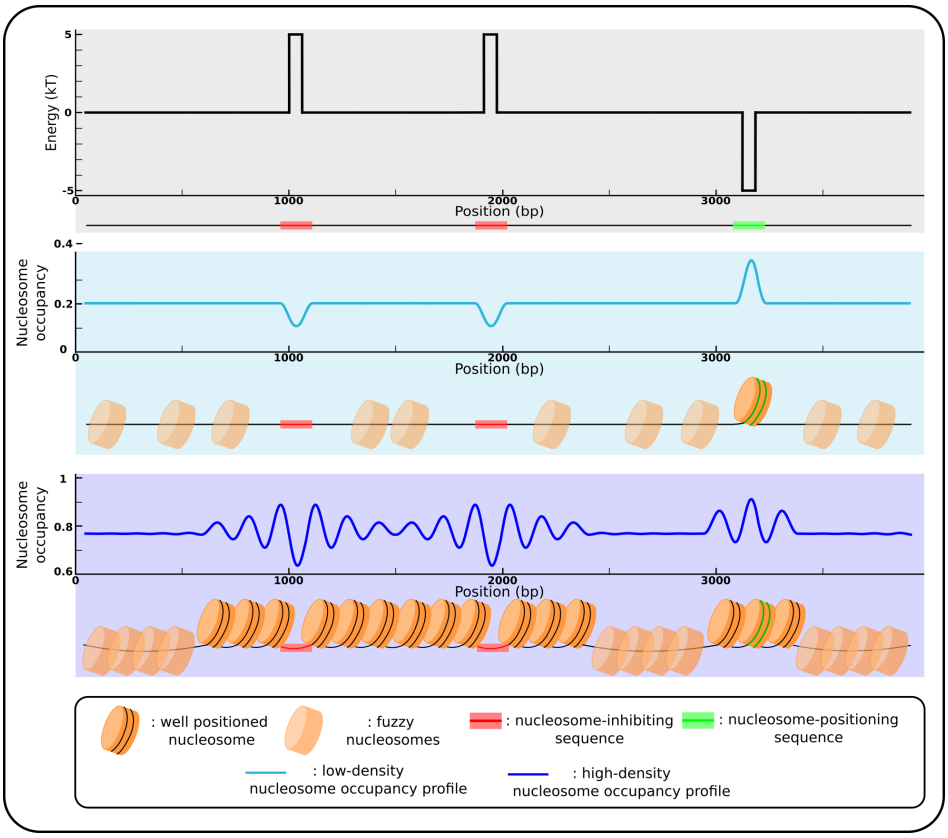


Figure 1. Nucleosomal positioning by sequence motifs. (**Top** panel) Landscape of the energy needed to bend the DNA fragment into the nucleosome depending on the sequence around the nucleosome dyad position (*x*-axis); the hypothetical landscape present two high energy peaks corresponding to two nucleosome-inhibiting sequence motifs (red), and a low energy well at a nucleosome-positioning sequence motif (green). (Mid panel) Nucleosome occupancy profile corresponding to the energy landscape for a low density of nucleosomes. Nucleosomes tend to avoid the inhibiting sequences (as represented by the minima in sky blue curve), and the only preferential nucleosome localisation is at the positioning sequence (peak in sky blue curve). (**Bottom** panel) Nucleosome occupancy profile to corresponding the energy landscape for a high nucleosome density. In this case, nucleosomes still avoid inhibiting sequences (minima in the blue curve), and a global positioning appears between and beside these nucleosomal barriers (oscillations in the blue curve), as a "parking" phenomenon resulting from the non overlapping property of nucleosomes (statistical positioning) (see Section 2.3). Nucleosome positioning also appears beside the well-positioned nucleosome formed on positioning sequence, according to the "anchor-positioning" model described in Section 2.3. Transparent nucleosomes represent fuzzy positioning, meaning that at these loci, nucleosomes have no preferential locations and can be formed more or less anywhere on the DNA.

2.2. Sequence Motifs with 10 Base Pair Periodicity as Nucleosome Positioning Signals

In the 1980s, the analysis of 32 coding and non-coding sequences (representing about 36,000 nucleotides) that were known to fold in chromatin-like structures (i.e., nucleosomes) exhibited a periodicity of ∼10.5 base pair (bp) in the distribution of dinucleotides along their sequences [62]. Dinucleotides GG, TA, TG and TT were found to be the strongest contributors to this observed periodicity. In other words, in sequences that fold in chromatin-like structures, dinucleotides GG, TA, TG and TT tend to be regularly spaced by 10 or 11 bp whereas other dinucleotides are more randomly positioned. Interestingly, no 10.5 bp periodicity was found for prokaryotic sequences. Further analysis showed a symmetry in the phasing of the preferential positionning of complementary dinucleotides within the 10.5 bp periodicity [63]. An explanation proposed for these observations was about the affinity of the DNA sequence for histone core. It was suggested that sequence periodicity

and their symmetries facilitates the bending of the DNA molecule around the nucleosome core histones proteins [62,63]. It was even expected that it would be possible to predict nucleosome positioning from these sequence properties.

The "periodicity model" successfully predicted the curved shape of a 423 bp DNA restriction fragment containing a strong periodicity of AA and TT dinucleotides [64]. Sequence-encoded bending of DNA was explored in several studies [65–68], from which nucleosomal DNA bending tables were derived. In the nucleosome, A/T-rich sequences are preferred where the minor groove is facing inward, and G/C-rich sequences where it is facing outward of the structure [67]. In addition, homopolymers tend to be excluded from the nucleosome, especially from the dyad position [65–68]. Finally, it was observed that linker DNA regions between nucleosomes are cut poorly by DNAse I enzyme, that is known to cut poorly in homopolymers, probably revealing their strong occurrence in linker DNA [68], in accordance with the previous observation.

The sequence periodicities described here facilitate the bending of DNA around the histone octamer to form a nucleosome. Such sequences could have a positioning effect. During the course of evolution, some selective pressure could have acted on genomes to select those sequences at specific loci where the presence of a nucleosome is necessary. Periodicities associated to nucleosomal sequences have been found in several species, in chicken, but also in yeast, human and worm [53,56,61,68–70]. However, among genomic sequences, even the most powerful positioning sequences only have a weak positioning power [33]. Sequences optimized for wrapping into the nucleosome, like the sequence of the clone 601 established by Lowary and Widom in their SELEX experiment on artificial DNA [33], are not found in genomic DNA. In addition, the global positioning power of genomic DNA is not much higher than that of random DNA sequences [33]. Thus, positioning sequences and their periodicities in the dinucleotide distributions fail to explain the genome-wide sequence-encoded nucleosomal positioning [33]. However, periodic distribution of sequence motifs is not the only way to encode nucleosome position.

2.3. Sequence-Encoded Nucleosome Depleted Regions and Statistical Positioning

In yeast, it has been showed that promoters are enriched in what are called nucleosome-depleted regions (NDRs) [43]. In several yeast species, these NDRs are found both in vivo and in vitro, indicating that they are directly encoded in the DNA sequence, mainly through poly(dA:dT) sequences that are known to inhibit nucleosome formation [1]. The strength of the depletion depends mainly on the length and purity of the poly(dA:dT) sequence [1], allowing a fine tune regulation of gene expression in yeast [71]. Positioning of nucleosomes can arise from these NDRs, following a statistical positioning model [72,73], where nucleosomes stack against a fixed object (either a NDR or a highly positioned nucleosome) that serves as an anchor, forming an array of positioned nucleosomes (Figure 1). The closer a nucleosome is to the anchor, the better it is positioned. Thus, counter-intuitively, sequence-encoded nucleosome positioning could arise not from positioning sequences but rather from anti-positioning sequences that anchor the position of nucleosomal arrays. In the case of yeast promoters, if NDRs are observed both in vivo and in vitro, arrays of nucleosomes are only observed in vivo, on the side of the transcribed units [74]. In this case, the in vivo nucleosomal organization results from the combination of the sequence effect (mainly specifying the NDRs and probably the +1 nucleosomes) and the ATP-dependent chromatin remodelers (for the ordering of nucleosomes). Another type of arrays of nucleosomes relying only on sequences have been observed in yeast, where nucleosomes are confined between sequence-encoded NDRs when these NDRs are close to one another [55,57]. Indeed, when two NDRs are close enough to each other, constraints appear on the nucleosomal positioning, mainly because of the exclusion interaction between nucleosomes since two nucleosomes cannot superimpose. For example, if two sequence-encoded NDRs are separated by a distance of about 300 bp (∼2 nucleosomes), and one nucleosome is formed between the NDRs, it can be formed quite anywhere along the 300 bp. However, if 2 nucleosomes are formed, taking about 147 bp each, then the possi-

bilities are greatly reduced and preferential positioning appears. Sequence-encoded arrays of nucleosomes can thus result from sequence-encoded NDRs and a high density of nucleosomes. This "statistical positioning between NDRs" model was experimentally validated with atomic force microscopy (AFM) visualization of nucleosome positioning along a DNA fragment bounded by two sequence-encoded NDRs separated by a two-nucleosomes long distance [55,75]. When either one or two nucleosomes were reconstituted on this fragment, single nucleosomes were observed anywhere between the barriers, but as predicted, the position of nucleosome pairs were very constrained.

In human, part of the genome-wide nucleosomal positioning follows this scenario of statistical positioning between NDRs [28,76]. Indeed, a physical model of nucleosome formation based on sequence-dependent bending properties of the DNA double helix revealed about 1.6 million nucleosome-inhibiting energy barriers (NIEBs) along the human genome. These NIEBs correspond to NDRs, both among in vivo and in vitro data. In both conditions, when NIEBs are close enough to each other (about four nucleosomes or less), a constrained positioning of nucleosomes is observed, just as described above in yeast. The in vitro observation indicates that this positioning is not dependent of the action of remodelers, but relies only on the sequence-encoded NIEBs/NDRs and high density of nucleosomes. in vitro map of nucleosomes also showed that a nucleosome-favoring sequence flanked by two nucleosome-deterring sequences can form what is called a "container" site in which a nucleosome is trapped [10]. Taken alone, each of these sequences do not have any significant positioning or anti-positioning power, but taken together, they form a highly positioned nucleosome at a specific locus. These container sites were also found in the in vivo nucleosomes maps, where they can serve as anchors to form nucleosomal arrays by stacking of the other nucleosomes against the well positioned one. The situation is similarly found at the promoters of yeast genome: a fixed object (here, a highly positioned nucleosome, a NDR in yeast) serves as an anchor for regularly spaced nucleosomal arrays. The difference is that the formation of the array is not associated with transcription as in yeast. However, these arrays are also only observed in vivo, indicating that if the anchor is sequence-encoded, the action of remodelers is needed to fluidify the movement of nucleosomes and allow statistical positioning. Note that isolated NIEBs can also serve as anchors: two to three positioned nucleosomes have been observed on their borders in human, both in vivo and in vitro [28,76], illustrating that the "stacking against an anchor" model does not always need the activity of remodelers.

2.4. Predicting Nucleosomal Positioning from Sequences

Nucleosome occupancy encoded in the sequence can presumably be predicted through sequence-based modeling. This was achieved using mainly two types of approaches: bioinformatic models relying on machine learning [22,53,56,58], and physical models relying on energy calculations [55,57,59,60,77]. The general idea of the bioinformatic models is to detect, genome-wide, the sequence features associated with nucleosomal positioning. For example, the model detailed in [53] is based on an in vitro map of yeast nucleosomes. From this map, the sequence preferences for nucleosomes are extracted to establish a probabilistic model that assigns a score to each 147 bp fragment. This score is based on the 5-mers observed along the sequence of the fragment. From the score landscape, and taking into account the impossibility to superimpose two nucleosomes, nucleosomal positioning can be predicted. This approach reproduced well experimental mapping of nucleosomes [53]. A simpler approach has been developed in [22], in which the over 2000 parameters of [53] are reduced down to only 14 parameters. It was even claimed that a model taking into account only the GC content and poly(dA:dT) sequences is sufficient to achieve good predictions of nucleosome occupancy [22]. The GC content is tightly correlated to nucleosome occupancy [27,28]. It was in fact argued that the observation that the genomic GC content of Eukarya is way less variable than that of Bacteria and Archaea corroborates this observation. It was linked to the high level of conservation of histones between organisms, whereas nucleoid-associated proteins are more variable,

possibly allowing wider range for genomic GC content between species [78]. The physical modeling approach was considered independently by different groups [55,57,60,79]. It is based on intrinsic bending properties of the DNA and thus, its ability to be wrapped around histone octamers. The idea is to compute the energy needed to deform all 147 bp DNA fragments from their intrinsic conformation to the helical conformation adopted in the nucleosome, based on tabulated sequence-dependent elastic parameters. This provides an energy landscape for the formation potential of nucleosomes along the genome. The dynamic assembly of histone octamers along the DNA chain is then modeled as a fluid of rods of finite extension (the DNA wrapping length around the octamer), binding and moving in the nucleosome formation potential and respecting the exclusion relationship between nucleosomes. The nucleosome occupancy profile can then be deduced given a temperature and a chemical potential allowing to fix the average nucleosome density to the experimentally determined value. Nucleosome occupancy based on our implementation of the model [55,57] fits well the experimental occupancy data in yeasts, in the nematode *C. elegans* and the fly *D. melanogaster* [55,59,80], and in human [28,76].

3. Nucleosome Positioning during Evolution

3.1. Nucleosome Position as a Darwinian Feature

Nucleosome occupancy influences the binding of transcription factors by controlling the accessibility to DNA [25]. The modulation of nucleosome occupancy is thus a critical feature for gene transcription regulation. Indeed, the distribution of nucleosomes around genes was associated with transcription levels in several species, including yeast [81], human [10,44], mouse [51], drosophila [49], and plants such as the thale cress [50], rice [50] and maize [26]. For example, highly expressed genes are associated with a more pronounced nucleosome depletion at their promoter than lowly expressed genes. The transcriptional changes during cell life processes such as differentiation, reprogramming, stress or even aging are associated with changes in nucleosome occupancy [82–85]. Modifying the nucleosome organization at some loci is thus expected to have either a positive or a negative impact on the fitness of an individual [86]. As nucleosome positions are at least partially sequence-encoded (Section 2), this strongly suggests that natural selection on DNA sequence could have an impact on the nucleosomal positioning. In other words, mutations could be selected or counter-selected, not for their direct effect on coding sequences, but for their influence on the position of nucleosomes at some specific loci, indirectly influencing features under selection such as gene expression. Following this hypothesis, natural selection could favor nucleosome inhibiting sequences where sequences need to be constantly available to transcription factors (at the regulating sequences of constitutive genes for example). It could also favor certain nucleosomal organization on the body of genes according to the basal level of transcription needed. The latter possibility question the compatibility between the nucleosomal and the genetic codes, to allow encoding of both a protein sequence and the nucleosomal organization in the same sequences. This compatibility has been explored by Eslami-Mossallam et al. [87], revealing the possibility of multiplexing genetic and mechanical information along a single sequence. Indeed, it is achievable to change the nucleosomal organization on the body of a gene without changing the protein(s) associated with the gene, thanks to the redundancy of the genetic code [87].

3.2. Nucleosome Positioning and the Evolution of Gene Regulation

In yeast, "growth genes" are identified as genes almost constantly expressed during growth, often associated with the metabolic pathways used in *ideal* growth conditions. In contrast, "stress genes" are genes expressed only in certain specific conditions, for example to respond to an environmental change. At the nucleosomal level, differences have been observed between growth and stress genes. The prediction of the nucleosomal organization at the promoter of these different types of genes in two yeast species, *Candida albicans* and *Saccharomyces cerevisiae*, showed that on average growth genes exhibit an intrinsically open chromatin at their promoter, when stress genes harbor a more closed patterns [23].

The experimental confirmation of the predicted organizations, both in vitro and in vivo, demonstrated that they are encoded directly in both genomes. Thus, in these two yeasts, we have two distinct sequence-encoded nucleosomal patterns associated with the two modes of gene expression. These two species display major metabolism differences when grown in a high glucose environment: *C. albicans* that grows mainly using respirative metabolism is identified as an aerobic yeast, as oppose to *S. cerevisiae* that grows mainly using fermentative metabolism, identified as an anaerobic yeast. From an evolutionary standpoint, orthologous genes associated with respiration are growth genes in the former, that switched to stress genes in the latter during the evolution of yeasts. By comparing the nucleosomal organization at the promoter of these genes in these two species, it was shown that they exhibit an intrinsically open chromatin in *C. albicans*, and a closed chromatin in *S. cerevisiae* [23]. This pattern was also observed in 10 other yeast species for which the nucleosome occupancy was predicted genome-wide from the DNA sequence. These results were confirmed experimentally with the direct comparison of experimental nucleosome positioning and gene expression data in the same 10 yeast species [24]. It showed that gain or loss of poly(dA:dT) tracts are associated with modifications of the nucleosomal organization at several phylogenetic branch points [24]. For example, the promoters of mitochondrial ribosomal protein (mRP) genes have lost their poly-A-like sequences in anaerobic yeasts, changing the chromatin organization on these genes from an open conformation (in aerobic yeasts) to a closed one (in anaerobic yeasts) [23,24]. These experiments show that in the course of yeast evolution, nucleosomes located at the promoter of genes have been repositioned, notably through the modification of the DNA sequence, and it was associated to a major change in yeast metabolisms, such as the switch from an aerobic to an anaerobic metabolism. This is a very good example of sequence selection not acting directly on coding properties, but for their affinity to nucleosomes, allowing a fine tuning of gene regulation from growth expression to stress expression pattern.

A similar dichotomy is present in multi-cellular organisms, such as maize, in the form of constitutive genes that are expressed regardless of the cell type, versus tissue-specific genes that are expressed only in some specific cell types. Sequences selected for nucleosomal positioning have been observed in this species [25,26]. In maize, the expression level between tissues show only minor differences in constitutive genes which contrast with tissue-specific genes that show higher differences. This difference shows that tissue-specific genes have higher transcriptional plasticity than constitutive genes. It was proposed that the sequence-encoded nucleosomal organization of each gene controls its transcriptional plasticity instead of directly its level of expression [25,26]. Indeed, the level of expression can change between cell types and conditions, particularly for tissue-specific genes. If the level of expression was directly sequence-encoded through nucleosomal positioning, transcriptional plasticity could not be achieved, since the gene sequence is the same in each cell and condition. In maize, the prediction from sequences of the nucleosomal organization of different set of genes showed that constitutive genes have the lowest sequence-encoded global nucleosome occupancy, while tissue-specific genes have the highest [26]. Compared to tissue-specific genes, constitutive genes have bigger and stronger NDRs at their transcription start site (TSS) as well as longer distances between both their 5′ NDR and TSS, and their 3′ NDR and transcription termination site. All these predicted features have been confirmed experimentally with MNase experiments. These two types of genes have different nucleosomal organization resulting in different transcriptional plasticity. In maize, it was also observed that the sequence of constitutive genes has a lower GC content than the sequence of tissue-specific genes, both in introns and exons where it is mainly driven by different codon usage. This likely illustrates selective pressures acting on the nucleosome positioning. The redundancy of the genetic code, allowing the multiplexing of genetic and structural informations [87], is used in this species to promote AT-rich codons in constitutive genes and GC-rich codons in tissue-specific genes, to reduce the GC content of the former and raise the GC content of the latter. This leads to differences in maize genes

nucleosomal organization, with a reduced occupancy on constitutive genes, associated with lower transcriptional plasticity. In contrast, the nucleosome occupancy is higher in tissue-specific genes, and associated with higher transcriptional plasticity. This interplay between nucleosome and transcriptional plasticity has also been observed in several other species such as *C. elegans* and *S. cerevisiae*. In *C. elegans*, a time-course of MNase digestion showed that the AT content in the promoter influences nucleosome stability [88]. In this type of experiments, various levels of chromatin digestion are obtained using different concentrations of MNase or different digestion times, providing information about the stability of nucleosomes [52,88,89]. Fragile nucleosomes are identified as nucleosomes only apparent in low-digestion data, as they are more easily destabilized by the MNase than stable nucleosomes [52,88]. Such experiment in *C. elegans* showed that fragile nucleosomes are associated with high AT content of the underlying DNA sequence, and low expression plus high transcriptional plasticity when they are localized at the promoter of genes [52]. In *S. cerevisiae*, it has been shown that genes can be classified according to their nucleosomal organization [55,80,90]. Some genes have a "cristal" nucleosomal organization, with n nucleosomes on the body of the genes and a precise, constant NRL. Others have a "bistable" nucleosomal organization, with the possibility to put n or $n+1$ nucleosomes on the body of the gene, the $n+1$ organization being associated with a higher expression level. These two classes of nucleosomal organization are, like in maize, associated with different transcription plasticity. Indeed, growth genes are associated with "cristal" organization, where stress genes exhibit a "bistable" organization [55,80,90]. Finally, in human, about 70% of promoters are associated to CpG islands (GC rich regions with a CpG dinucleotide content higher than elsewhere on the genome) [91]. These CpG islands have been described to be accessible without the need for ATP-dependent remodeling [92]. This could be due to their DNA sequence inhibiting nucleosome formation, although the well described nucleosome-free region surrounding the TSS of eukarotic genes could also be implicated [93]. All the examples mentioned here show that in a range of organisms, sequence-encoded nucleosomal organization at genes is strongly linked to expression pattern.

Selection of nucleosomal positioning at genes has also been linked to the complexity of organisms (Figure 2) [29,90]. In yeast, the majority of promoter exhibit a NDR, both in vivo and in vitro, indicating that this nucleosomal conformation is encoded directly in the DNA sequence. In contrast, if NDRs can be found in human at the promoter of expressed genes in vivo, it has a rare occurrence in vitro, and sequence-encoded NDRs are typically absent from promoters. In fact, in human, prediction of nucleosomal positioning from sequence showed that promoters are generally occupied by nucleosome attracting regions (NAR), that are the opposite of NDR. One explanation of this difference could lie in the fact that yeasts are unicellular organisms when humans are complex multicellular ones. Most of yeast genes are supposed to be used almost constantly, unlike human genes that are mostly tissue-specific. Following this hypothesis, it could be advantageous for yeast to have a default organization of "open and ready to transcribe" chromatin at their promoter, and to actively close the promoters of the genes that need to be expressed in specific conditions only. In contrast, it could be advantageous in human to adopt the opposite default organization of "closed and repressed" chromatin at promoters and to open specifically the few genes needed in each cell. The comparison of sequence-predicted chromatin conformation at promoters of several species confirmed this hypothesis [29,90]. The nucleosomal organization at promoters follows a gradient, from "mostly NDR" to "mostly NAR", that corresponds to the complexity of the organisms (identified as the number of different tissues composing the organism) [29]. In other words, yeast, a simple unicellular organisms, exhibited the most sequence-encoded open chromatin at their promoters. Interestingly, the same rule applies in archaea possessing nucleosome-like structures, where the histone core is tetrameric instead of octameric in eukaryotes, leading to the wrapping of only about 80 bp of DNA in archeal nucleosomes instead of 147 bp in eukaryotic nucleosomes [94]. Inversely, vertebrates like zebrafish and mammals, which are

multicellular complex organisms, exhibited the most sequence-encoded closed chromatin at their promoters. Between them a range of intermediate signals was found, but with a clear progression from full NDR model for unicellular, to hybrid NDR-NAR and full NAR model in multicellular organisms, according to the increase in organism complexity. This result seems to confirm the hypothesis mentioned earlier about the two models of chromatin at promoter. However, following this hypothesis, genes that are expressed in all cell types of complex multicellular organisms should exhibit a NDR at their promoter, because the "open and ready to transcribe" model would then be advantageous for these genes. Interestingly, this is not the case, and the promoters of these gene are even stronger NAR than cell-type specific genes. To explain this result, it has been proposed that the presence of NAR at promoters could also be linked to a retention of nucleosomes at promoters in cells generally depleted in nucleosomes such as sperm cells, to ensure transmission of epigenetic informations [29]. Regardless of the real biological meaning of these different sequence-encoded nucleosome organizations at promoters, this example shows that it has been modified during the evolution, and that these changes are mainly the result of sequence modifications, with NDR in yeast and NAR in mammals.

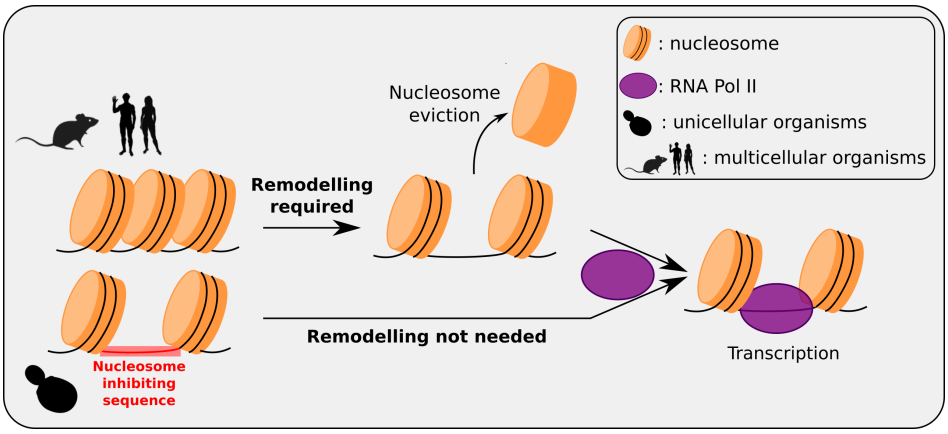


Figure 2. Multicellular/unicellular strategy for intrinsic nucleosomal organization at promoters. In human and most multicellular organisms, typical promoters are intrinsically occupied by nucleosomes, with no inhibiting sequences at these loci; genes are "repressed-by-default" and activated only when needed by the active removal of a nucleosome at the promoter making it accessible to the transcription machinery. In unicellular Eukaryotes such as yeast, inhibiting sequences have been selected at typical gene promoters, to avoid nucleosome formation at these loci; genes are "activated-by-default" since promoters are directly accessible to the transcription machinery, avoiding a remodeling step. These different strategies can be understood as in multicellular organisms, most genes (except housekeeping genes) present tissue-specificity, whereas in a unicellular Eukaryotes most genes are susceptible to be used in every cell. (RNA Pol II is not drawn to scale).

3.3. Is Chromatin Organization Selected Genome-Wide?

Examples of selection on specific nucleosomal organization at genes through selection of DNA sequences were described in Section 3.2. As nucleosome organization has a direct impact on the expression of genes, either driving expression level or transcriptional plasticity, it has a direct consequence on the fitness of individuals. Hence, selection of the corresponding sequence motifs in the course of evolution makes sense. However, genes represent only a very small fraction of the genome of most multicellular organisms. At numerous loci, nucleosomes are positioned by the intrinsic properties of the DNA sequence on which they are formed (Section 2). For example, nucleosomes are encoded in the DNA sequence over about 37% of the human genome through the statistical positioning at the border of NIEBs [28]. This genome-wide encoding of nucleosomes through nucleosomal barriers seems universal among vertebrates, as predicted in human but also in mouse, cow, pig, chicken and zebrafish [95]. This raises the question of the selection of this nucleosome positioning mechanism. In other words, are nucleosome positions also selected

at the genome-wide level? One NIEB feature that is common across vertebrates is the oscillating GC-content profile at NIEB borders, with very low GC at the internal border of NIEBs, then high GC on the ∼140 bp adjacent to the barrier (corresponding to the first stacked nucleosome position), then again low GC over ∼10 bp (first linker), then high GC over the second nucleosome location, low GC on the second linker, and so on. The oscillating pattern becomes less and less pronounced as we move away from the NIEBs, with barely no oscillation detectable after the third nucleosome. However, in the vicinity of NIEBs (∼500 bp of each border), the oscillations are very clear and observed across vertebrates species. As low GC is associated with inhibition of nucleosome formation, and higher GC content in general is associated with nucleosome positioning, the nucleosome organization at the border of NIEBs should also conserved be across these species, through the conservation of GC content. It was indeed observed that there is a link between a higher GC content at the location of nucleosome dyads compared to linker regions and sequence evolution [27,28]. By comparing the interspecies mutations between human and chimpanzee to intraspecies mutations obtained from the 1000 Genomes project [96] in human, several types of selection reinforcing the oscillation of GC content at the border of NIEBs have been observed [28]. First, signature of positive selection for mutations towards A and T nucleotides were described at the internal border of NIEBs and at the linker loci. Inversely, signatures of purifying selection (counterselection) were observed against these mutations at the positions corresponding to nucleosomal DNA. This confirmed an earlier observation of C-to-T mutations favored in linkers and disfavored in nucleosomes [27]. Second, mutations towards G and C nucleotides followed the exact opposite pattern, with purifying selection in NIEBs and in linkers, and positive selection in nucleosomal DNA. Finally, mutations disrupting TTT or AAA sequences (tTt-to-tAt or aAa-to-aTa mutations) were highly counter-selected in NIEBs and linkers, and favored in nucleosomal DNA. As these sequences strongly impair nucleosome formation, this suggests that natural selection is acting on NIEBs to maintain the nucleosomal organization at their borders. In a nutshell, evolution at human NIEBs loci favored mutations towards A and T in non-nucleosomal DNA, and mutations toward C and G in nucleosomal DNA, leading to the oscillating GC content also observed in each vertebrate analyzed, and reinforcing the positioning of two to three nucleosomes at these loci.

3.4. Are Transposable Elements Involved in Chromatin Organization?

For now, most studies about the interplay between sequence evolution and nucleosome positioning focused on single nucleotide variations (SNVs), analyzing their position relative to the nucleosomes. However, little is known about other types of mutations such as insertions or deletions in this context. The insertions of transposable elements (TEs) could in fact be important to fully capture the coupling between sequence-mediated nucleosome organization and genome evolution. Indeed, TEs are able to integrate and spread within genomes through a mechanism called transposition [97,98]. They are major components of Eukaryotic genomes, representing for example at least 45% of the human genome [99], although there is a high diversity in terms of TE composition in vertebrate genomes [100]. There are many families of TEs, according to their transposition mechanism, size, DNA base composition, etc. [101]. Some of these elements have been associated to a biological function. For example, a TE insertion can be at the origin of the formation of a new gene, an event called TE domestication (reviewed in [102]). Some TE copies were found to be implicated in various biological processes, for example in the sexual development and function in various animal species [103]. In contrast, some TE insertions have been found to have deleterious effects, with TEs being associated with various diseases [104]. Thus, TEs are major components of the evolution of genomic sequences, their transposition bringing DNA fragments to new locations, inserting from a few tens to several thousands of base pairs of DNA at the insertion site. If the sequence effects of these insertions have been largely investigated such as TFBS transport or coding sequence disruption, the effect of the insertion on the nucleosomal organization remains largely unknown. The insertion

of TEs, by disrupting the sequence at the insertion site, could either disrupt or reinforce the sequence-encoded nucleosomal organization, according to the nucleosome-associated properties of the inserted sequence. Thus, apart from being drivers of sequence evolution, TEs could also be drivers of the evolution of nucleosomal organization. Some results already point into this direction such as the presence in human of Alu transposable elements at the border of about half of the NIEBs mentioned in Section 2.3 [95]. The family of Alu TEs is specific to primate genomes [105]. They are short retrotransposons of about 300 bp, with a DNA sequence compatible with the positioning of two nucleosomes [106]. One hypothesis to explain the distribution of Alu TE at the border of human NIEBs is that NIEBs being NDRs and thus accessible to external factors, they could represent preferential target sites for the insertion of Alu TEs. Another hypothesis is that Alu TEs could be at the origin of new NIEBs formation, i.e., nucleosome organization would be a consequence of Alu insertion. Note that these hypothesis are not mutually exclusive, and the link that was observed between NIEBs and Alu TEs in human could result from the interplay between several mechanisms and selection. Moreover, strongly positioned nucleosomes were observed on newly inserted TEs, possibly participating to their regulation [107]. The presence of these nucleosome could both decrease the accessibility to these TEs for transposition machinery, making new transpositions more difficult, and increase the mutation rates on them, because DNA repair is less efficient in nucleosomes than in naked DNA [107]. In a general fashion, there seems to be an interconnection between TEs and the evolution of nucleosomal positioning that still needs to be investigated to fully understand the coupling between sequence evolution and chromatin evolution.

4. Feedback of Nucleosomal Positioning on Mutational Patterns

As we saw in Section 3.3, signatures of selection have been clearly identified at the borders of NIEBs. However, another phenomenon could participate to the observed reinforcement of the local GC content. In fact, profiles of mutational rates at NIEB borders were calculated, for both inter- and intra-specific human mutations [28]. This showed for example that interspecies mutation rates towards A and T were higher in non-nucleosomal DNA than in nucleosomal DNA. As discussed above, positive selection would favor these mutations in non-nucleosomal DNA while counterselection would act in nucleosomal DNA. In addition, some oscillations of mutation rates were also observed for intraspecies mutations, for which selection had way less time to influence the mutational pattern. Thus, it seems that even in the presence of weak to no selection, the mutations are not randomly distributed at the borders of the NIEBs. This suggests that nucleosome occupancy has a direct influence on the mutational patterns. The presence of a well-positioned nucleosome, meaning that it almost always covers the same DNA fragment, could then create a mutational bias on this DNA fragment, favoring some mutations type in the nucleosomal DNA with respect to the linker DNA. Nucleosome could bias mutations towards some specific nucleotides on the nucleosomal DNA, by its interaction with DNA damage mechanisms, or the DNA repair machinery. Next generation sequencing progress now permits to establish cartographies of specific DNA damage mechanisms on the genome, and to quantify the efficiency of DNA repair machinery. This made it possible to explore the direct influence of nucleosomes on mutational processes.

Early in the 2000s, it was shown that the excision repair mechanisms of DNA such as base excision repair (BER) or nucleotide excision repair (NER) are hampered by the presence of nucleosomes [108]. It was confirmed a decade later that DNA damages are more persistent in nucleosomal DNA [109]. As DNA damages can lead to mutations, notably during replication, the inhibition of BER and NER has a direct influence on mutational patterns. Nucleosomes also directly modulate the formation rate of certain type of DNA lesions [110]. These properties can be related to the stability of the DNA double helix in the nucleosomal context, as illustrated by the lower degradation rate after cell death of nucleosomal DNA compared to linker DNA in ancient DNA samples [111,112]. Nonetheless, it is crucial to decipher the interplay between nucleosomes and DNA lesion formation

and repair mechanisms to understand the influence of nucleosomes on the mutational pattern, and take it into account in evolutionary approaches. Two types of mutational biases have been described in relation to nucleosome positioning [31], associated to nucleosomal occupancy and the rotational positioning of nucleosomal DNA in regards to the histone core, respectively.

Concerning nucleosome occupancy, it has been shown that C-to-T mutations were depleted in nucleosomal DNA relative to linker DNA [113]. As discussed in Section 3.3, natural selection is implicated in the mutational biases [27,28], but a mutational mechanism itself could also be implicated. Indeed, C-to-T mutations usually results from spontaneous deamination of cytosines and 5mCs [31]. This mechanism is more efficient when there is a local opening, called "breathing", of the DNA double helix. Such breathing of DNA is inhibited in nucleosomal DNA, due to strong structural constraints imposed in the wrapping of DNA around histones, but remains possible in linker DNA, which is free from these constraints [114]. This wrapping is a hindrance to mutations leading to a depletion of the main C-to-T mutations in nucleosomal DNA as compared to linker DNA [113]. Similarly, experiments to map oxidatively induced DNA damages such as 8-oxoguanine (8-oxoG) in *S. cerevisiae* showed that they are modulated by nucleosome occupancy [115]. However, as 8-oxoG persistence depends on the equilibrium between DNA susceptibility to oxidation damage and efficiency of BER, it is still unclear whether the cause of the modulation by nucleosome occupancy is the influence on damage formation or on the efficiency of the repair mechanism [115]. Both hypotheses are not mutually exclusives. Further studies in yeast BER-deficient strains should provide insights about this question.

The effect of nucleosome occupancy on the mutational patterns has also been investigated in cancers where whole genome sequencing of tumors allows to examine the interplay between nucleosomes and mutational signatures [116–118]. These signatures correspond to unique combinations of mutation types, generated by specific mutational processes, in one or several types of cancers [119]. For example, mutational signature 1 found in all cancer types results from spontaneous deamination of 5-methylcytosine, and the type of mutation is mainly C-to-T mutation, with preferences for ACG, CCG, GCG and TCG contexts [119]. Mutations from signatures 17 and 18 are mainly T-to-G and C-to-A mutations, respectively, for which the mutational processes involved are unknown. In breast tumors, these two mutational signatures have been found to be more frequent in nucleosomes than it was expected from the sequence composition of the associated DNA fragments [116]. At transcription factor binding sites (TFBSs) flanked by regularly ordered nucleosomes following the model of statistical positioning by anchors (Section 2.3), melanoma mutations (principally induced by UV light) exhibit a periodic distribution associated to nucleosome positioning with a maximal density at nucleosome dyads, which differs from the expected pattern based on sequence composition [120]. More generally, a pan-cancer analysis revealed that for many cancer mutational processes, there are differences in mutation rates between nucleosomal DNA and linker DNA [121]. It also brought new observations, like tobacco-linked mutations occurring more frequently in linker than in nucleosomal DNA. The inhibition of both BER and NER repair systems is hypothesized to be a major player of UV-induced mutational biases. For tobacco-induced mutational bias, the mutational process (bulky DNA adducts at guanines (BPDE-dG)) is known to be inhibited in nucleosomes, leading to the "linker preference" for this type of mutations. The different examples mentioned here show that nucleosome dyad position (the so-called translational positioning of nucleosomes) has an influence on mutational patterns, through the modulation of the efficiency of either the DNA damage processes, or the repair mechanisms, or both, altogether leading to differences in mutation rates and biases between nucleosomal DNA and linker DNA.

Mutations are also modulated at a higher resolution than the nucleosome-linker dichotomy. Indeed, depending on which of the minor or the major groove of a DNA base pair faces the histones (the so-called rotational positioning of DNA within the nucleosome), mutation rates can be variable and, because DNA histone contact points are specifically with

the minor groove, each nucleosome translational positioning imposes a specific rotational positioning. A comparison between different *D. melanogaster* populations and between this species and a closely related species showed that C-to-T substitutions were more frequent where the minor groove of DNA faces the nucleosome (minor-in), than where the minor groove of DNA faces away from histones (minor-out) [122]. As at minor-in loci, the DNA is structurally constrained by chemical groups of histones H3 and H4, A/T (or WW) di-nucleotides could be favored for their higher flexibility and low steric hindrance [4]. The periodic occurrence of C-to-T mutations in nucleosomal DNA has been interpreted as a sign of selection on more favorable DNA fragments for nucleosome. However, an alternative hypothesis is that the interaction ability between DNA and mutagenic agents or repair machinery are different at minor-in and minor-out stretches of DNA, resulting in different mutation rates between these loci. This hypothesis is supported by the demonstrated decreased activity of BER at minor-in loci, resulting in lower repair efficiency of methylated guanines, the corollary of this being a higher mutation rate [123]. Experiments with DNase I showed that the accessibility to DNA could be a reason for the decreased activity of BER [124].

Another example of modulation of mutational processes along nucleosomes is for the UV-induced formation of cyclobutane pyrimidine dimers (CPDs) and (6-4) photoproducts (6,4-Pps) in DNA. Both DNA lesions are formed on TT, TC, CT and CC di-nucleotides. In nucleosomal DNA, a ∼10 bp periodicity has been observed in CPDs formation [125]. In fact, this periodic pattern correlates with the rotational positioning of nucleosomes, with preferential CPD formation at minor-out loci [125,126]. The 10 bp periodic pattern and the correlation have been observed genome-wide in yeast and human thanks to a NGS-based damage mapping method named CPD-seq [123,127,128]. The UV-irradiation of the same naked DNA fragment (without nucleosomes) resulted in an opposite CPD formation pattern, with CPDs occurring at positions corresponding to minor-in loci, probably because of the increase of TT dinucleotides at these regions (Section 2) [127]. This means that the underlying sequence is not the cause of the periodic formation pattern of CPDs in nucleosomal DNA, in fact the sequence would even favor the opposite pattern. The presence of a nucleosome, and the structural constraints associated with its formation, override the sequence preferences of CPDs to promote UV-damage at minor-out regions, where the DNA is more accessible. So, nucleosomes have a strong influence on this DNA damage process.

Distribution patterns favoring the minor-out stretches of DNA such as the CPD distribution described above are also found in some types of cancer. In melanoma, the vast majority of somatic mutations are C-to-T transitions at dimers of pyrimidine, characteristic of the UV mutational signature [129]. Analysis of melanoma mutations showed the same ∼10 bp periodicity in well-positioned nucleosomes as the one described for CPD mutations above [121,128]. The same pattern has been retrieved in lung cancer mutations, with high density at minor-out and low density at minor-in [121]. A high resolution genome-wide mapping of DNA damage process implicated in lung cancer mutations would help to understand if it is inhibited at minor-in stretches like CPD formation, or if the DNA damage distribution is constant, which would suggest an implication of the DNA repair mechanisms. On the other hand, other cancers exhibit an opposite mutational pattern, with high mutation densities at minor-in stretches and low mutation densities at minor-out stretches [121]. It has been proposed that a lower efficiency of BER mechanism at minor-in stretches could explain this periodicity. DNA damage at these loci would then be more persistent than at minor-out loci, leading to an increase in mutation rate [121]. Moreover, the presence of nucleosomes impair the recognition of single-strand breaks localized on the non-template DNA strand (NT-SSBs) by poly(adenosine diphosphate-ribose) polymerase 1 (PARP1), and in turn their reparation through BER [130]. Undetected nucleosomal NT-SSBs can be repaired during transcription through transcription-coupled nucleotide excision repair (TC-NER). The efficiency of this reparation mechanism is higher when the NT-SSBs is localized on the side of nucleosomal DNA facing the histone octamer than when it

is localized at more accessible loci in nucleosomal DNA. This could be because these accessible loci are more prone to be repaired by BER [130]. Although in the case of NT-SSBs, the interactions between nucleosome and DNA damage and repair mechanisms do not seem to lead to modulated mutation rates, it is another good example of the importance of nucleosome in DNA damage and repair processes.

Hence, the nucleosome has a strong influence on mutational patterns (Figure 3). It affects the effectiveness of excision repair mechanisms like BER and NER [131–134]. A lot of DNA damages are repaired through BER or NER, like UV-induced, tobacco-induced and oxydative DNA damage. The resulting mutational densities are generally increased in nucleosomal DNA compared to naked linker DNA. However, the activity of BER and NER are also modulated within the nucleosome, with a higher efficiency at minor-out loci, where DNA is more accessible than at minor-in loci, where structural constraints are stronger. The accessibility to DNA inside the nucleosome has also an influence on DNA damage mechanisms, some of them having a preference for minor-out stretches of DNA, and other for minor-in stretches [121]. All these possible influences of nucleosome on mutational patterns need to be taken into account in further attempts to decipher the evolution of DNA sequence regarding nucleosome positioning, to avoid considering as the sign of selection pressure some mutation biases induced by the presence of a nucleosome and its interplay with mutational processes.

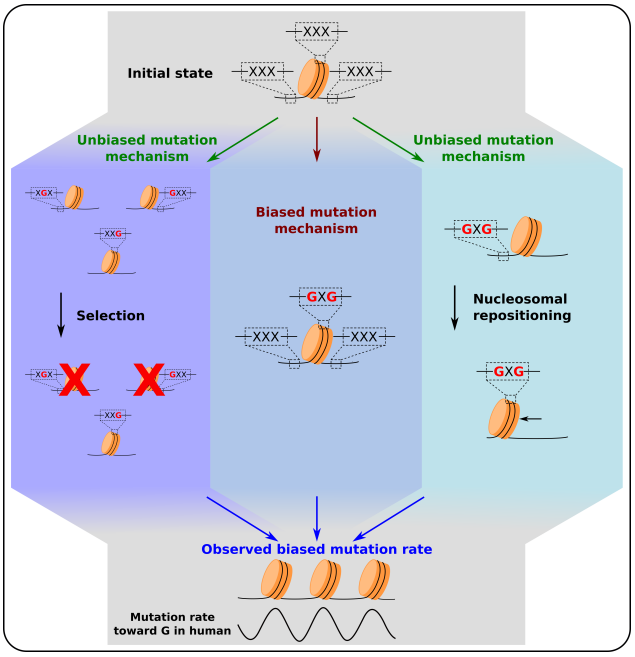


Figure 3. Source of biased mutation rates relative to nucleosomal positioning. Three mechanisms that can lead to biased mutation rates relative to nucleosome positioning as described in Sections 3 and 4. (**Left**) Mutations that facilitate the positioning of nucleosomes at specific loci are positively selected, those favoring alternative positions are purified. Such mechanism is for example observed at yeast promoters (Section 3.2). (**Center**) A biased mutation mechanism where the presence of a nucleosome drives mutations notably through interactions between nucleosomes and DNA damage and repair mechanisms (Section 4). (**Right**) A nucleosome repositioning model, in which mutations lead to the repositioning of nucleosomes that can also explain the observed biased mutation rates relative to nucleosome positioning when the latter is assumed to remain unchanged during evolution (Section 5). Since all three mechanisms have been observed at the genome scale, the global biased rate of mutations observed is likely to come from a combination of all three mechanisms. The cartoons illustrate possible evolutionary scenarios for 3 trinucleotides (XXX) located in the nucleosomal DNA, and the linker DNA upstream and downstream of a nucleosome. The figure only represents mutations toward G, but these three models are also valid for the other mutational biases (Sections 3–5).

5. Concluding Remarks

Nucleosome positions in genomes are at least partially encoded in the DNA sequence, through two main mechanisms (Section 2; Figure 1). The first one consists of an interplay between anti-positioning sequences (such as homopolymers like poly(dA:dT)) and high density of nucleosomes, leading to positioning by confinement between nucleosomal barriers [28,55,57,75]. The second mechanism consists in a fine-tuning of nucleosome positioning at the base-pair resolution, with preferences for A/T rich sequences where the DNA is making contact with histone proteins at minor-in positions, and G/C rich sequences where the DNA minor groove is facing away from histones [67]. In vivo and in vitro maps of nucleosomes present high similarities, indicating that the sequence properties are relevant even in the presence of other factors influencing the position of nucleosomes such as ATP-dependent remodelers. As nucleosomes have a functional importance, notably by modulating the accessibility to regulatory regions of genes, sequence evolution should be constrained by the effect of mutations on nucleosome positions (Figure 3). We reviewed several cases of sequence selection for their nucleosomal properties, in yeast, but also in multicellular organisms such as maize and human (Section 3). However, some caveats need to be taken into account when we try to decipher the evolution of sequence regarding its effect on nucleosomes. The first bias that must be considered is the feedback of nucleosome positioning on the mutational patterns. Nucleosomes influence both the mechanisms of DNA damage and DNA repair, leading to difference in the mutational patterns, either between nucleosomal DNA and linker DNA, and within the nucleosomes, between the minor-in and minor-out positions (Section 4; Figure 3). Now that these biases are described, one must be careful interpreting sequence changes in regards to nucleosomal positioning, and properly separate the contribution of selective pressure from the mutational biases induced by the presence of nucleosomes. A second caveat needs to be considered about the interpretation of some observations such as the signature of selective pressure on nucleosomal positioning, as initially raised in [30]. In many studies, mutational data obtained through the comparison of phylogenetically related species assume that the nucleosome organization was identical in the ancestor of the extant species. This "static" view of nucleosome positions in the course of evolution may not be a correct assumption [30], because nucleosomes are frequently repositioned following the evolution of sequences (Section 3). For example, in Eukaryotes, it was observed that A/T-to-G/C mutations are more frequent at the nucleosome dyad. It was interpreted as either a mutational bias caused by the nucleosome, or selection acting on these mutations to reinforce nucleosome positioning, assuming an evolutionary stable nucleosome organization. However, another scenario is compatible with the observations, where the A/T-to-G/C mutations would have repositioned the nucleosomes because of the preference of the dyad for GC-rich motifs [30], i.e., nucleosome positioning would follow the mutations towards G/C (Figure 3). To properly address this possibility, one needs to reconstruct the in vivo nucleosome organization at the time of the mutation. In regions where sequence-encoded nucleosome positioning is relevant in vivo, this can directly be done by applying the nucleosome position prediction tools available (Section 2.4) on the phylogenetically reconstructed ancestral sequences. Otherwise, one would need to compare experimental maps of nucleosome positioning in germline cells of all the species considered. However, for now, just about a handful of species have their nucleosomes mapped experimentally, mainly in somatic or cancer cell lines [54]. Thus, regions where the nucleosomal array appears not to be remodeled, such as NIEB loci in vertebrates, are the best candidates to distinguish between selection, repositioning, and the biased mutation events, and to estimate the relative importance of each mechanism to explain the mutational patterns at nucleosomes.

In this article, we reviewed here findings about the interplay between sequence-encoded nucleosome positioning and evolutionary constraints. Yet, the contribution of the collective properties and functions of the nucleosomal array depending on the position of nucleosomes relative to other nucleosomes have not been addressed. In genomes, the formation of nucleosomal arrays with regularly spaced nucleosomes is conserved across

Eukaryotic organisms [135]. These arrays are associated with various functions, such as chromatin condensation in higher order structure, but also with long-range contacts between enhancers and promoters, or inhibition of cryptic transcripts or protection of DNA from double-strand breaks [135]. The formation of nucleosomal arrays depends on various external factors, including remodeling, but also on DNA-binding factors creating nucleosomal barriers against which nucleosomes are stacked, following the model described in Section 2.3. Some sequence motifs such as NIEBs can also act as barriers. If one NIEB does not seem to be sufficient to position more than two to three nucleosomes at each of its borders, two close NIEBs can lead to a regularly spaced array of up to six nucleosomes between them [28]. In vertebrates, the relative position of NIEBs is constrained, with consecutive NIEBs being spaced by distances that are multiple of ∼153 bp [28,95], which was interpreted as a constraint that an integer numbers of compact nucleosomes fits between two close NIEBs. In this way, consecutive NIEBs could form regularly spaced nucleosomal array of controlled NRL following the statistical positioning model between two barriers [55,73,75]. The constraint of NIEB positioning regarding other NIEBs could arise to favor the apparition of such arrays, to use their properties on chromatin condensation and long-range contacts as described above. For example, short NRLs would assure that the intrinsic nucleosome arrays are in an open state, permissive to epigenetic regulation, allowing cell type specific regulation [28]. The influence of DNA sequence on nucleosome positioning and its interplay with the evolution of both sequences and nucleosome positions must then be considered not only at the nucleosome scale, but also at the scale of the nucleosomal array, thus taking into account higher order chromatin structures.

Author Contributions: Conceptualization, J.B., C.V., J.-N.V., F.G.B. and B.A.; writing—original draft preparation, J.B. and B.A.; writing—review and editing, J.B., C.V., J.-N.V., F.G.B. and B.A.; supervision, J.-N.V., F.G.B. and B.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Agence Nationale de la Recherche (ANR-20-CE12-013), and the ENS de Lyon "Projets Emergents" program. J.B. acknowledges support from the PhD funding program of ENS de Lyon.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank the French CNRS network GDR "Architecture et Dynamique Nucléaire" (ADN) for stimulating workshops.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

BER	Base excision repair
bp	Base pair
CPD	Cyclobutane pyrimidine dimer
MNase	Micrococcal nuclease
NAR	Nucleosome attracting region
NER	Nucleotide excision repair
NDR	Nucleosome depleted region
NIEB	Nucleosome-inhibiting energy barrier
NRL	Nucleosome repeat length
PTM	Post-translational modification

SNV	Single nucleotide variation
TE	Transposable element
TFBS	Transcription factor binding site
TSS	Transcription start site

References

1. Segal, E.; Widom, J. What controls nucleosome positions? *Trends Genet.* **2009**, *25*, 335–343. [CrossRef]
2. Kornberg, R.D.; Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **1999**, *98*, 285–294. [CrossRef]
3. Finch, J.T.; Lutter, L.C.; Rhodes, D.; Brown, R.S.; Rushton, B.; Levitt, M.; Klug, A. Structure of nucleosome core particles of chromatin. *Nature* **1977**, *269*, 29–36. [CrossRef] [PubMed]
4. McGinty, R.K.; Tan, S. Nucleosome Structure and Function. *Chem. Rev.* **2015**, *115*, 2255–2273. [CrossRef] [PubMed]
5. Lantermann, A.B.; Straub, T.; Strålfors, A.; Yuan, G.C.; Ekwall, K.; Korber, P. Schizosaccharomyces pombe genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of Saccharomyces cerevisiae. *Nat. Struct. Mol. Biol.* **2010**, *17*, 251–257. [CrossRef] [PubMed]
6. Athey, B.D.; Smith, M.F.; Rankert, D.A.; Williams, S.P.; Langmore, J.P. The diameters of frozen-hydrated chromatin fibers increase with DNA linker length: Evidence in support of variable diameter models for chromatin. *J. Cell Biol.* **1990**, *111*, 795–806. [CrossRef]
7. Kurumizaka, H.; Kujirai, T.; Takizawa, Y. Contributions of Histone Variants in Nucleosome Structure and Function. *J. Mol. Biol.* **2021**, *433*, 166678. [CrossRef]
8. Szenker, E.; Ray-Gallet, D.; Almouzni, G. The double face of the histone variant H3.3. *Cell Res.* **2011**, *21*, 421–434. [CrossRef]
9. Correll, S.J.; Schubert, M.H.; Grigoryev, S.A. Short nucleosome repeats impose rotational modulations on chromatin fibre folding. *EMBO J.* **2012**, *31*, 2416–2426. [CrossRef]
10. Valouev, A.; Johnson, S.M.; Boyd, S.D.; Smith, C.L.; Fire, A.Z.; Sidow, A. Determinants of nucleosome organization in primary human cells. *Nature* **2011**, *474*, 516–520. [CrossRef]
11. Pisano, S.; Galati, A.; Cacchione, S. Telomeric nucleosomes: Forgotten players at chromosome ends. *Cell. Mol. Life Sci.* **2008**, *65*, 3553–3563. [CrossRef] [PubMed]
12. Tommerup, H.; Dousmanis, A.; de Lange, T. Unusual chromatin in human telomeres. *Mol. Cell. Biol.* **1994**, *14*, 5777–5785. [CrossRef] [PubMed]
13. Makarov, V.L.; Lejnine, S.; Bedoyan, J.; Langmore, J.P. Nucleosomal organization of telomere-specific chromatin in rat. *Cell* **1993**, *73*, 775–787. [CrossRef]
14. Muyldermans, S.; De Jonge, J.; Wyns, L.; Travers, A.A. Differential association of linker histones H1 and H5 with telomeric nucleosomes in chicken erythrocytes. *Nucleic Acids Res.* **1994**, *22*, 5635–5639. [CrossRef]
15. Smilenov, L.B.; Dhar, S.; Pandita, T.K. Altered telomere nuclear matrix interactions and nucleosomal periodicity in ataxia telangiectasia cells before and after ionizing radiation treatment. *Mol. Cell. Biol.* **1999**, *19*, 6963–6971. [CrossRef]
16. Lejnine, S.; Makarov, V.L.; Langmore, J.P. Conserved nucleoprotein structure at the ends of vertebrate and invertebrate chromosomes. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 2393–2397. [CrossRef]
17. Fajkus, J.; Kovarík, A.; Královics, R.; Bezděk, M. Organization of telomeric and subtelomeric chromatin in the higher plant Nicotiana tabacum. *Mol. Gen. Genet.* **1995**, *247*, 633–638. [CrossRef]
18. Sýkorová, E.; Fajkus, J.; Ito, M.; Fukui, K. Transition between two forms of heterochromatin at plant subtelomeres. *Chromosome Res.* **2001**, *9*, 309–323. [CrossRef] [PubMed]
19. Vershinin, A.V.; Heslop-Harrison, J.S. Comparative analysis of the nucleosomal structure of rye, wheat and their relatives. *Plant Mol. Biol.* **1998**, *36*, 149–161. [CrossRef] [PubMed]
20. Rando, O.J.; Ahmad, K. Rules and regulation in the primary structure of chromatin. *Curr. Opin. Cell Biol.* **2007**, *19*, 250–256. [CrossRef]
21. Struhl, K.; Segal, E. Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.* **2013**, *20*, 267–273. [CrossRef] [PubMed]
22. Tillo, D.; Hughes, T.R. G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinform.* **2009**, *10*, 442. [CrossRef] [PubMed]
23. Field, Y.; Fondufe-Mittendorf, Y.; Moore, I.K.; Mieczkowski, P.; Kaplan, N.; Lubling, Y.; Lieb, J.D.; Widom, J.; Segal, E. Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. *Nat. Genet.* **2009**, *41*, 438–445. [CrossRef] [PubMed]
24. Tsankov, A.M.; Thompson, D.A.; Socha, A.; Regev, A.; Rando, O.J. The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol.* **2010**, *8*, e1000414. [CrossRef]
25. Chen, J.; Li, E.; Lai, J. The coupled effect of nucleosome organization on gene transcription level and transcriptional plasticity. *Nucleus* **2017**, *8*, 605–612. [CrossRef] [PubMed]
26. Chen, J.; Li, E.; Zhang, X.; Dong, X.; Lei, L.; Song, W.; Zhao, H.; Lai, J. Genome-wide Nucleosome Occupancy and Organization Modulates the Plasticity of Gene Transcriptional Status in Maize. *Mol. Plant* **2017**, *10*, 962–974. [CrossRef]
27. Prendergast, J.G.D.; Semple, C.A.M. Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. *Genome Res.* **2011**, *21*, 1777–1787. [CrossRef]

28. Drillon, G.; Audit, B.; Argoul, F.; Arneodo, A. Evidence of selection for an accessible nucleosomal array in human. *BMC Genom.* **2016**, *17*, 526. [CrossRef]
29. Tompitak, M.; Vaillant, C.; Schiessel, H. Genomes of Multicellular Organisms Have Evolved to Attract Nucleosomes to Promoter Regions. *Biophys. J.* **2017**, *112*, 505–511. [CrossRef]
30. Warnecke, T.; Becker, E.A.; Facciotti, M.T.; Nislow, C.; Lehner, B. Conserved substitution patterns around nucleosome footprints in eukaryotes and Archaea derive from frequent nucleosome repositioning through evolution. *PLoS Comput. Biol.* **2013**, *9*, e1003373. [CrossRef]
31. Gonzalez-Perez, A.; Sabarinathan, R.; Lopez-Bigas, N. Local Determinants of the Mutational Landscape of the Human Genome. *Cell* **2019**, *177*, 101–114. [CrossRef] [PubMed]
32. Widlund, H.R.; Cao, H.; Simonsson, S.; Magnusson, E.; Simonsson, T.; Nielsen, P.E.; Kahn, J.D.; Crothers, D.M.; Kubista, M. Identification and characterization of genomic nucleosome-positioning sequences. *J. Mol. Biol.* **1997**, *267*, 807–817. [CrossRef] [PubMed]
33. Lowary, P.T.; Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J. Mol. Biol.* **1998**, *276*, 19–42. [CrossRef]
34. Thåström, A.; Bingham, L.M.; Widom, J. Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. *J. Mol. Biol.* **2004**, *338*, 695–709. [CrossRef] [PubMed]
35. Gencheva, M.; Boa, S.; Fraser, R.; Simmen, M.W.; A Whitelaw, C.B.; Allan, J. In Vitro and in Vivo nucleosome positioning on the ovine beta-lactoglobulin gene are related. *J. Mol. Biol.* **2006**, *361*, 216–230. [CrossRef]
36. Parmar, J.J.; Marko, J.F.; Padinhateeri, R. Nucleosome positioning and kinetics near transcription-start-site barriers are controlled by interplay between active remodeling and DNA sequence. *Nucleic Acids Res.* **2014**, *42*, 128–136. [CrossRef] [PubMed]
37. Sexton, B.S.; Avey, D.; Druliner, B.R.; Fincher, J.A.; Vera, D.L.; Grau, D.J.; Borowsky, M.L.; Gupta, S.; Girimurugan, S.B.; Chicken, E.; et al. The spring-loaded genome: Nucleosome redistributions are widespread, transient, and DNA-directed. *Genome Res.* **2014**, *24*, 251–259. [CrossRef]
38. Vavouri, T.; Lehner, B. Chromatin organization in sperm may be the major functional consequence of base composition variation in the human genome. *PLoS Genet* **2011**, *7*, e1002036. [CrossRef]
39. Zhang, Z.; Wippo, C.J.; Wal, M.; Ward, E.; Korber, P.; Pugh, B.F. A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science* **2011**, *332*, 977–980. [CrossRef]
40. Sulkowski, E.; Laskowski, M. Mechanism of action of micrococcal nuclease on deoxyribonucleic acid. *J. Biol. Chem.* **1962**, *237*, 2620–2625. [CrossRef]
41. Axel, R. Cleavage of DNA in nuclei and chromatin with staphylococcal nuclease. *Biochemistry* **1975**, *14*, 2921–2925. [CrossRef]
42. Pajoro, A.; Muiño, J.M.; Angenent, G.C.; Kaufmann, K. Profiling Nucleosome Occupancy by MNase-seq: Experimental Protocol and Computational Analysis. *Methods Mol. Biol.* **2018**, *1675*, 167–181. [CrossRef]
43. Yuan, G.C.; Liu, Y.J.; Dion, M.F.; Slack, M.D.; Wu, L.F.; Altschuler, S.J.; Rando, O.J. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* **2005**, *309*, 626–630. [CrossRef] [PubMed]
44. Schones, D.E.; Cui, K.; Cuddapah, S.; Roh, T.Y.; Barski, A.; Wang, Z.; Wei, G.; Zhao, K. Dynamic regulation of nucleosome positioning in the human genome. *Cell* **2008**, *132*, 887–898. [CrossRef]
45. Gaffney, D.J.; McVicker, G.; Pai, A.A.; Fondufe-Mittendorf, Y.N.; Lewellen, N.; Michelini, K.; Widom, J.; Gilad, Y.; Pritchard, J.K. Controls of nucleosome positioning in the human genome. *PLoS Genet.* **2012**, *8*, e1003036. [CrossRef] [PubMed]
46. Lai, B.; Gao, W.; Cui, K.; Xie, W.; Tang, Q.; Jin, W.; Hu, G.; Ni, B.; Zhao, K. Principles of nucleosome organization revealed by single-cell micrococcal nuclease sequencing. *Nature* **2018**, *562*, 281–285. [CrossRef]
47. Oberbeckmann, E.; Wolff, M.; Krietenstein, N.; Heron, M.; Ellins, J.L.; Schmid, A.; Krebs, S.; Blum, H.; Gerland, U.; Korber, P. Absolute nucleosome occupancy map for the Saccharomyces cerevisiae genome. *Genome Res.* **2019**, *29*, 1996–2009. [CrossRef] [PubMed]
48. Lee, W.; Tillo, D.; Bray, N.; Morse, R.H.; Davis, R.W.; Hughes, T.R.; Nislow, C. A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.* **2007**, *39*, 1235–1244. [CrossRef]
49. Mavrich, T.N.; Jiang, C.; Ioshikhes, I.P.; Li, X.; Venters, B.J.; Zanton, S.J.; Tomsho, L.P.; Qi, J.; Glaser, R.; Schuster, S.C.; et al. Nucleosome organization in the Drosophila genome. *Nature* **2008**, *453*, 358–362. [CrossRef] [PubMed]
50. Zhang, T.; Zhang, W.; Jiang, J. Genome-Wide Nucleosome Occupancy and Positioning and Their Impact on Gene Expression and Evolution in Plants. *Plant Physiol.* **2015**, *168*, 1406–1416. [CrossRef] [PubMed]
51. Teif, V.B.; Vainshtein, Y.; Caudron-Herger, M.; Mallm, J.P.; Marth, C.; Höfer, T.; Rippe, K. Genome-wide nucleosome positioning during embryonic stem cell development. *Nat. Struct. Mol. Biol.* **2012**, *19*, 1185–1192. [CrossRef]
52. Jeffers, T.E.; Lieb, J.D. Nucleosome fragility is associated with future transcriptional response to developmental cues and stress in *C. elegans*. *Genome Res.* **2017**, *27*, 75–86. [CrossRef]
53. Kaplan, N.; Moore, I.K.; Fondufe-Mittendorf, Y.; Gossett, A.J.; Tillo, D.; Field, Y.; LeProust, E.M.; Hughes, T.R.; Lieb, J.D.; Widom, J.; et al. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **2009**, *458*, 362–366. [CrossRef] [PubMed]
54. Teif, V.B. Nucleosome positioning: Resources and tools online. *Brief. Bioinform.* **2016**, *17*, 745–757. [CrossRef] [PubMed]
55. Chevereau, G.; Arnéodo, A.; Vaillant, C. Influence of the genomic sequence on the primary structure of chromatin. *Front. Life Sci.* **2011**, *5*, 29–68. [CrossRef]

56. Segal, E.; Fondufe-Mittendorf, Y.; Chen, L.; Thåström, A.; Field, Y.; Moore, I.K.; Wang, J.P.Z.; Widom, J. A genomic code for nucleosome positioning. *Nature* **2006**, *442*, 772–778. [CrossRef] [PubMed]
57. Vaillant, C.; Audit, B.; Arneodo, A. Experiments confirm the influence of genome long-range correlations on nucleosome positioning. *Phys. Rev. Lett.* **2007**, *99*, 218103. [CrossRef]
58. Field, Y.; Kaplan, N.; Fondufe-Mittendorf, Y.; Moore, I.K.; Sharon, E.; Lubling, Y.; Widom, J.; Segal, E. Distinct Modes of Regulation by Chromatin Encoded through Nucleosome Positioning Signals. *PLoS Comput. Biol.* **2008**, *4*. [CrossRef]
59. Miele, V.; Vaillant, C.; d'Aubenton Carafa, Y.; Thermes, C.; Grange, T. DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Res.* **2008**, *36*, 3746–3756. [CrossRef]
60. Scipioni, A.; De Santis, P. Predicting nucleosome positioning in genomes: Physical and bioinformatic approaches. *Biophys. Chem.* **2011**, *155*, 53–64. [CrossRef]
61. Kaplan, N.; Hughes, T.R.; Lieb, J.D.; Widom, J.; Segal, E. Contribution of histone sequence preferences to nucleosome organization: Proposed definitions and methodology. *Genome Biol.* **2010**, *11*, 140. [CrossRef]
62. Trifonov, E.N.; Sussman, J.L. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl. Acad. Sci. USA* **1980**, *77*, 3816–3820. [CrossRef] [PubMed]
63. Trifonov, E.N. Sequence-dependent deformational anisotropy of chromatin DNA. *Nucleic Acids Res.* **1980**, *8*, 4041–4053. [CrossRef]
64. Levene, S.D.; Crothers, D.M. A computer graphics study of sequence-directed bending in DNA. *J. Biomol. Struct. Dyn.* **1983**, *1*, 429–435. [CrossRef] [PubMed]
65. Rhodes, D. Nucleosome cores reconstituted from poly (dA-dT) and the octamer of histones. *Nucleic Acids Res.* **1979**, *6*, 1805–1816. [CrossRef]
66. Simpson, R.T.; Künzler, P. Cromatin and core particles formed from the inner histones and synthetic polydeoxyribonucleotides of defined sequence. *Nucleic Acids Res.* **1979**, *6*, 1387–1415. [CrossRef]
67. Drew, H.R.; Travers, A.A. DNA bending and its relation to nucleosome positioning. *J. Mol. Biol.* **1985**, *186*, 773–790. [CrossRef]
68. Satchwell, S.C.; Drew, H.R.; Travers, A.A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **1986**, *191*, 659–675. [CrossRef]
69. Johnson, S.M.; Tan, F.J.; McCullough, H.L.; Riordan, D.P.; Fire, A.Z. Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. *Genome Res.* **2006**, *16*, 1505–1516. [CrossRef]
70. Albert, I.; Mavrich, T.N.; Tomsho, L.P.; Qi, J.; Zanton, S.J.; Schuster, S.C.; Pugh, B.F. Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature* **2007**, *446*, 572–576. [CrossRef] [PubMed]
71. Raveh-Sadka, T.; Levo, M.; Shabi, U.; Shany, B.; Keren, L.; Lotan-Pompan, M.; Zeevi, D.; Sharon, E.; Weinberger, A.; Segal, E. Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat. Genet.* **2012**, *44*, 743–750. [CrossRef]
72. Fedor, M.J.; Lue, N.F.; Kornberg, R.D. Statistical positioning of nucleosomes by specific protein-binding to an upstream activating sequence in yeast. *J. Mol. Biol.* **1988**, *204*, 109–127. [CrossRef]
73. Kornberg, R.D.; Stryer, L. Statistical distributions of nucleosomes: Nonrandom locations by a stochastic mechanism. *Nucleic Acids Res.* **1988**, *16*, 6677–6690. [CrossRef]
74. Iyer, V.R. Nucleosome positioning: Bringing order to the eukaryotic genome. *Trends Cell Biol.* **2012**, *22*, 250–256. [CrossRef]
75. Milani, P.; Chevereau, G.; Vaillant, C.; Audit, B.; Haftek-Terreau, Z.; Marilley, M.; Bouvet, P.; Argoul, F.; Arneodo, A. Nucleosome positioning by genomic excluding-energy barriers. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 22257–22262. [CrossRef] [PubMed]
76. Drillon, G.; Audit, B.; Argoul, F.; Arneodo, A. Ubiquitous human 'master' origins of replication are encoded in the DNA sequence via a local enrichment in nucleosome excluding energy barriers. *J. Phys. Condens. Matter* **2015**, *27*, 064102. [CrossRef] [PubMed]
77. Tompitak, M.; Barkema, G.T.; Schiessel, H. Benchmarking and refining probability-based models for nucleosome-DNA interaction. *BMC Bioinform.* **2017**, *18*, 157. [CrossRef] [PubMed]
78. Nishida, H. Nucleosome Positioning. *ISRN Mol. Biol.* **2012**, *2012*, 245706. [CrossRef]
79. Chevereau, G.; Palmeira, L.; Thermes, C.; Arneodo, A.; Vaillant, C. Thermodynamics of intragenic nucleosome ordering. *Phys. Rev. Lett.* **2009**, *103*, 188103. [CrossRef]
80. Vaillant, C.; Palmeira, L.; Chevereau, G.; Audit, B.; d'Aubenton Carafa, Y.; Thermes, C.; Arneodo, A. A novel strategy of transcription regulation by intragenic nucleosome ordering. *Genome Res.* **2010**, *20*, 59–67. [CrossRef]
81. Brogaard, K.; Xi, L.; Wang, J.P.; Widom, J. A map of nucleosome positions in yeast at base-pair resolution. *Nature* **2012**, *486*, 496–501. [CrossRef]
82. Shivaswamy, S.; Bhinge, A.; Zhao, Y.; Jones, S.; Hirst, M.; Iyer, V.R. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol.* **2008**, *6*, e65. [CrossRef]
83. Bochkis, I.M.; Przybylski, D.; Chen, J.; Regev, A. Changes in nucleosome occupancy associated with metabolic alterations in aged mammalian liver. *Cell Rep.* **2014**, *9*, 996–1006. [CrossRef] [PubMed]
84. Hu, Z.; Chen, K.; Xia, Z.; Chavez, M.; Pal, S.; Seol, J.H.; Chen, C.C.; Li, W.; Tyler, J.K. Nucleosome loss leads to global transcriptional up-regulation and genomic instability during yeast aging. *Genes Dev.* **2014**, *28*, 396–408. [CrossRef]
85. Sexton, B.S.; Druliner, B.R.; Vera, D.L.; Avey, D.; Zhu, F.; Dennis, J.H. Hierarchical regulation of the genome: Global changes in nucleosome organization potentiate genome response. *Oncotarget* **2016**, *7*, 6460–6475. [CrossRef] [PubMed]
86. Weghorn, D.; Lässig, M. Fitness landscape for nucleosome positioning. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 10988–10993. [CrossRef] [PubMed]

87. Eslami-Mossallam, B.; Schram, R.D.; Tompitak, M.; van Noort, J.; Schiessel, H. Multiplexing Genetic and Nucleosome Positioning Codes: A Computational Approach. *PLoS ONE* **2016**, *11*, e0156905. [CrossRef] [PubMed]
88. Mieczkowski, J.; Cook, A.; Bowman, S.K.; Mueller, B.; Alver, B.H.; Kundu, S.; Deaton, A.M.; Urban, J.A.; Larschan, E.; Park, P.J.; et al. MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nat. Commun.* **2016**, *7*, 1–11. [CrossRef] [PubMed]
89. Chereji, R.V.; Bryson, T.D.; Henikoff, S. Quantitative MNase-seq accurately maps nucleosome occupancy levels. *Genome Biol.* **2019**, *20*, 198. [CrossRef]
90. Arneodo, A.; Drillon, G.; Argoul, F.; Audit, B. 2-The Role of Nucleosome Positioning in Genome Function and Evolution. In *Nuclear Architecture and Dynamics*; Translational Epigenetics; Lavelle, C., Victor, J.M., Eds.; Academic Press: Boston, MA, USA, 2018; Volume 2, pp. 41–79. [CrossRef]
91. Saxonov, S.; Berg, P.; Brutlag, D.L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 1412–1417. [CrossRef]
92. Ramirez-Carrozzi, V.R.; Braas, D.; Bhatt, D.M.; Cheng, C.S.; Hong, C.; Doty, K.R.; Black, J.C.; Hoffmann, A.; Carey, M.; Smale, S.T. A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. *Cell* **2009**, *138*, 114–128. [CrossRef]
93. Deaton, A.M.; Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **2011**, *25*, 1010–1022. [CrossRef]
94. Sandman, K.; Reeve, J.N. Structure and functional relationships of archaeal and eukaryal histones and nucleosomes. *Arch. Microbiol.* **2000**, *173*, 165–169. [CrossRef]
95. Brunet, F.G.; Audit, B.; Drillon, G.; Argoul, F.; Volff, J.N.; Arneodo, A. Evidence for DNA Sequence Encoding of an Accessible Nucleosomal Array across Vertebrates. *Biophys. J.* **2018**. [CrossRef]
96. Durbin, R.M.; Altshuler, D.; Durbin, R.M.; Abecasis, G.R.; Bentley, D.R.; Chakravarti, A.; Clark, A.G.; Collins, F.S.; De La Vega, F.M.; Donnelly, P.; et al. A map of human genome variation from population-scale sequencing. *Nature* **2010**, *467*, 1061–1073. [CrossRef]
97. Biémont, C.; Vieira, C. Junk DNA as an evolutionary force. *Nature* **2006**, *443*, 521–524. [CrossRef]
98. Bourque, G.; Burns, K.H.; Gehring, M.; Gorbunova, V.; Seluanov, A.; Hammell, M.; Imbeault, M.; Izsvák, Z.; Levin, H.L.; Macfarlan, T.S.; et al. Ten things you should know about transposable elements. *Genome Biol.* **2018**, *19*, 199. [CrossRef]
99. Pace, J.K.; Feschotte, C. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. *Genome Res.* **2007**, *17*, 422–432. [CrossRef] [PubMed]
100. Chalopin, D.; Naville, M.; Plard, F.; Galiana, D.; Volff, J.N. Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates. *Genome Biol. Evol.* **2015**, *7*, 567–580. [CrossRef] [PubMed]
101. Wicker, T.; Sabot, F.; Hua-Van, A.; Bennetzen, J.L.; Capy, P.; Chalhoub, B.; Flavell, A.; Leroy, P.; Morgante, M.; Panaud, O.; et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **2007**, *8*, 973–982. [CrossRef] [PubMed]
102. Etchegaray, E.; Naville, M.; Volff, J.N.; Haftek-Terreau, Z. Transposable element-derived sequences in vertebrate development. *Mob. DNA* **2021**, *12*, 1. [CrossRef]
103. Dechaud, C.; Volff, J.N.; Schartl, M.; Naville, M. Sex and the TEs: Transposable elements in sexual development and function in animals. *Mob. DNA* **2019**, *10*. [CrossRef] [PubMed]
104. Payer, L.M.; Burns, K.H. Transposable elements in human genetic disease. *Nat. Rev. Genet.* **2019**, *20*, 760–772. [CrossRef] [PubMed]
105. Deininger, P. Alu elements: Know the SINEs. *Genome Biol.* **2011**, *12*, 236. [CrossRef]
106. Tanaka, Y.; Yamashita, R.; Suzuki, Y.; Nakai, K. Effects of Alu elements on global nucleosome positioning in the human genome. *BMC Genom.* **2010**, *11*, 309. [CrossRef] [PubMed]
107. Li, C.; Luscombe, N.M. Nucleosome positioning stability is a modulator of germline mutation rate variation across the human genome. *Nat. Commun.* **2020**, *11*, 1363. [CrossRef] [PubMed]
108. Hara, R.; Mo, J.; Sancar, A. DNA damage in the nucleosome core is refractory to repair by human excision nuclease. *Mol. Cell. Biol.* **2000**, *20*, 9173–9181. [CrossRef] [PubMed]
109. Rodriguez, Y.; Smerdon, M.J. The structural location of DNA lesions in nucleosome core particles determines accessibility by base excision repair enzymes. *J. Biol. Chem.* **2013**, *288*, 13863–13875. [CrossRef]
110. Smerdon, M.J.; Conconi, A. Modulation of DNA damage and DNA repair in chromatin. *Prog. Nucleic Acid Res. Mol. Biol.* **1999**, *62*, 227–255. [CrossRef]
111. Pedersen, J.S.; Valen, E.; Velazquez, A.M.V.; Parker, B.J.; Rasmussen, M.; Lindgreen, S.; Lilje, B.; Tobin, D.J.; Kelly, T.K.; Vang, S.; et al. Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. *Genome Res.* **2014**, *24*, 454–466. [CrossRef]
112. Hanghøj, K.; Seguin-Orlando, A.; Schubert, M.; Madsen, T.; Pedersen, J.S.; Willerslev, E.; Orlando, L. Fast, Accurate and Automatic Ancient Nucleosome and Methylation Maps with epiPALEOMIX. *Mol. Biol. Evol.* **2016**, *33*, 3284–3298. [CrossRef]
113. Chen, X.; Chen, Z.; Chen, H.; Su, Z.; Yang, J.; Lin, F.; Shi, S.; He, X. Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. *Science* **2012**, *335*, 1235–1238. [CrossRef] [PubMed]
114. Makova, K.D.; Hardison, R.C. The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.* **2015**, *16*, 213–223. [CrossRef]

115. Wu, J.; McKeague, M.; Sturla, S.J. Nucleotide-Resolution Genome-Wide Mapping of Oxidative DNA Damage by Click-Code-Seq. *J. Am. Chem. Soc.* **2018**, *140*, 9783–9787. [CrossRef]
116. Morganella, S.; Alexandrov, L.B.; Glodzik, D.; Zou, X.; Davies, H.; Staaf, J.; Sieuwerts, A.M.; Brinkman, A.B.; Martin, S.; Ramakrishna, M.; et al. The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **2016**, *7*, 11383. [CrossRef] [PubMed]
117. Nik-Zainal, S.; Morganella, S. Mutational Signatures in Breast Cancer: The Problem at the DNA Level. *Clin. Cancer Res.* **2017**, *23*, 2617–2629. [CrossRef]
118. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **2020**, *578*, 82–93. [CrossRef] [PubMed]
119. Alexandrov, L.B.; Kim, J.; Haradhvala, N.J.; Huang, M.N.; Tian Ng, A.W.; Wu, Y.; Boot, A.; Covington, K.R.; Gordenin, D.A.; Bergstrom, E.N.; et al. The repertoire of mutational signatures in human cancer. *Nature* **2020**, *578*, 94–101. [CrossRef] [PubMed]
120. Sabarinathan, R.; Mularoni, L.; Deu-Pons, J.; Gonzalez-Perez, A.; López-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **2016**, *532*, 264–267. [CrossRef]
121. Pich, O.; Muiños, F.; Sabarinathan, R.; Reyes-Salazar, I.; Gonzalez-Perez, A.; Lopez-Bigas, N. Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. *Cell* **2018**, *175*, 1074–1087.e18. [CrossRef]
122. Langley, S.A.; Karpen, G.H.; Langley, C.H. Nucleosomes shape DNA polymorphism and divergence. *PLoS Genet.* **2014**, *10*, e1004457. [CrossRef]
123. Mao, P.; Brown, A.J.; Malc, E.P.; Mieczkowski, P.A.; Smerdon, M.J.; Roberts, S.A.; Wyrick, J.J. Genome-wide maps of alkylation damage, repair, and mutagenesis in yeast reveal mechanisms of mutational heterogeneity. *Genome Res.* **2017**, *27*, 1674–1684. [CrossRef] [PubMed]
124. Mao, P.; Wyrick, J.J. Organization of DNA damage, excision repair, and mutagenesis in chromatin: A genomic perspective. *DNA Repair* **2019**, *81*, 102645. [CrossRef] [PubMed]
125. Gale, J.M.; Nissen, K.A.; Smerdon, M.J. UV-induced formation of pyrimidine dimers in nucleosome core DNA is strongly modulated with a period of 10.3 bases. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 6644–6648. [CrossRef] [PubMed]
126. Mao, P.; Wyrick, J.J.; Roberts, S.A.; Smerdon, M.J. UV-Induced DNA Damage and Mutagenesis in Chromatin. *Photochem. Photobiol.* **2017**, *93*, 216–228. [CrossRef] [PubMed]
127. Mao, P.; Smerdon, M.J.; Roberts, S.A.; Wyrick, J.J. Chromosomal landscape of UV damage formation and repair at single-nucleotide resolution. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 9057–9062. [CrossRef]
128. Brown, A.J.; Mao, P.; Smerdon, M.J.; Wyrick, J.J.; Roberts, S.A. Nucleosome positions establish an extended mutation signature in melanoma. *PLoS Genet.* **2018**, *14*. [CrossRef]
129. Hayward, N.K.; Wilmott, J.S.; Waddell, N.; Johansson, P.A.; Field, M.A.; Nones, K.; Patch, A.M.; Kakavand, H.; Alexandrov, L.B.; Burke, H.; et al. Whole-genome landscapes of major melanoma subtypes. *Nature* **2017**, *545*, 175–180. [CrossRef]
130. Pestov, N.A.; Gerasimova, N.S.; Kulaeva, O.I.; Studitsky, V.M. Structure of transcribed chromatin is a sensor of DNA damage. *Sci. Adv.* **2015**, *1*, e1500021. [CrossRef]
131. Adar, S.; Hu, J.; Lieb, J.D.; Sancar, A. Genome-wide kinetics of DNA excision repair in relation to chromatin state and mutagenesis. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E2124–E2133. [CrossRef]
132. Li, W.; Hu, J.; Adebali, O.; Adar, S.; Yang, Y.; Chiou, Y.Y.; Sancar, A. Human genome-wide repair map of DNA damage caused by the cigarette smoke carcinogen benzo[a]pyrene. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 6752–6757. [CrossRef]
133. Polak, P.; Karlić, R.; Koren, A.; Thurman, R.; Sandstrom, R.; Lawrence, M.; Reynolds, A.; Rynes, E.; Vlahoviček, K.; Stamatoyannopoulos, J.A.; et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **2015**, *518*, 360–364. [CrossRef] [PubMed]
134. Hu, J.; Lieb, J.D.; Sancar, A.; Adar, S. Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 11507–11512. [CrossRef] [PubMed]
135. Singh, A.K.; Mueller-Planitz, F. Nucleosome Positioning and Spacing: From Mechanism to Function. *J. Mol. Biol.* **2021**, *433*, 166847. [CrossRef] [PubMed]

1.1.4 Les barrières nucléosomales, un moyen de positionnement des nucléosomes à l'échelle génomique

1.1.4.1 Un encodage des nucléosomes dans la séquence qui semble ubiquitaire

Dans cette revue bibliographique, nous avons vu que le principal moyen d'encodage du positionnement nucléosomal dans la séquence d'ADN est de générer des zones inhibant la formation des nucléosomes, sur les bords desquelles les nucléosomes sont calés en suivant les règles du positionnement statistique (Barbier et al., 2021; Drillon et al., 2016). Ces séquences inhibitrices, également appelées NIEBs (pour Nucleosome Inhibitory Energy Barriers) ou encore barrières nucléosomales, sont des séquences pour lesquelles l'énergie nécessaire à la courbure de l'ADN pour l'enrouler autour des protéines histones et former le nucléosome est bien plus élevée que celle des séquences adjacentes, ce qui rend très peu probable la formation d'un nucléosome à ce locus car trop coûteux en énergie (Chevereau et al., 2011; Miele et al., 2008). En revanche, les barrières nucléosomales peuvent contribuer au positionnement des nucléosomes. En effet, elles amènent des contraintes dans la formation des nucléosomes, sous la forme de zones ou cette formation est inhibée. Le nucléosome est donc préférentiellement formé hors de ces zones. Aux bords des barrières nucléosomales, le positionnement statistique amène les nucléosomes à se former plus souvent à certaines positions précise, particulièrement à celle positionnant le nucléosome contre la barrière nucléosomale (Kornberg & Stryer, 1988). Un second nucléosome se formera alors préférentiellement contre le premier, et ainsi de suite. C'est ce qu'on définit ici comme un effet de parcage des nucléosomes contre les barrières nucléosomales. Cet effet est responsable du positionnement préférentiel des nucléosomes aux bords des NIEBs. Il a été observé expérimentalement grâce à des images de microscopie à force atomique, confirmant expérimentalement le principe théorique (Milani et al., 2009).

À l'aide d'un modèle élastique calculant le coût énergétique pour courber un fragment d'ADN dans la configuration nucléosomale à partir de sa conformation libre dépendant de sa séquence, il est donc possible de prédire la localisation des barrières nucléosomales, et également d'en déduire la position des nucléosomes par positionnement statistique aux bords de ces barrières. Cela a été réalisé dans diverses espèces, mettant en évidence des NIEBs chez la levure, principalement au niveau des promoteurs de gènes (Chevereau et al., 2011; Vaillant et al., 2010), chez le nématode et la drosophile (Chevereau et al., 2011; Miele et al., 2008), et également chez l'humain, d'abord au niveau des origines de réplication conservées entre les types cellulaires (Drillon et al., 2015), puis de manière générale dans l'ensemble du génome humain (Drillon et al., 2016). Dans toutes ces espèces, la prédiction de l'occupation nucléosomale basée sur notre modèle a été validée par l'analyse de jeux de données expérimentales de positionnement nucléosomal. Des barrières nucléosomales ont également été détectées dans une variété d'autres espèces représentatives des vertébrés comme la souris, la vache, le porc, la poule ainsi que le poisson-zèbre, même si aucune confirmation expérimentale des prédictions du modèle n'a encore été rapportée. Le **Chapitre 3** de ce manuscrit sera en partie consacré à la validation expérimentale des prédictions du modèle dans les espèces pour lesquelles des données expérimentales de positionnement des nuclésomes sont disponibles.

1.1.4.2 Des caractéristiques partagées chez les vertébrés

Chez l'humain, une analyse détaillée des barrières nucléosomales a permis d'en définir les caractéristiques (Drillon et al., 2016). Tout d'abord, l'analyse des données expérimentales de positionnement de nucléosomes de type MNase-seq a mis en évidence que les NIEBs correspondent *in vivo* à des régions déplétées en nucléosomes (NDRs), en accord avec les prédictions du modèle physique de positionnement des nucléosomes. Toujours en accord avec notre modèle, le positionnement prédit aux bords des NIEBs reproduit bien le positionnement observé expérimentalement, avec deux à trois nucléosomes très positionnés aux bords des barrières. En d'autres termes, aux bords des barrières nucléosomales, la position des nucléosomes semble contrainte et suit un positionnement statistique, comme observé dans la Figure 1 de l'article de Drillon et al. (Drillon et al., 2016). Dans cet article (et cette Figure en particulier), l'analyse des séquences au niveau des NIEBs a mis en évidence un pourcentage GC oscillant aux bords des NIEBs entre des valeurs très basses au niveau des barrières et des linkers, et des valeurs bien plus élevées dans les séquences nucléosomales. Cette oscillation du contenu en GC a d'ailleurs également été observée chez les autres vertébrés où les NIEBs ont été prédits (Brunet et al., 2018, Figure 3).

Concernant la distribution des NIEBs dans les génomes, aucune régionalisation particulière n'a été observée chez l'humain, avec une densité en NIEBs plutôt homogène tout le long du génome. En effet, que ce soit vis à vis du pourcentage GC, du timing de réplication, du taux de recombinaison ou encore des régions géniques ou intergéniques, on n'observe que peu de différence dans la densité en NIEBs des différentes régions (Drillon et al., 2016, Tableau 1). La seule exception à cette règle se trouve aux promoteurs des gènes, où significativement moins de NIEBs sont retrouvés que dans le reste du génome (Drillon et al., 2016). En revanche, si la distribution des NIEBs dans le génome humain ne semble pas contrainte, le positionnement de ces barrières nucléosomales les unes par rapport aux autres semble l'être fortement. Particulièrement lorsque les barrières sont proches les unes des autres (à moins d'un millier de paires de bases), on observe une distribution quantifiée de la distance entre deux barrières consécutives, avec des pics séparés de 153 pb (Drillon et al., 2016). En fait, lorsque les barrières sont proches les unes des autres, leur positionnement est tel qu'un nombre entier de nucléosomes peut se former entre deux barrières. Ainsi, deux barrières proches l'une de l'autre sont séparées par une distance correspondant à exactement 1, 2, 3, 4 ou 5 nucléosomes, avec les distances intermédiaires largement sous-représentées (Drillon et al., 2016, Figure 2). Cette hypothèse de nombres entiers de nucléosomes entre deux barrières consécutives a été confirmée expérimentalement (Drillon et al., 2016, Figure 3). Enfin, la distribution des distances inter-NIEBs est, à l'image de celle du pourcentage GC aux bords des NIEBs, conservée dans les autres espèces de vertébrés étudiées (Brunet et al., 2018, Figure 2). On note cependant quelques spécificités propres à certaines espèces, particulièrement chez l'humain et chez le porc. Chez l'humain, cette spécificité a été associée à une famille d'éléments transposables, les éléments Alu. Les caractéristiques des éléments transposables ainsi que celles propres aux éléments Alu seront détaillées plus loin dans cette introduction (**Partie 1.2**). La relation entre éléments transposables Alu et barrières nucléosomales est au cœur de mon travail de thèse. Le **Chapitre 4** est entièrement dédié à mes travaux autour de cette question.

1.1.4.3 Un positionnement nucléosomal sélectionné au cours de l'évolution

Chez l'humain, l'analyse des profils de mutations aux bords des NIEBs a mis en évidence une hétérogénéité des différents taux de mutations. De manière générale, on observe que les mutations de G ou C vers A ou T ont un taux plus important au niveau des NIEBs et des séquences inter-nucléosomales que dans les séquences nucléosomales. A l'inverse, les mutations de A ou T vers G ou C ont un taux plus important à l'intérieur des nucléosomes que dans les linkers et les NIEBs (Drillon et al., 2016, Figure 4). La comparaison des taux de mutation entre l'humain et le chimpanzé (obtenus par comparaison des deux génomes et de deux groupes externes) avec ceux spécifiques à l'humain (obtenus à partir des données du projet 1000Genomes (Durbin et al., 2010)) suggère une sélection de ces patrons de mutations (Drillon et al., 2016, Figure 5). Il semble donc que le pourcentage GC oscillant observé aux bords des NIEBs chez l'humain soit une caractéristique sélectionnée et renforcée au cours de l'évolution des séquences. De plus, l'étude des mutations interrompant une séquence de type polyA ou polyT indique que ces interruptions sont favorisées au niveau des séquences nucléosomales. A l'inverse, les mutations menant à la formation de polyT ou polyA sont, elles, favorisées au niveau des séquences inter-nucléosomales (Drillon et al., 2016, Figure 6). Or, les séquences de type polyA et polyT ont été identifiées comme particulièrement inhibitrices de la formation du nucléosome (Segal & Widom, 2009b). Ainsi, la distribution de ces mutations suggère également un renforcement de la structure nucléosomale aux bords des NIEBs. Il semble donc que chez l'humain, la sélection des mutations aux bords des barrières nucléosomales participe au renforcement de l'oscillation du taux de GC à ces loci. Dans cette thèse, les patrons de mutations de la lignée du chimpanzé sont analysés pour déterminer si ce qui est observé dans la lignée humaine est retrouvé dans une lignée proche (**Chapitre 2**).

1.2 Les éléments transposables, moteurs de l'évolution des séquences

Dès 1957, Francis Crick a décrit ce qu'il a alors nommé le dogme central de la biologie moléculaire [1], dans lequel il associe un gène (une séquence d'ADN) à la production d'un ARN qui sert d'intermédiaire pour être traduit en protéine ayant une fonction précise dans la vie cellulaire (Cobb, 2017; Crick, 1958). Déjà il y a près de 65 ans, et alors que l'obtention de séquences protéiques était très difficile, et celle de séquences d'ADN indisponible pour encore au moins 20 ans, Crick pressentait qu'une grande partie de l'étude de l'évolution des espèces serait faite par comparaison de séquences (Cobb, 2017; Crick, 1958). Aujourd'hui, les progrès effectués dans le séquençage des génomes nous a permis de déterminer que chez l'humain, environ 2 % des 3.2 milliards de paires de bases constituant le génome sont considérées comme codantes, à savoir traduites en protéines fonctionnelles. C'est sur ces 2 % du génome, contenus dans les gènes, que s'est concentrée la grande majorité des études évolutives, le reste du génome ayant longtemps été considéré comme de "l'ADN poubelle". Petit à petit, on s'aperçoit néanmoins que la prétendue poubelle est en réalité emplie d'objets génomiques divers, on pourrait même dire, pour continuer l'analogie, qu'elle se vide au fur et à mesure qu'on identifie la fonction de ces différents objets. Par exemple, on sait aujourd'hui que certaines séquences éloignées de plusieurs dizaines de kilobases des gènes peuvent influencer leur expression (Smith & Shilatifard, 2014), ou que d'autres sont impliquées dans la formation de ce qui a été identifié comme de larges portions de génome présentant un même état chromatinien tout le long du domaine (McArthur & Capra, 2021; Tena & Santos-Pereira, 2021). L'importance fonctionnelle de la majorité des séquences génomiques, si elle existe, reste néanmoins à découvrir. Parmi ces séquences à la fonction inconnue, des éléments en particulier sont retrouvés à la fois en grand nombre dans les génomes et dans un très grand nombre de génomes, à tels points qu'ils sont parfois considérés comme des "envahisseurs de génomes"(Le Rouzic & Capy, 2005). Il s'agit des éléments transposables (ETs).

1.2.1 Qu'est-ce qu'un élément transposable ?

1.2.1.1 Une découverte précoce, acceptée tardivement

Dès 1950, Barbara McClintock détermine que des "éléments de contrôle" sont impliqués dans la variabilité de coloration des grains d'épis de maïs (McClintock, 1950). Mal reçue à cette époque, cette découverte sera remise au goût du jour dans les années 1970 avec la mise en évidence des séquences d'insertion chez *Escherichia coli* (Malamy et al., 1972), ainsi que de ce qui sera appelé "éléments transposables" chez l'humain et la drosophile (Ohno, 1972). Ces éléments ne codant pas toujours pour des protéines, et lorsque c'est le cas, ces protéines n'étant pas nécessaires au fonctionnement de la cellule, les éléments transposables ont d'abord été considérés comme de "l'ADN poubelle" (Biémont, 2010; Biémont & Vieira, 2006). Depuis, cette vision a été largement remise en cause, avec la mise en évidence de divers impacts fonctionnels des ETs, qui seront détaillés plus loin dans cette

1. L'appellation "dogme" est cependant un piètre choix de mot, admis par Crick lui-même, dans le sens où un dogme est, par définition, établi comme une vérité fondamentale et incontestable. Ici, on préférera parler de "théorie", qui peut par définition être confrontée aux données pour évoluer (ce qui fut le cas à la suite de la découverte de l'épissage alternatif qui a indiqué qu'un même gène pouvait être à l'origine de plusieurs protéines différentes), ou même être infirmée.

introduction (**Partie 1.2.2**). Cette remise en cause a notamment commencé dans les années 1990, lorsque le potentiel de ces séquences en terme d'évolution des génomes a été remarqué (Brosius, 1991; Makałowski et al., 1994). L'intérêt pour les ETs a ensuite grandi avec le séquençage du génome humain, puis l'apparition de nouvelles technologies de séquençage qui ont permis de mettre au point un génome de référence pour de nombreuses espèces. On y a découvert que la proportion d'ETs était très variable, de presque secondaire (~ 12 % chez le nématode *Caenorhabditis elegans* (C. elegans Sequencing Consortium, 1998)) à grandement majoritaire (plus de 80 % chez certaines plantes (Makałowski et al., 2019; SanMiguel et al., 1996)). Aujourd'hui, les éléments transposables sont reconnus comme des acteurs majeurs de l'évolution des séquences, et Barbara McClintock a finalement été récompensée du prix Nobel en 1983 pour leur découverte.

1.2.1.2 Définition d'un élément transposable

Les éléments transposables sont des séquences d'ADN ayant la capacité de se répliquer dans les génomes via un mécanisme appelé transposition. Cers séquences sont généralement répétées, avec un nombre de copies pouvant varier de une à plusieurs centaines de milliers voire plus d'un million pour certaines familles comme les éléments *Alu* chez l'humain (Deininger, 2011). Une classification hiérarchique de ces éléments a été mise au point à la fin des années 2000 (**Figure 1.2**) (Kapitonov & Jurka, 2008; Seberg & Petersen, 2009; Wicker et al., 2007). Elle comporte plusieurs niveaux qui seront détaillés dans la partie suivante. Post-insertion, certaines copies peuvent perdre leur capacité à se multiplier dans les génomes, par mutation de leur séquence. Elles peuvent tout de même parfois être mobilisées par d'autres copies du même élément qui, elles, sont restées intègres dans le génome. Enfin, certaines familles ne sont pas capable de transposer par elles mêmes, et nécessitent la mobilisation de la machinerie de transposition d'autres familles pour leur propre multiplication. C'est notamment le cas des éléments *Alu* qui utilisent la machinerie de transposition des éléments *LINE-1* pour se multiplier dans les génomes de primates. Ce mécanisme sera détaillé ultérieurement dans cette introduction.

1.2.1.3 Une classification hiérarchique des éléments transposables

Pour leur transposition, certains éléments sont d'abord transcrits en ARN avant d'être rétro-transcrits en ADN lors de l'insertion. D'autres peuvent se passer de cet intermédiaire ARN par exemple en excisant directement leur séquence d'ADN de leur locus originel et en l'insérant ailleurs dans le génome. La présence/absence d'un intermédiaire ARN lors de la transposition est à l'origine du premier niveau de classification des ETs (Kapitonov & Jurka, 2008; Seberg & Petersen, 2009; Wicker et al., 2007). Les ETs sont ensuite classifiés en ordre selon leur mécanisme d'insertion, puis en superfamilles selon la structure de leur séquence (notamment la présence et l'ordre d'apparition de certaines régions codantes) et la présence/absence ainsi que la taille de leurs duplications du site cible qui correspondent à de petites répétitions du site d'insertion générées lors de l'insertion de chaque côté de l'élément (TSD pour target site duplication) (Wicker et al., 2007). Enfin, les ETs sont divisés en familles, définies par la conservation de leur séquence d'ADN. Les familles d'ETs regroupent donc les copies issues d'une même séquence ancestrale. Pour certaines familles, particulièrement celles comportant un grand nombre de copies ayant pu largement diverger au

cours de l'évolution, des sous-familles sont établies, toujours en regroupant les copies selon leur similarités de séquence. Ces sous-familles peuvent elles même être encore subdivisées (on parle alors de sous-sous-familles et ainsi de suite) selon la complexité de l'histoire évolutive de l'élément.

La classe I : les rétrotransposons

La première classe d'éléments transposables regroupe les éléments capables de transposer par la transcription inverse de leur intermédiaire ARN en un ADN qui est inséré dans le génome à un nouveau locus. Ainsi, la séquence d'origine qui a été transcrite en un ARN intermédiaire reste en place à son locus d'origine. On parle donc d'un mécanisme de transposition de type "copier-coller" (Han & Boeke, 2005; Kumar & Bennetzen, 1999; Sabot & Schulman, 2006; SanMiguel et al., 1996). Les rétrotransposons peuvent être autonomes ou non-autonomes. Lorsqu'ils sont autonomes, ils codent directement pour les protéines nécessaires à leur mobilisation. Quant aux non-autonomes, ils sont mobilisés en *trans* par d'autres éléments autonomes. Par exemple, les éléments SINEs (Short Interspersed Nuclear Elements) sont mobilisés par les éléments LINEs (Long Interspersed Nuclear Elements) (Dewannieux et al., 2003; Kajikawa & Okada, 2002; Kramerov & Vassetzky, 2005). C'est notamment le cas de la famille des éléments *Alu*, qui sont des SINEs ayant besoin de la machinerie de transposition des éléments *LINE-1* (des LINEs) pour se multiplier (Deininger, 2011). Certains éléments autonomes peuvent aussi perdre cette autonomie, notamment par la mutation ou la perte des séquences codant pour les protéines nécessaires à leur transposition. Ces éléments, à l'image des éléments non-autonomes, peuvent également être mobilisés par leurs homologues autonomes qui codent à leur place les protéines nécessaires.

La classe des rétrotransposons est principalement constituée de deux superfamilles, différenciées par la présence/absence de longues répétitions terminales (LTRs pour "Long Terminal Repeats"). Les éléments à LTRs comportent ces répétitions de quelques centaines de nucléotides à chacune de leurs extrémités, lorsqu'ils sont complets (Wicker et al., 2007). Ils contiennent également des séquences codant pour les protéines nécessaires à leur transposition, notamment Gag qui permet la formation d'une capside dans laquelle se replie l'ARNm, et Pol, qui code à la fois pour une transcriptase inverse, une ribonucléase et une intégrase. Les superfamilles de l'ordre des éléments à LTRs diffèrent d'ailleurs principalement par l'ordre des séquences codantes inclues dans Pol. La transcription inverse des éléments à LTRs a lieu dans la capside formée après transcription et traduction de Gag (Makałowski et al., 2019), et l'ADN ainsi formé est ensuite directement inséré dans le génome, formant une nouvelle copie de l'élément.

L'autre grande catégorie d'ETs de classe I est celle des rétrotransposons non-LTRs. On y trouve notamment les LINEs et les SINEs, déjà évoqués précédemment, et dont certaines familles ont connu de formidables succès de transposition dans certaines espèces (plusieurs centaines de milliers de copies de LINE-1 et plus d'un million de copies d'Alu sont retrouvées dans le génome humain). La principale différence avec les éléments à LTRs réside dans le mécanisme de transposition. En effet, si les LINEs produisent également un ARN intermédiaire, celui-ci est directement rétro-transcrit au site d'intégration lors de l'insertion grâce à la transcriptase inverse et l'endonucléase encodées dans les éléments LINEs autonomes. Les SINEs suivent le même mécanisme, étant mobilisés en *trans* par les éléments LINEs autonomes. Le mécanisme de transposition des éléments *Alu*, qui seront étudiés en détail dans cette thèse, sera détaillé plus loin dans cette introduction.

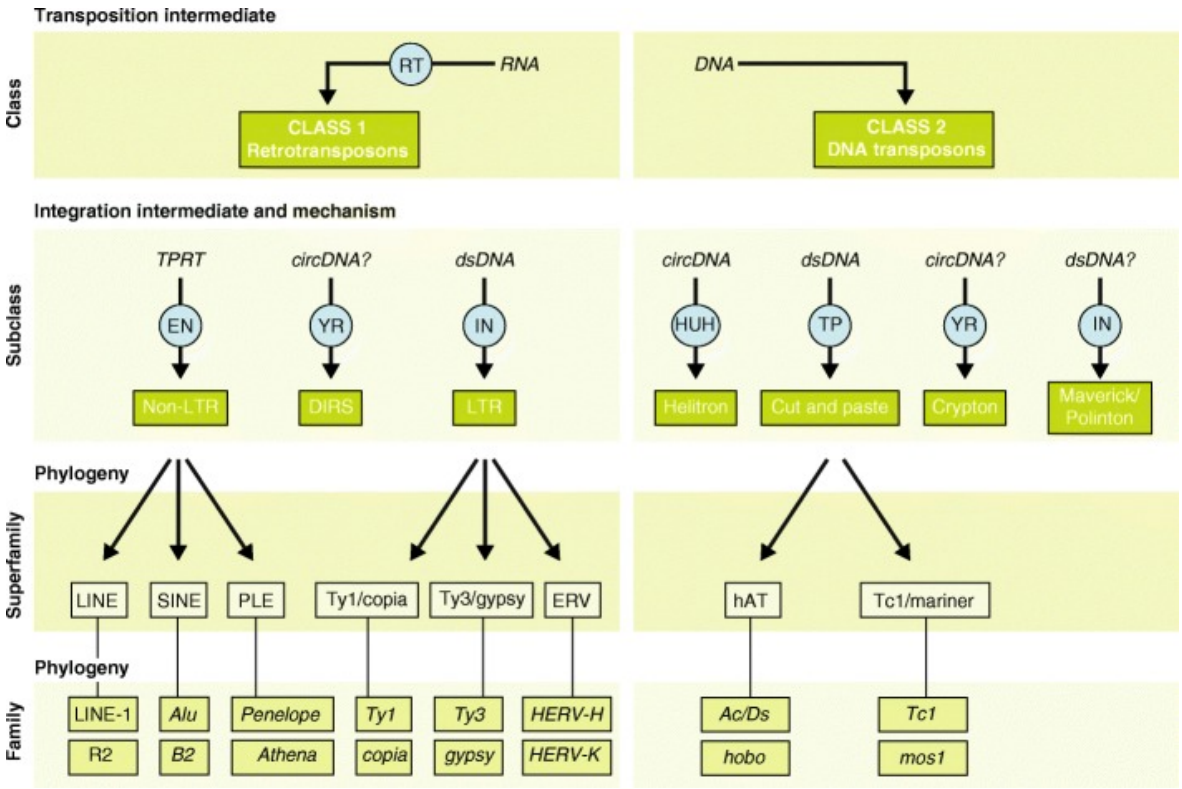


FIGURE 1.2 – **Classification des éléments transposables eucaryotes.** Schéma et exemples montrant les caractéristiques principales et la relation entre les classes, sous-classes, superfamilles et familles d'ETs. Les cercles bleus représentes les enzymes codées par des ETs. circDNA : ADN circulaire intermédiaire ; DIRS : Dictyostelium repetitive sequence ; dsDNA : ADN double-brin linéaire intermédiaire ; EN : endonucléase ; IN : intégrase ; PLEs : Penelope-like elements ; HUH : Protéines à activité endonucléase HUH ; RT : reverse-transcriptase ; TP : transposase ; TPRT : Target primed reverse transcription ; YR : tyrosine recombinase. © Bourque et al. (2018).

Si les éléments à LTRs et les LINEs/SINEs forment les deux principales catégories de rétrotransposons, d'autres ordres existent également. C'est le cas des éléments DIRS qui possédent une tyrosine recombinase à la place de l'intégrase présente dans les éléments à LTRs, et ne dupliquent pas leur site d'insertion (pas de TSD) (Cappello et al., 1985; Goodwin & Poulter, 2004; Wicker et al., 2007). C'est le cas aussi des éléments de type Penelope-like (PLE) qui codent pour une transcriptase inverse et une endonucléase comme les LINEs mais possèdent des séquences ressemblant à des LTRs (Evgen'ev & Arkhipova, 2005; Evgen'ev et al., 1997; Wicker et al., 2007).

La classe II : les transposons ADN

La seconde classe d'éléments transposables correspond à ceux n'ayant pas d'intermédiaire ARN dans leur mécanisme de transposition. Par opposition aux rétrotransposons, cette classe devrait s'appeler les "transposons". Néanmoins, cette appellation est déjà souvent utilisée par abus de langage pour parler des éléments transposables en général, incluant les rétrotransposons. Pour éviter les confusions, on précise donc souvent transposons "ADN" lorsqu'on veut parler des éléments appartenant à la classe II.

Cette classe est séparée en deux sous-classes, selon le nombre de brins d'ADN coupés lors de la transposition. La première sous-classe correspond principalement aux éléments de l'ordre des TIRs, qui sont caractérisés par la présence de répétitions terminales inversées (TIR, pour "Terminal

Inverted Repeats"). Ces éléments codent pour une transposase, qui reconnaît ces TIRs, et coupe les deux brins d'ADN à leur niveau, excisant complètement l'élément, pour l'insérer à un autre locus du génome. Ce mécanisme de transposition est identifié comme du "couper-coller". Pour se multiplier dans les génomes, ces éléments doivent donc effectuer leur transposition en parallèle de la réplication, en transposant d'un locus déjà répliqué à un locus non-répliqué (Greenblatt & Brink, 1962). Ainsi, à l'issue de la réplication, la nouvelle séquence possèdera les deux copies de l'élément. Les mécanismes de réparation post-excision permettent également parfois de conserver la copie au site d'excision (Nassif et al., 1994).

La deuxième sous-classe de transposons ADN correspond aux éléments dont le mécanisme de transposition ne coupe qu'un seul des deux brins d'ADN. Ainsi, ils se multiplient par un mécanisme de "copier-coller", mais sans passer par un intermédiaire ARN (Kapitonov & Jurka, 2001). On trouve notamment dans cette sous-classe les Helitrons et les éléments Maverick.

1.2.1.4 Une distribution variable au sein des génomes

La distribution des ETs au sein du génome est très variable selon les espèces et les familles d'ETs. Elle est la résultante d'un processus en deux étapes, à savoir d'abord les préférences d'intégration de l'élément, qui définissent où se font les insertions, puis la sélection post-insertion, qui élimine les insertions nocives pour ne conserver que celles qui sont neutres ou bénéfiques. A noter qu'il peut y avoir une conservation d'insertions délétères lorsqu'elles sont associées à des mutations bénéfiques mais ce ne sera pas détaillé ici. Selon les familles, les préférences d'intégration ainsi que la sélection post-insertion sont plus ou moins importantes dans la distribution finale des ETs dans le génome de l'hôte. On observe plusieurs cas de figures :

— Les préférences d'intégration peuvent être faibles, et associées à une sélection post-insertion forte. Dans ce cas, les éléments sont souvent insérés *de novo* uniformément tout le long du génome, mais la sélection post-insertion élimine les insertions de manière non-aléatoire, ce qui mène à des particularités dans la distribution de ces éléments dans les génomes. C'est notamment ce mécanisme qui explique que les *Alu* sont retrouvés dans les régions riches en GC et les *LINE-1* dans les régions pauvres en GC, alors que ces deux familles partagent le même mécanisme d'insertion (encodé par les éléments *LINE-1*). En effet, l'étude d'insertions *de novo* d'éléments *Alu* (Wagstaff et al., 2012) et *LINE-1* (Sultana et al., 2019) a montré que c'est principalement de la sélection post-insertion qui explique la différence de distribution entre les deux familles d'ETs (Sultana et al., 2017, 2019; Wagstaff et al., 2012). De la même manière, les ETs sont généralement cantonnés aux régions péricentromériques chez *Arabidopsis Thaliana*, et ce alors que les copies les plus récentes montrent une intégration équitablement distribuée le long des chromosomes, illustrant ici encore un effet important de sélection post-intégration (Quadrana et al., 2016).

— Les préférences d'intégration peuvent être très fortes, comme par exemple pour les éléments *R2*, des rétrotransposons non-LTR qui s'insèrent spécifiquement dans l'ADN ribosomal (Burke et al., 1987), ou encore les éléments HeT-A qui s'insèrent dans les télomères, expliquant à elles seules les distributions de ces éléments dans les génomes (Sultana et al., 2017).

— Enfin, il peut y avoir un équilibre entre préférences d'intégration et sélection. Certains ETs ont des préférences d'intégration marquées, mais qui sont assez faibles pour autoriser une

certaine dispersion des insertions, commes les éléments *Ty1*, qui ciblent à la fois l'ADN codant pour les ARNt et les séquences subtélomériques (Sultana et al., 2017). La sélection post-insertion peut alors jouer un rôle dans la distribution génomique des éléments en purifiant une partie des insertions.

Les préférences d'intégration associées aux différentes familles d'ETs sont très variables, et dépendantes de nombreux facteurs, qui ne seront pas décrits précisément ici mais qui ont fait l'objet d'une revue détaillée par Sultana et al. (Sultana et al., 2017). Parmi ces facteurs, il y a la séquence du site d'insertion, qui peut être plus ou moins variable selon le mécanisme d'insertion. En jeu se trouve aussi le contexte chromatinien, avec certains ETs préférant la chromatine ouverte, et d'autres la chromatine plus fermée, certains les séquences dépourvues de nucléosomes et d'autres les séquences nucléosomales. Enfin, les mécanismes de transposition et pas seulement d'insertion, impliquant notamment l'accès des protéines nécessaires à la transposition au noyau, et donc au génome, sont également déterminant pour les préférences d'intégration.

1.2.2 L'impact des éléments transposables sur les génomes

1.2.2.1 La valeur sélective

Avant de parler d'impact d'une insertion d'un ET dans un génome, il est nécessaire de définir le terme de valeur sélective (ou *fitness* en anglais). Il s'agit de la capacité d'un individu à transmettre son patrimoine génétique à sa descendance. Elle est le produit de la survie par la fécondité. En d'autres termes, pour qu'un individu transmette ses gènes à la génération suivante, il doit survivre au moins jusqu'à être en âge de se reproduire, puis effectuer cette reproduction. Cette valeur sélective est fortement influencée par le génome de l'individu. Aussi les mutations de ce génome peuvent avoir un impact sur celle-ci. La transposition peut donc avoir une grande influence sur la valeur sélective de l'hôte.

1.2.2.2 Des effets parfois délétères ...

Historiquement, les éléments transposables ont longtemps été considérés comme des parasites génomiques qui ne faisaient qu'apporter toujours plus "d'ADN poubelle" dans les génomes, sans aucun bénéfice (Ohno, 1972). Cette vision a été étayée par la découverte d'une variété de mutations délétères pouvant être générées par l'insertion d'ETs (Kazazian, 1998). Elle a depuis largement été remise en cause par la découverte d'effets bénéfiques liés à l'insertion d'ETs qui seront détaillés dans la partie suivante (Biémont, 2010).

Les effets délétères des ETs peuvent aller de la mutation létale à une diminution plus ou moins importante de la valeur sélective d'un individu. Les sources de ces effets peuvent être variables. La plus évidente est l'insertion d'un ET directement dans la partie codante d'un gène essentiel à la survie de l'hôte faisant apparaître un codon stop prématuré dans le cadre de lecture par exemple, ou créant des décalages de cadre de lecture (Hancks & Kazazian, 2016). Mais les ETs peuvent avoir toute une variété d'effets. Ils peuvent par exemple s'insérer dans les régions régulatrices de gènes de l'hôte et entraîner des modifications de l'expression qui peuvent avoir un impact important sur la valeur sélective selon le gène et l'importance du changement d'expression. Les effets délétères des

ETs ne sont également pas toujours liés à l'insertion en elle même mais à leur nature répétée. En effet, la similarité de séquence entre les copies d'ETs peut entraîner de la recombinaison homologue non-allélique, et provoquer de grands réarrangements génomiques comme des délétions et des duplications (Hancks & Kazazian, 2016).

Les effets délétères des ETs ont été largement confirmés notamment chez l'humain par l'association de certaines insertions avec des maladies génétiques. Par exemple, deux insertions d'éléments SVA dans l'intron 8 du gène *NF1* ont été associée à des délétions de 867 kilobases et 1 mégabase entrainant une inactivation du gène et provoquant la neurofibromatose de type I chez les individus porteurs d'une de ces délétions (Hancks & Kazazian, 2016; Vogt et al., 2014). Également, l'hyperactivité des éléments LINE-1 dans les neurones a été associée à la susceptibilité d'apparition de schizophrénie. Enfin, les éléments *Alu*, particulièrement prospères chez les primates, sont associés avec environ 0.1 % des maladies génétiques humaines (Kim et al., 2016).

1.2.2.3 Et parfois bénéfiques . . .

Si beaucoup d'insertions d'ETs ont été identifiées comme ayant un effet plus ou moins délétère pour leur hôte, il arrive parfois que des insertions aient un impact positif sur la valeur sélective des individus. Ce type d'évènement est appelé exaptation, et place les ETs comme une source importante d'innovations évolutives (**Figure 1.3**) (Etchegaray et al., 2021). A l'image des effets délétères pouvant être variables, les effets bénéfiques peuvent être de plusieurs types. On détaillera ici les processus de domestication, d'exonisation et les effets de régulation de gènes. Néanmoins, cette liste n'est pas exhaustive et les ETs peuvent être également, entre autres, liés à la formation de rétrogènes ou à la plasticité génomique des espèces (Schrader & Schmitz, 2018).

La domestication, un apport de nouveaux gènes

Une grande partie des ETs comportent des séquences codant pour des protéines nécessaires à leur transposition. Durant l'évolution des génomes, il est arrivé de manière récurrente que ces protéines, dont la fonction première est liée à la transposition, soient recrutées pour servir au fonctionnement des cellules de l'hôte. C'est ce qu'on appelle la domestication d'éléments transposables. Ce mécanisme entraîne la formation de nouveautés génétiques directement issues d'insertion d'ETs. Les exemples sont nombreux, et ont notamment été passés en revue dans la publication d'Etchegaray et al. (Etchegaray et al., 2021). On peut citer celui des télomères de la drosophile, qui sont maintenus non pas par une télomérase mais par deux éléments transposables domestiqués (HeT-A et TART), qui ajoutent leurs LTRs au niveau des régions terminales des chromosomes pour compenser la perte de nucléotides associée à la réplication (Pardue & DeBaryshe, 2003). Un autre exemple de domestication d'ETs est celui des syncitines chez les mammifères, qui sont issues de multiples événements indépendants de domestication du gène codant pour l'enveloppe d'un rétrovirus endogène (ERV pour Endogenous RetroVirus). La domestication de ce gène permet la fusion cellulaire lors du développement placentaire (Etchegaray et al., 2021; Kaneko-Ishino & Ishino, 2012).

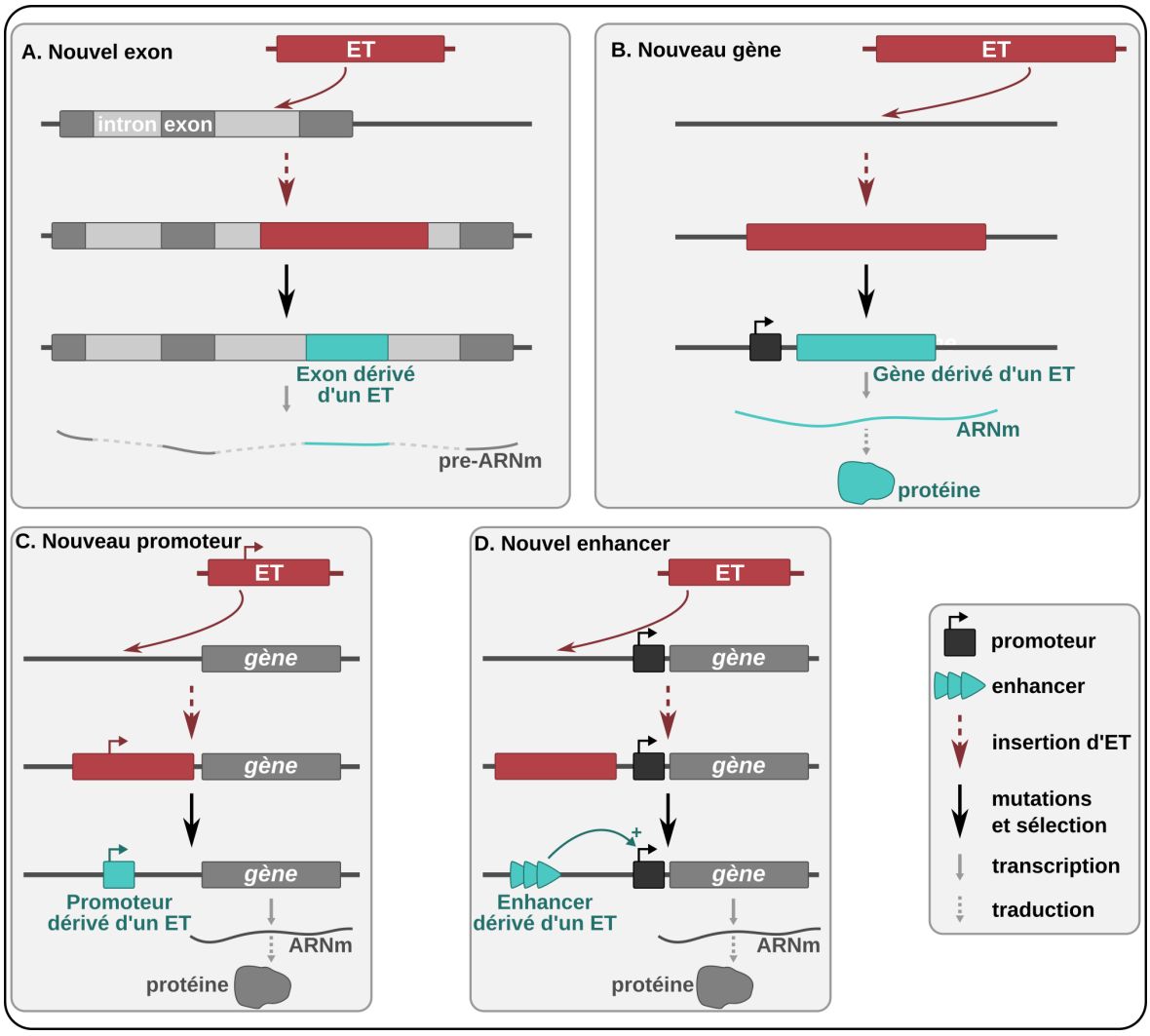


FIGURE 1.3 – **Exemples d'exaptation d'éléments transposable.** Après l'insertion d'un ET : (**A**) dans un intron, une partie de l'ET peut devenir un nouvel exon (exonisation). Les sites d'épissages peuvent être déjà présents dans la séquence de l'ET ou acquis par mutation. (**B**) Une partie de l'ET peut former un nouveau gène chez l'hôte, et être transcrit grâce à un promoteur de l'hôte dans les séquences flanquant l'insertion ou à un promoteur dérivé de la séquence de l'ET. (**C-D**) en amont d'un gène, l'ET peut former un nouveau promoteur (C) ou amplificateur (enhancer, D). On note que ce modèle fonctionne aussi pour la répression de gènes associée à des ETs. Les ETs sont représentés en rouge, et les séquences d'ETs exaptées en bleu. Adapté d'Etchegaray et al., 2021 (Etchegaray et al., 2021).

L'exonisation, source de nouveaux exons

Si l'insertion d'ET dans un gène peut être délétère par disruption de la séquence codante ou des séquences régulatrices, elle peut aussi parfois être bénéfique, notamment lorsqu'elle permet d'apporter de nouveaux exons dans un mécanisme appelé exonisation. Par exemple, l'insertion d'un élément SINE de type MIR-b est associée à deux isoformes du gène IGF-1 spécifiques aux mammifères par l'ajout d'un cinquième exon au gène (Annibalini et al., 2016). L'exonisation d'ETs est d'ailleurs un processus récurrent chez les vertébrés, et on estime qu'environ 1 % des exons humains sont dérivés d'éléments transposables (Piriyapongsa et al., 2007; Sela et al., 2007).

Les éléments transposables comme régulateurs de gènes

Afin de limiter l'impact de la transposition, les ETs sont souvent contrôlés par le génome hôte, par exemple par l'ajout de marques épigénétiques empêchant l'expression des éléments nouvellement

FIGURE 1.4 – **Photographies des deux types de coloration de la phalène du bouleau (*Biston betularia*).** (**A**) Forme *typica*, présente avant la révolution industrielle en Angleterre. (**B**) Forme *carbonaria*, apparue pendant la révolution industrielle en adaptation au noircissement du tronc des bouleaux, à la suite de l'insertion d'un ET au niveau du gène *cortex*. ©Olaf Leillinger, Vincent Guili. CC-BY-SA 4.0

insérés (Barau et al., 2016; Jansz, 2019; Lisch, 2009; Xie et al., 2013). Ce *silencing* des ET peut également affecter les régions adjacentes au site d'insertion. Ainsi, une insertion proche d'une séquence régulatrice de gène peut, sans modifier cette séquence, avoir un impact sur l'expression du gène associé (Rebollo et al., 2012a).

Les ETs contribuent également directement à l'apport de séquences régulatrices telles que des sites de fixation de facteurs de transcription. Ces sites, présents à l'origine pour permettre l'expression de l'ET lui-même, peuvent également influencer l'expression des gènes situés dans le voisinage de l'insertion. De plus, des ETs peuvent aussi être porteurs de séquences proches de ces sites de fixations, ne nécessitant que peu de mutations pour les rendre fonctionnelles, là où la création d'un site *de novo* nécessite un grand nombre de mutations ponctuelles au même endroit. Ainsi, ces propriétés de transport de séquences régulatrices confèrent aux ETs un avantage sélectif en améliorant la valeur sélective de l'hôte, ce qui a probablement favorisé leur persistance au cours de l'évolution (Bourque, 2009; Jacques et al., 2013; Marnetto et al., 2018; Sundaram & Wysocka, 2020). Elles positionnent également les ETs comme potentiellement impliqués dans l'évolution des réseaux de régulation de gènes, où des facteurs codés par certains gènes agissent sur l'expression d'autres gènes, formant parfois des cascades de régulation où chaque gène est activé par le produit du gène précédent. En effet, en dispersant des séquences régulatrices dans les génomes, les ETs pourraient permettre de modifier rapidement la régulation de tout un réseau de gènes. C'est notamment le cas d'éléments ERV comportant dans leur LTRs un site de fixation pour la protéine p53 impliquée dans le processus d'apoptose. Ces éléments ont participé à la dispersion de ce site de fixation dans le génome humain (Wang et al., 2007).

L'exemple de la phalène du bouleau, largement utilisé pour illustrer le principe de sélection naturelle, est d'ailleurs également un excellent exemple de régulation de gène par l'insertion d'ET (**Figure 1.4**). La phalène du bouleau est un papillon nocturne dont la couleur des ailes imite celle du tronc des bouleaux, ce qui lui permet de se camoufler de ses prédateurs en se posant sur ces arbres. Suite à une industrialisation massive au Royaume-Uni, la pollution au charbon noircit le tronc des bouleaux. Les phalènes aux ailes claires ne sont alors plus camouflées en se posant sur ces arbres, ce qui a induit une sélection des individus aux ailes plus foncées. Ainsi, dans les zones polluées où les bouleaux ont le tronc noir, on a observé des phalènes aux ailes noires également, là où dans les zones moins polluées, elles ont conservé leur couleur initiale. On sait aujourd'hui que le phénotype des ailes sombres est induit par une forte expression du gène *cortex*. Or, c'est l'insertion

d'un ET dans le premier intron du gène qui a été identifiée comme la cause de la surexpression du gène, même si le mécanisme associé est encore inconnu (Van't Hof et al., 2016). Cet exemple illustre le potentiel régulateur des ETs, et place ces éléments comme des sources d'adaptation rapide face à un environnement variable.

1.2.2.4 Mais principalement neutres . . . Jusqu'à preuve du contraire

Si dans les deux sections précédentes, on a pu détailler toute une variété d'effets non-neutres associés aux insertions d'ETs, il faut néanmoins rappeler que de telles influences des ETs sur la valeur sélective des individus restent l'exception plutôt que la règle (Schrader & Schmitz, 2018). La plupart des insertions d'éléments transposables sont identifiées comme neutres d'un point de vue évolutif. Cette identification est sujette à débat, car ce n'est pas parce qu'on n'a pas encore identifié d'effet pour une copie d'un élément que l'insertion n'en a aucun. Par exemple, si les effets de séquences comme l'exaptation des ETs ou la disruption de séquences géniques ou régulatrices par les insertions d'ETs ont été au centre de l'attention jusqu'ici, l'impact des insertions d'ETs sur la structure des génomes, particulièrement la structure 3D, et le repliement de l'ADN en nucléosome n'a été que très peu étudié. Or, on a vu dans la première partie de cette introduction que l'évolution de la séquence pouvait mener à des modifications de la structure du génome avec des conséquences directes sur le phénotype des individus (Barbier et al., 2021, Section 3). Les ETs étant des moteurs très importants de l'évolution des séquences, leur implication potentielle dans l'évolution de la structure de la chromatine est à considérer très sérieusement. Jusqu'à présent, la mise en évidence d'une importance fonctionnelle des ETs a majoritairement été faite à l'échelle de la copie unique, ou de quelques copies. L'impact des ETs dans les génomes pourrait, en plus de ces effets "ponctuels", être plus global. On note par exemple que la prolifération d'ETs est corrélée à l'augmentation de la taille des génomes, ce qui a été remarqué chez les drosophiles (Sessegolo et al., 2016), les urochordées (Naville et al., 2019), de manière plus générale chez les vertébrés (Chalopin et al., 2015), les plantes (Vitte & Panaud, 2005), et s'étend même aux eucaryotes (Chénais et al., 2012; Kidwell, 2002). Dans cette thèse, la relation entre insertion d'ET et structure nucléosomale sera étudiée à travers l'exploration de la relation entre éléments transposables *Alu* et les barrières nucléosomales (**Chapitre 4**).

1.2.3 Les éléments Alu, une famille d'éléments transposables qui a réussi

Les éléments Alu ont été découverts en 1979 (Houck et al., 1979), et doivent leur nom à la présence dans leur séquence de sites de restriction pour l'enzyme AluI de la bactérie *Arthrobacter luteus*, ce qui a aidé à leur découverte(Stenz, 2021). Parmi les différentes familles de SINEs, les éléments Alu forment l'une des plus prospères, avec un nombre de copies dépassant le million dans plusieurs espèces de primates, dont ils sont d'ailleurs spécifiques. Les quelques 1.2 millions de copies chez l'humain représentent à elles seules plus de 10 % de notre génome, ce qui fait aujourd'hui des Alu les ETs les plus abondants du génome humain (Batzer & Deininger, 2002).

1.2.3.1 Qu'est-ce qu'un élément Alu ?

A l'origine, un ARN 7SL

L'ARN 7SL, codé chez l'humain par le gène RN7SL1, est traduit en une protéine appelée SRP (pour Signal Recognition Particle), essentielle pour la translocation des protéines sécrétées (Stenz, 2021). La comparaison de la séquence de cet ARN avec celle des Alu a mis en évidence que cette famille d'ETs dérive de cet ARN. Cet ARN 7SL est à l'origine d'un monomère ancêtre des Alu (FAM, pour Fossil Alu Monomer), qui a ensuite dérivé en deux séquences distinctes, les FRAM (Free Right Alu Monomer) et les FLAM (Free Left Alu Monomer), qui sont sensiblement similaires, exceptée la présence d'une séquence de 11 pb issue du gène 7SL spécifiquement dans les FRAM (Stenz, 2021). La fusion de ces deux monomères est à l'origine des Alu, qui sont donc des éléments dimériques (Quentin, 1992). Pendant 65 millions d'années, ces éléments se sont propagés dans les génomes de primates (Kriegs et al., 2007). Ils y sont encore actifs aujourd'hui, et on estime qu'une insertion d'Alu survient toutes les vingt naissances (Cordaux et al., 2006).

Trois familles, et de nombreuses sous-familles

La comparaison des séquences d'Alu entre elles a mené à l'établissement de trois grandes familles d'Alu, à savoir les AluJ, les AluS et les AluY. Les AluJ sont identifiés comme les éléments les plus anciens, c'est-à-dire ceux insérés le plus tôt dans l'évolution des primates. Ils ont pour beaucoup perdu leur capacité de transposition à cause de l'accumulation de mutations (Stenz, 2021). La famille des AluS, dont l'âge est estimé à environ 30 millions d'années, contient quelques éléments encore actifs. Cependant, c'est dans la famille des AluY, qui sont apparus le plus récemment, il y a 2 à 4 millions d'années, que l'on retrouve le plus d'éléments actifs, particulièrement dans la sous-famille AluYb (Ahmed et al., 2013; Martinez-Gomez et al., 2020). Aujourd'hui, environ 6000 copies d'éléments Alu sont actifs chez l'humain, appartenant aux sous-familles Y et S (Bennett et al., 2008). En plus de la classification en trois grandes familles, les éléments Alu sont séparés en un certain nombre de sous-familles. Les AluJ sont composés de 2 sous-familles, les AluS de 6, et les AluY de 23, soit 31 sous-familles en tout. L'arbre phylogénétique des Alu ainsi formé est disponible en **Figure 1.5**. L'histoire évolutive des éléments Alu pourrait cependant être beaucoup plus complexe, des comparaisons de séquences plus poussées mettant en évidence une séparation en 213 sous-familles plutôt que 31 (Price et al., 2004). Dans cette thèse, nous nous en tiendrons à la séparation en 3 grandes familles précédemment cités (J, S et Y).

Une structure dimérique et un mécanisme de transposition emprunté aux LINE-1

Les éléments Alu sont composés de deux monomères riches en GC séparés par un court polyA interne. A l'extrémité 3' de l'élément, on retrouve un polyA terminal, plus long que le premier, et de taille assez variable (de quelques paires de bases à quelques dizaines) (**Figure 1.6 - A**). Le corps d'un élément Alu est long d'environ 280 pb, et l'élément entier (corps + polyA terminal) a une taille d'environ 300 pb. Sur le premier bras de l'élément, on retrouve les deux composants d'un promoteur de l'ARN polymérase III, ce qui permet la transcription des Alu. Il est à noter que ces éléments ne contiennent pas de terminateur de l'ARN polymérase III, aussi la transcription d'un Alu se termine en aval de l'élément lorsqu'une séquence TTTT est rencontrée. Ainsi, chaque transcrit issus d'un Alu est unique :

— Par les mutations présentes sur l'élément lui même.

— Par la taille du polyA de l'élément.

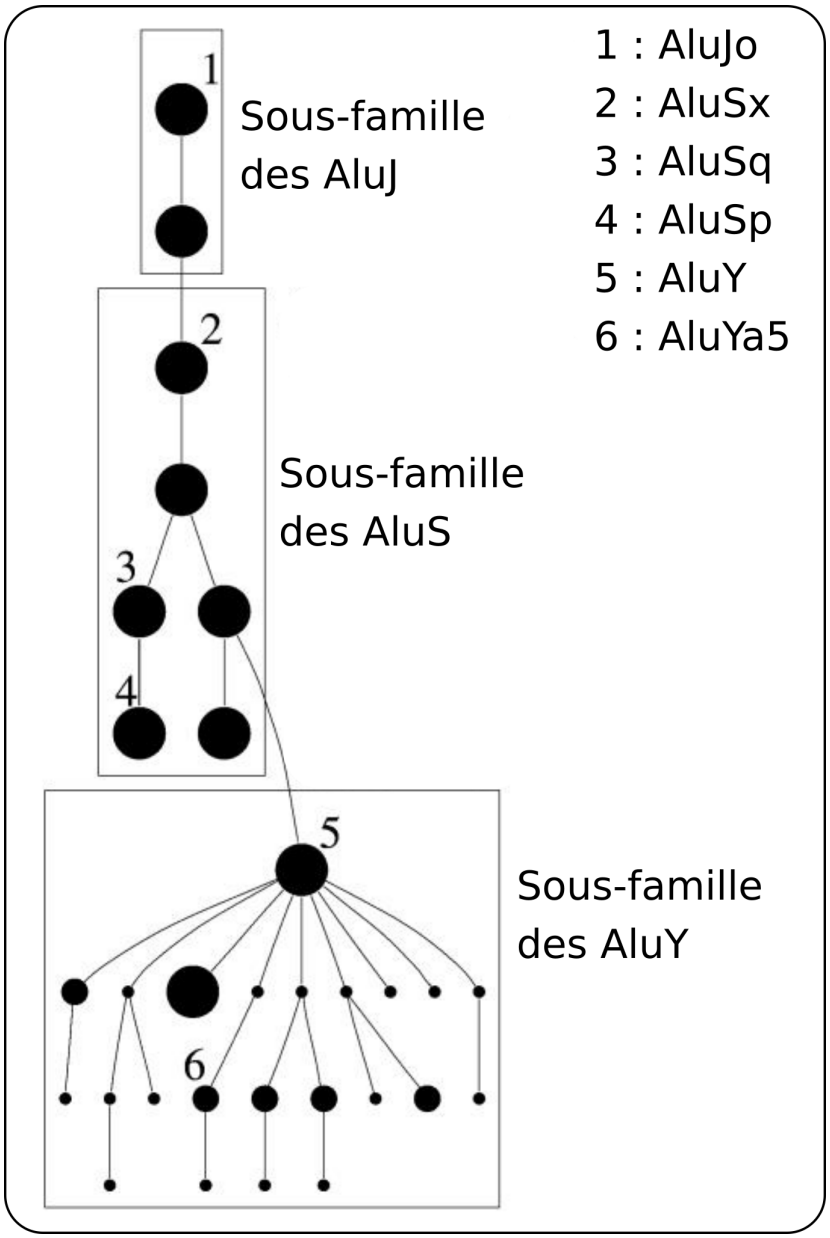


FIGURE 1.5 – **Arbre phylogénétique des 31 sous-familles d'Alu.** Les noeuds les plus gros représentent les sous-familles comportant au moins 10 000 éléments. Les noeuds les plus petits représentent les sous-familles comportant moins de 1000 éléments. Les noeuds intermédiaires représentent donc les sous-familles comportant entre 1000 et 10 000 éléments. Les cadres correspondent aux trois familles d'Alu (J, S et Y). Adapté de Price et al. (2004).

— Par la séquence présente en 3' qui varie d'un élément à l'autre en fonction de son site d'insertion.

Une fois transcrit, l'ARN issu d'un élément Alu peut être intégré au génome grâce aux protéines produites par les éléments LINE-1 pour leur propre transposition. Cependant, il est à noter que si la transposition des LINE-1 nécessite à la fois l'expression de leur ORF1 et de leur ORF2, celle des Alu ne nécessite que les protéines issues de l'ORF2 (Dewannieux et al., 2003; Roy-Engel et al., 2002). Cela pourrait d'ailleurs expliquer pourquoi le nombre de copies d'Alu est presque deux fois plus important que celui de LINE-1 dans les génomes (Deininger, 2011). En effet, il existe des éléments LINE-1 ne produisant que les protéines issues de leur ORF2, contribuant donc directement à la mobilisation en *trans* les éléments Alu mais nécessitant l'expression de LINE-1 complets pour leur

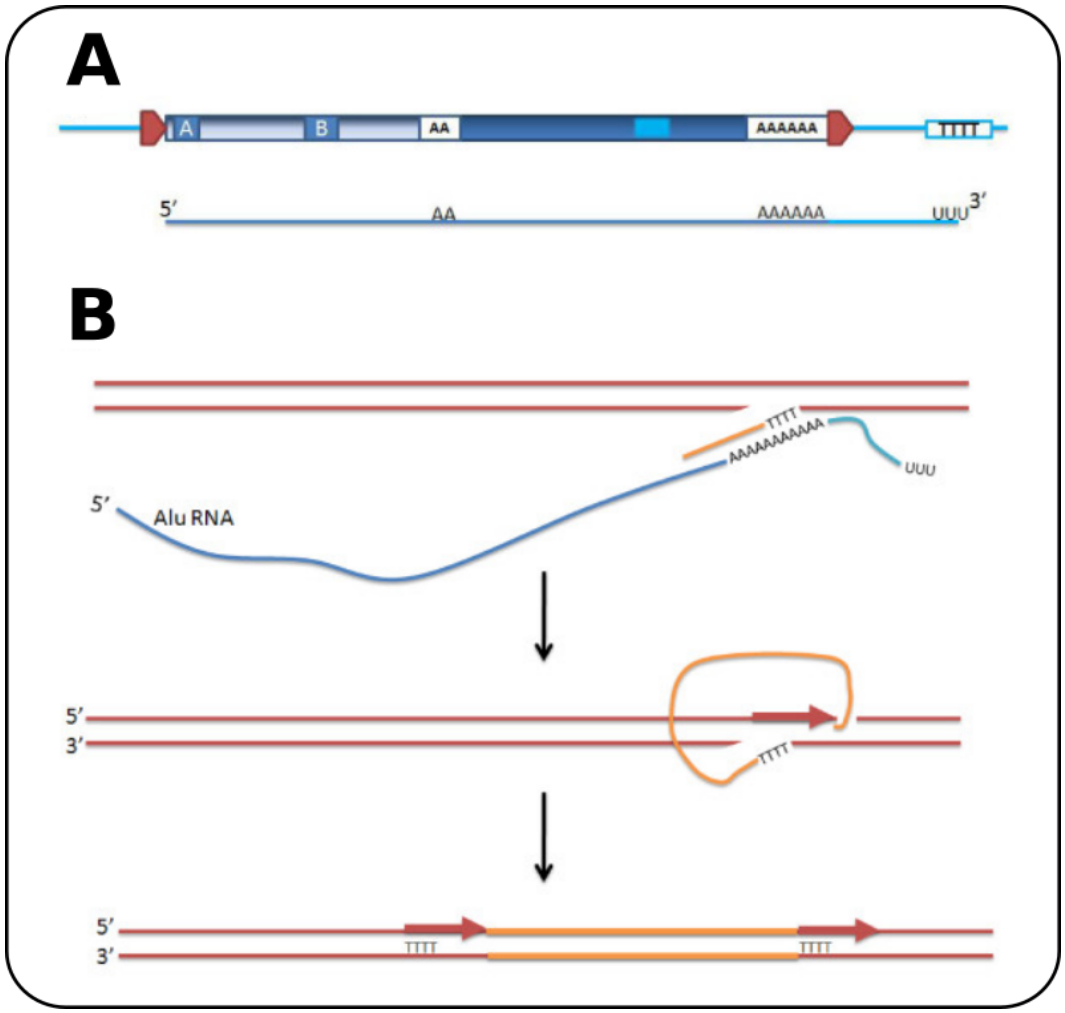


FIGURE 1.6 – **Schéma de la structure et du mécanisme d'insertion des éléments Alu.** (**A**) Structure d'un élément Alu. Sont représentées les deux séquences polyA (interne et terminale), ainsi que le promoteur de l'ARN polymérase III (boîtes A et B). La séquence polyT en aval de l'élément représente le terminateur de la transcription qui n'est pas inclus dans l'élément Alu (**Partie 1.2.3.1**). Sous le schéma de l'élément Alu génomique est représenté un transcrit d'Alu canonique, avec sa queue polyA ainsi que la région 3' spécifique de chaque élément (en bleu clair). (**B**) Représentation schématique de la TPRT (target-primed reverse transcription). L'ARN d'Alu amène l'ORF2p à cliver le génome au niveau d'une séquence consensus riche en T, servant de primer à la transcription inverse après appariement du polyA de l'élément Alu. Cela produit un ADNc d'Alu. La seconde partie du schéma représente le clivage du second brin d'ADN, dont le mécanisme est inconnu, ainsi que la synthèse du second brin utilisant comme modèle l'ADNc précédemment synthétisé. L'élément Alu final (3ème partie) est alors flanqué de TSDs (target site duplications) correspondant à la séquence entre le site de clivage du premier et du second brin. Ces TSDs sont représentés ici par les flèches rouges. Adapté de Deininger (2011)

propre mobilisation. D'autres différences dans les mécanismes de rétrotransposition de ces deux familles d'ETs pourraient participer à l'inégalité du nombre de copies. Elles ont fait l'objet de revues détaillées, par Prescott Deininger ainsi que par Catherine Ade et al. (Ade et al., 2013; Deininger, 2011). L'intégration de l'élément Alu est faite par un mécanisme appelé TPRT (pour Target Primed Reverse Transcription, **Figure 1.6 - B**), durant lequel une séquence cible riche en nucléotides T est reconnue et coupée par l'endonucléase de l'ORF2 des LINE-1, puis l'Alu s'apparie via son polyA terminal à la séquence cible et est rétrotranscrit en ADN par la transcriptase inverse de l'ORF2 des LINE-1 (Luan et al., 1993; Stenz, 2021). Le mécanisme par lequel la synthèse du second brin d'ADN est amorcée est encore inconnu (Deininger, 2011).

Un élément enrichi dans les régions riches en GC

La distribution des Alu dans le génome humain n'est pas homogène. En effet, ces éléments sont enrichis dans les régions riches en gènes, ce qui peut être étonnant sachant qu'ils utilisent la machinerie de transposition des éléments LINE-1 qui sont, eux, enrichis dans les régions pauvres en gènes (Lander et al., 2001). Cette différence s'explique par de la sélection post-insertion plutôt que par des préférences d'intégration (**Partie 1.2.1.4**). La contre-sélection des insertions d'éléments LINE-1 plus forte que celle des éléments Alu au niveau des gènes pourrait d'ailleurs s'expliquer par la différence de taille entre les deux familles d'éléments (un éléments LINE-1 complet est presque 20 fois plus grand qu'un élément Alu) (Deininger, 2011).

1.2.3.2 Des éléments aux effets génomiques divers

A l'image des éléments transposables en général dont les effets ont été présentés précédemment (**Partie 1.2.2**), toute une variété d'effets ont été décrits pour l'insertion d'éléments Alu dans les génomes, certains délétères, d'autres bénéfiques. Cependant, comme pour les ETs, l'influence des éléments Alu sur la valeur sélective n'est décrite que pour quelques copies. Ici, nous allons détailler quelques-uns des exemples qui montrent que l'insertion d'un élément Alu peut ne pas être neutre d'un point de vue évolutif. Il est cependant à noter que les éléments Alu peuvent également avoir une activité somatique, qui ne sera pas détaillée ici, mais qui pourrait être associée par exemple au développement de certains cancers, même si l'association est moins claire que pour d'autres familles d'ETs comme les LINE-1 (Ade et al., 2013).

Des influences multiples sur la transcription

Les éléments Alu sont majoritairement situés dans des régions riches en gènes, même si cette distribution semble plutôt due à de la sélection post-insertion qu'à une préférence d'intégration. Toujours est-il que leur proximité avec les régions géniques multiplie les possibilités d'interaction entre éléments Alu et expression des gènes. Les Alu peuvent influencer l'expression des gènes de nombreuses manières. Tout d'abord, ils peuvent être insérés directement dans les régions exoniques, ce qui amène généralement à une disruption du gène (Kim et al., 2016). Les éléments Alu insérés au niveau des introns peuvent également avoir une influence sur la fonction du gène associé, en provoquant de l'épissage alternatif via les multiples sites donneurs et accepteurs situés sur l'élément, ou bien l'acquisition de tels sites post-insertion par mutations ponctuelles (**Figure 1.7**). Cela peut amener à une exonisation de l'élément Alu, ou encore à de la rétention d'intron, modifiant voire rendant non fonctionnelle la ou les protéines associées au gène (Kim et al., 2016). Les insertions d'Alu dans les gènes peuvent également amener à un "saut d'exon" (exon skipping) où un des exons du gène est ignoré lors de l'épissage de l'ARN messager. Ils peuvent également amener à l'apparition de nouveaux sites de polyadénylation, dont trois sont présents sur l'élément, dont deux actifs chez l'humain (Kim et al., 2016). Enfin, la présence d'un élément Alu dans l'ARN messager est également associée à la désamination des adénosines dans l'ARN (ADAR, pour adenosine deamination that acts on RNA), un processus pouvant amener à de l'épissage alternatif ou à de la rétention de l'ARN messager dans le noyau, empêchant alors sa traduction (Deininger, 2011; Kim et al., 2016). De nombreux exemples ont été documentés illustrant tous ces effets potentiels, passés en revue plusieurs fois durant ces dernières années (Burns, 2020; Gussakovsky & McKenna,

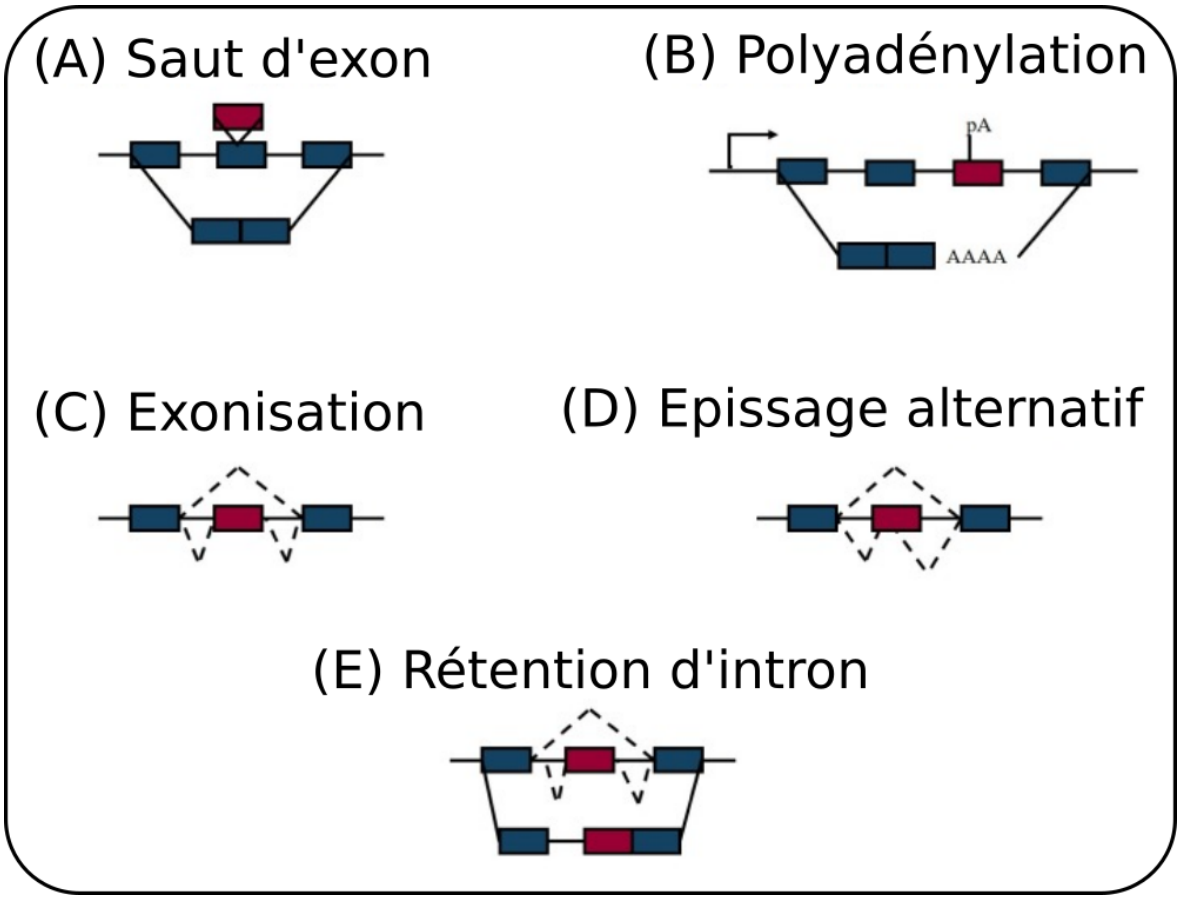


FIGURE 1.7 – **Schéma des impacts des éléments Alu sur la transcription alternative.** L'insertion d'Alu au niveau d'un gène peut avoir des effets variés. L'insertion dans un exon peut provoquer un saut d'exon (**A**). L'insertion dans un intron peut provoquer l'apparition d'un site de polyadénylation (**B**), l'exonisation de tout ou d'une partie de l'éléments (**C**), l'apparition de sites d'épissage alternatifs (**D**) ou encore de la rétention d'intron (**E**). Les éléments Alu sont représentés ici en rouge, les exons en bleu. Les lignes pointillées indiquent l'épissage alternatif. Adapté de Kim et al., 2016 (Kim et al., 2016).

2021; Kim et al., 2016; Ule, 2013). On peut notamment citer celui d'une insertion d'un AluYb9 dans l'intron 18 du gène du facteur VIII, menant au saut de l'exon 19 lors de l'épissage de l'ARN messager, identifié comme une cause de l'hémophilie A (Ganguly et al., 2003), ou encore l'insertion d'un Alu dans le troisième intron du gène OAT (ornithine aminotransferase, qui code pour une enzyme transformant l'ornithine en glutamate), amenant à un épissage alternatif modifiant la protéine ce qui entraîne une atrophie du choroïde et de la rétine (Mitchell et al., 1991).

Des transporteurs de sites de fixation de facteurs de transcription

Les éléments Alu peuvent également être impliqués dans des processus de régulation des gènes, notamment car ils sont porteurs de plusieurs sites de fixation de facteurs de transcription, dont certains sont spécifiques à des familles d'Alu, ainsi que d'un promoteur pour l'ARN polymérase III (Deininger, 2011; Polak & Domany, 2006). Plusieurs exemples viennent illustrer le potentiel des Alu comme transporteurs de sites de fixation de facteurs de transcription. Une sous-famille de ces éléments a notamment été identifiée comme pouvant avoir une activité d'amplificateurs liés aux récepteurs de l'œstrogène (estrogen receptor-dependent enhancers) (Norris et al., 1995). Une autre sous-famille d'Alu contient, elle, un site de fixation pour les récepteurs de l'acide rétinoïque (Vansant & Reynolds, 1995). Les éléments Alu peuvent donc évoluer pour contenir une variété

de sites de fixation pour des récepteurs nucléaires (Cotnoir-White et al., 2011), mais également pour d'autres facteurs de transcription comme NF-κB (un régulateur de la réponse immunitaire), dont les sites de fixations sont dispersés dans les génomes de primates par les éléments Alu (Antonaki et al., 2011). Cette dernière association illustre le rôle que pourrait jouer les Alu, et plus généralement les ETs, dans l'élargissement du répertoire de sites de fixation potentiels pour des facteurs de transcription dans les génomes, avec pour conséquence d'ajouter rapidement de nouveaux gènes dans un réseau de régulation. Pour finir, la présence le long des éléments Alu de nombreux dinucléotides CpG, un contexte permettant la méthylation des cytosines qui favorise les mutations C vers T, est associée à la formation de nombreux sites de fixation pour le suppresseur de tumeur p53 (Zemojtel et al., 2009).

Une source importante de recombinaison

Le nombre très important de copies d'éléments Alu dans les génomes et la conservation des séquences entre éléments Alu en fait une source majeure de recombinaisons non-alléliques, ce qui peut amener à de grands réarrangements et à la délétion de larges portions des génomes. Chez l'humain, plus de 500 délétions liées aux éléments Alu ont été documentées, la grande majorité dues à de la recombinaison non-allélique/ectopique entre ces éléments (Kim et al., 2016). En effet, la comparaison des génomes de l'humain et de celui du chimpanzé a mis en évidence que près de 400 000 pb ont été délétées chez l'humain en lien avec la recombinaison induite par les éléments Alu (Callinan et al., 2005; Kim et al., 2012; Sen et al., 2006; Srikanta et al., 2009). La proximité de séquence entre deux copies joue un rôle dans ce type d'évènements. En effet, si la recombinaison peut survenir entre copies n'appartenant pas à la même sous-famille, elle est plus fréquente entre les AluY qu'entre les AluJ parce que les premiers sont plus conservés entre eux que les seconds, notamment à cause de l'accumulation de mutations aléatoires liées à l'âge des éléments (Sen et al., 2006). La recombinaison induite par les éléments Alu peut également provoquer, en plus des délétions, des duplications et des translocations, par exemple en alignant deux copies d'Alu non-alléliques durant la méiose (Kim et al., 2016).

1.2.3.3 Les éléments Alu et le nucléosome

Dans la partie précédente, on a vu que les éléments Alu sont fortement liés à l'évolution des génomes, à la fois par leurs effets sur l'expression des gènes mais également sur l'évolution des séquences, par leur insertion ainsi que par la recombinaison non-allélique induite par leur grand nombre de copies très similaires, qui peut mener à des réarrangements génomiques tels que larges délétions, duplications et translocations. Cependant, comme présenté dans la première partie de cette introduction, l'évolution de la séquence a une influence importante sur l'évolution de la structure du génome, et particulièrement sur sa brique de base que représente le nucléosome. Les éléments Alu étant des moteurs très importants de l'évolution des génomes de primates (Liu et al., 2009), il parait crucial de comprendre dans quelle mesure leur insertion peut avoir un impact sur le positionnement nucléosomal. C'est précisément un des enjeux de cette thèse, et les analyses et résultats obtenus seront présentés dans le **Chapitre 4**. Cette dernière partie introductive vise à faire un état des lieux de ce qui a été démontré en amont de cette thèse quant à la relation entre éléments Alu et nucléosomes.

Les premières études de l'influence de la séquence sur le positionnement nucléosomal ont mis en évidence une potentielle importance de la périodicité 10 pb de certains nucléotides et di-nucléotides, comme les AA/TT (Lowary & Widom, 1998). De telles périodicités faciliteraient l'enroulement de l'ADN autour des protéines histones pour former le nucléosome, dans des séquences identifiées comme "positionnantes". Cependant, comme on l'a vu dans la première partie de cette introduction, la contribution de ces périodicités à l'encodage du positionnement nucléosomal dans les séquences génomiques est très faible (Barbier et al., 2021). L'analyse spectrale de la distribution des nucléotides et di-nucléotides dans les génomes de primates confirme cette observation avec une absence de périodicité 10 pb (Tanaka et al., 2010). En revanche, cette analyse a mis en évidence que l'on retrouve des dinucléotides AA/TT et des nucléotides A/T et G/C périodiquement séparés par 84 ou 167 pb (Tanaka et al., 2010). Ces chiffres correspondent à la longueur d'un nucléosome additionné de son linker ($\sim 147+20$ pb). Cela a été interprété comme un marqueur de positionnement nucléosomal encodé dans la séquence génomique. Lorsque la même analyse a été effectuée en masquant les éléments transposables Alu, la périodicité observée a drastiquement diminué, indiquant que celle-ci est principalement liée à ces ETs (Tanaka et al., 2010). Cela a amené à l'analyse directe de données expérimentales de positionnement de nucléosomes le long des éléments Alu afin de déterminer dans quelle mesure le nucléosome est positionné par ces éléments. Cette analyse a montré que deux nucléosomes peuvent se former sur un élément, un sur chacun de ses deux bras riches en GC. Ce positionnement avait déjà été observé dans des expériences de reconstitution de nucléosomes *in vitro* sur des éléments Alu (Englander et al., 1993). On note que, positionnés ainsi, les nucléosomes ne se forment pas sur les séquences polyA internes ou terminales des éléments Alu, ce qui est en accord avec l'observation que ces séquences sont globablement inhibitrices de la formation du nucléosome (Barnes & Korber, 2021; Segal & Widom, 2009a). Dans cette étude, il a été remarqué que des nucléosomes étaient également positionnés en phase en amont et en aval de l'insertion. Cette dernière observation est à mettre en relation avec la proximité observée entre les éléments Alu et les barrières nucléosomales, aux bords desquelles on observe également un phasage des nucléosomes (Brunet et al., 2018; Drillon et al., 2016). Plus de la moitié de cette famille d'éléments transposables a été retrouvée positionnée directement aux bords d'une barrière nucléosomale, avec une claire préférence pour un positionnement du polyA terminal de l'élément face à la barrière nucléosomale (Brunet et al., 2018; Drillon et al., 2016).

Le positionnement des nucléosomes sur les éléments Alu pourrait être lié au contrôle de ces éléments par les génomes. En effet, des expériences *in vitro* ont montré que la formation des nucléosomes sur les éléments Alu diminuait la transcription de ces éléments par l'ARN polymérase III. De plus, la formation des nucléosomes associée à la méthylation des dinucléotides CpG sur les Alu semble complètement bloquer la transcription de ces éléments (Englander & Bruce, 1995; Englander et al., 1993). Le positionnement des nucléosomes sur les éléments Alu pourrait donc être le moyen utilisé par le génome pour contrôler la transcription de ces éléments. Il pourrait également être une simple conséquence du positionnement des insertions par rapport aux barrières nucléosomales qui sont à l'origine de nucléosomes périodiquement positionnés des deux côtés de chaque barrière (Brunet et al., 2018; Drillon et al., 2016). La séquence des éléments Alu semble également être partiellement inhibitrice de la formation des nucléosomes, au niveau des deux polyA des éléments. Or, les barrières nucléosomales sont également identifiées comme des séquences d'ADN inhibitrices de la formation des nucléosomes. Cela pose donc également la question des

interactions entre Alu et NIEBs, pour déterminer lequel des deux est antérieur à l'autre, et dans quelle mesure ces deux structures s'influencent mutuellement en terme de positionnement.

1.3 Objectifs de la thèse

Parmi les facteurs influençant le positionnement nucléosomal, la séquence de l'ADN enroulé autour des histones joue un rôle majeur (Segal & Widom, 2009b; Struhl & Segal, 2013). A l'échelle génomique, ce rôle se traduit principalement par la présence de séquences inhibitrices de la formation des nucléosomes, agissant comme des barrières nucléosomales. Aux bords de ces barrières, plusieurs nucléosomes se retrouvent positionnés, c'est-à-dire que leur localisation est très peu variable, contrainte par un effet de parcage aux bords des NIEBs (Drillon et al., 2016). La modification des séquences d'ADN peut donc amener à d'importants changements dans le positionnement nucléosomal (Field et al., 2008; Tsankov et al., 2011). Le nucléosome étant la brique de base de beaucoup de mécanismes de régulation du génome, de tels changements peuvent avoir une grande importance fonctionnelle. Ainsi, il apparaît nécessaire de tenir compte des conséquences sur l'évolution du positionnement nucléosomal lors de l'étude de l'évolution des séquences.

Les données expérimentales de positionnement nucléosomal ne sont malheureusement disponibles que dans trop peu d'espèces pour pouvoir étudier expérimentalement la co-évolution des séquences et de la structure de l'ADN. En revanche, il est possible, à l'aide de notre modèle physique de positionnement des nucléosomes, de prédire la position à la fois des barrières nucléosomales et celle des nucléosomes aux bords de ces barrières. Ce modèle ayant été validé dans plusieurs espèces (Chevereau et al., 2011; Drillon et al., 2016; Miele et al., 2008; Vaillant et al., 2010), il est un outil de choix pour étudier l'évolution conjointe de la séquence d'ADN et du positionnement nucléosomal, particulièrement en terme de barrières nucléosomales.

Un moteur majeur de l'évolution des séquences est l'insertion d'ETs (Biémont & Vieira, 2006). Ces séquences répétées, capable de se multiplier dans les génomes, peuvent avoir des conséquences importantes sur la valeur sélective des individus, de façon délétère ou bénéfique (Rebollo et al., 2012b). Cependant, la plupart des copies d'ETs sont, pour l'instant, décrites comme neutres d'un point de vue évolutif. L'étude des effets des ETs dans les génomes s'est pour l'instant focalisée sur ce qu'on pourrait résumer aux "effets de séquence", c'est-à-dire aux effets directement liés à l'apparition ou la disparition de telle ou telle séquence à tel ou tel locus, comme par exemple un site de fixation de facteur de transcription, un site d'épissage ou bien une délétion liée à de la recombinaison non-allélique. Jusqu'ici, les effets potentiels de l'évolution des séquences par l'insertion d'ETs sur l'évolution de la structure du génome, et particulièrement du positionnement nucléosomal, n'a été que très peu étudié.

L'objet de cette thèse est d'explorer cet aspect, à travers le prisme de la relation entre éléments transposables Alu et barrières nucléosomales. Les deux premiers chapitres de cette thèse seront consacrés principalement aux barrières nucléosomales, d'abord pour étudier leurs caractéristiques et leur conservation entre les différentes espèces, puis pour tenter de valider expérimentalement leur existence dans de nouvelles espèces. Cette dernière analyse amènera notamment à des indications quant à l'importance fonctionnelles des NIEBs dans les génomes. Enfin, la dernière partie de cette thèse est entièrement consacrée à la relation entre éléments transposables Alu et barrières nucléosomales, pour déterminer quelle peut être l'influence des NIEBs sur l'insertion des éléments Alu, et l'effet de ces insertions sur l'évolution des NIEBs.

2 Des barrières génomiques ubiquitaires pour la formation des nucléosomes

Sommaire

2.1 Introduction

Le mécanisme de positionnement des nucléosomes par effet de parcage contre des barrières inhibitrices a été mis en évidence d'abord chez la levure, puis chez l'humain (Drillon et al., 2016; Mavrich et al., 2008; Milani et al., 2009). Dans ces deux espèces, les prédictions du modèle physique de positionnement du nucléosome ont été confirmées expérimentalement par l'analyse de données de type MNase-seq. Chez l'humain, il a été notamment observé que le positionnement nucléosomal prédit aux bords des NIEBs corrèle avec les oscillations du profil de contenu en GC entre des valeurs hautes au niveau des séquences nucléosomales et basses au niveau des séquences inter-nucléosomales (Drillon et al., 2016). Une contrainte sur la position relative des NIEBs dans le génome a également été mise en évidence, comme détaillé dans l'introduction (**Partie 1.1.4**). En revanche, la densité en NIEBs selon les régions génomiques ne semble varier que très peu (Drillon et al., 2016).

La prédiction des NIEBs à partir du modèle physique de positionnement des nucléosomes a ensuite été effectuée dans d'autres génomes pour déterminer dans quelles espèces sont retrouvés les NIEBs, et également pour comparer leurs caractéristiques générales. Ainsi, la comparaison de la densité génomique en NIEBs, des distances entre deux barrières consécutives ainsi que du profil GC moyen aux bords des barrières a déjà été effectuée chez les vertébrés (Brunet et al., 2018). Elle a mis en évidence une conservation globale des caractéristiques. En effet, l'oscillation du % GC aux bords des NIEBs observée chez l'humain est retrouvée dans les autres espèces étudiées, tout comme la distribution quantifiée des distances entre deux barrières consécutives associée à des régions inter-NIEBs accommodant un nombre entier de nucléosomes (Brunet et al., 2018; Drillon et al., 2016). Cependant, quelques différences ont été notées, notamment concernant les inter-NIEBs de taille correspondant à deux nucléosomes, légèrement plus longs dans le génome humain que pour la plupart des autres vertébrés non-primates (272 bp vs. 259 bp). Ces inter-NIEBs de longueur spécifique ont été associés à la présence d'éléments Alu (Brunet et al., 2018).

Dans ce chapitre, ces comparaisons seront reprises et étendues à d'autres espèces non-vertébrés comme la drosophile ou l'arabette. Pour ceci, je disposais déjà des barrières nucléosomales dans les génomes de l'humain, du chimpanzé, de la souris, du porc et de la poule. A partir de notre modèle de positionnement de nucléosomes, j'ai pu prédire les barrières nucléosomales dans les espèces à ajouter à l'analyse, à savoir la drosophile, l'arabette et également le génome du rosier (disponible au Laboratoire de Reproduction et Développement des Plantes de l'ENS de Lyon). Il est à noter que j'ai également refait la prédiction sur le génome du poisson-zèbre afin d'en utiliser la dernière version. Le **Tableau 2.1** récapitule les versions, tailles et pourcentages GC des génomes analysés. Certaines caractéristiques associées aux NIEBs comme leur taille moyenne et la distance moyenne entre deux barrières consécutives sont disponibles dans le **Tableau 2.2**.

On peut séparer les caractéristiques des NIEBs en deux catégories :

— Les caractéristiques des NIEBs et de leur distribution génomique indépendamment de leur séquence telles que leur taille, la distance entre deux NIEBs consécutifs (qui donne un aperçu des corrélations spatiales de la distribution des NIEBs) ou encore la densité de NIEBs dans le génome.

— Les propriétés de séquence des NIEBs, telles que les profils aux bords des barrières des

Espèce	Nom scientifique	Version du génome	Taille du génome (Mb)	Taille du génome séquencé (Mb)	% GC moyen
Humain	*Homo sapiens*	hg38	3100	2938	40.9
Chimpanzé	*Pan troglodytes*	panTro5	3231	2871	40.9
Porc	*Sus scrofa*	susScr3	2809	2324	41.6
Souris	*Mus musculus*	mm10	2731	2648	41.7
Poule	*Gallus gallus*	galGal5	1230	1010	41.7
Poisson-zèbre	*Danio rerio*	danRer11	1373	1341	36.6
Drosophile	*Drosophila melanogaster*	dm6	144	137	42.1
Arabette	*Arabidopsis thaliana*	tair10	120	119	36.0
Rosier	*Rosa chinensis*	RcHm2	504	504	38.9
Levure	*Saccharomyces cerevisiae*	sacCer3	12	12	38.1

TABLEAU 2.1 – **Tableau récapitulatif des génomes utilisés lors des analyses du chapitre 1.** Les sources de ces données de séquence sont récapitulées dans la **Table A.1**.

contenus en GC et en polynucléotides, ou encore de l'énergie de formation des nucléosomes et de l'occupation nucléosomale prédits par le modèle.

Un analyse comparative de ces deux catégories de caractéristiques des NIEBs dans dix espèces eucaryotes sera présentée dans les **Parties 2.2** et **2.3**, respectivement.

En plus de ce travail comparatif, j'ai également étudié la conservation des NIEBs entre deux espèces proches, à savoir l'humain et le chimpanzé, pour déterminer dans quelle mesure la conservation des séquences était liée à une conservation du positionnement nucléosomal. Cette étude constitue la **Partie 2.4** de ce chapitre. La **Partie 2.5** est consacrée à l'étude des mutations ponctuelles aux bords des NIEBs pour les mutations non-CpG (**Partie 2.5.1**) et CpG (**Partie 2.5.2**). L'étude des patrons de mutations non-CpG chez le chimpanzé permettra de voir s'ils sont semblables aux patrons de mutations précédemment observés chez l'humain (Drillon et al., 2016) et détaillés en introduction. L'étude spécifique des mutations CpG aux bords des NIEBs, qui n'avait jusqu'ici jamais été réalisée, permettra de déterminer dans quelle mesure la distribution de ces mutations est en accord avec le positionnement nucléosomal prédit, particulièrement en prenant en compte l'hétérogénéité de ce type de mutations vis-à-vis des séquences nucléosomales et inter-nucléosomales.

2.2 Partage des caractéristiques des barrières aux nucléosomes entre 10 espèces eucaryotes

2.2.1 La densité de barrières varie, indépendamment de la phylogénie

Une des caractéristiques importantes pour comparer les barrières nucléosomales entre les génomes est leur densité génomique. On peut la mesurer directement en nombre de barrières par kilobase, en divisant le nombre de barrières total du génome par la taille du génome en kilobases. Elle donne une information sur la quantité de barrières dans chaque génome, avec un chiffre

Espèce	Nombre de NIEBs	Taille moyenne / médiane des NIEBs (pb)	Nombre d'inter-NIEBs	Taille moyenne / médiane des inter-NIEBs (pb)	Inter-NIEBs >= 1000 pb et NIEBs >= 70 pb
Humain	1 745 801	135 / 113	1 745 542	1436 / 997	585 339 (34%)
Chimpanzé	1 733 364	137 / 115	1 727 109	1404 / 1000	587 468 (34%)
Porc	1 494 058	141 / 118	1 485 284	1315 / 970	489 790 (33%)
Souris	1 465 549	119 / 103	1 465 383	1515 / 1063	481 918 (33%)
Poule	426 500	116 / 103	426 260	1919 / 1384	173 047 (41 %)
Poisson-zèbre	999 476	159 / 134	995 451	1171 / 870	342 839 (34%)
Drosophile	102 282	136 / 116	81 433	1207 / 919	57 966 (57%)
Arabette	61 353	168 / 140	61 328	1769 / 1234	28 691 (47%)
Rosier	293 783	133 / 115	293 776	1582 / 1070	110 878 (38%)
Levure	4735	120 / 108	4719	2399 / 1782	2382 (50%)

TABLEAU 2.2 – **Nombre et taille moyenne des NIEBs et des inter-NIEBs dans les espèces du Tableau 2.1**. La dernière colonne indique également le nombre d'inter-NIEBs de tailles >= 1000 pb, et dont les NIEBs gauche et droit sont de tailles supérieures à 70 pb.

directement comparable entre les espèces car la taille des génomes est prise en compte dans le calcul, ce qui permet de s'affranchir de la grande différence de tailles entre les génomes analysés (**Tableau 2.1**). Ainsi, une densité plus importante dans un génome par rapport à un autre indique donc simplement que, proportionnellement à la taille des deux génomes, il y a plus de barrières dans l'un que dans l'autre. La comparaison des densités de barrières dans les différents génomes nous indiquera donc si certaines espèces ont un génome particulièrement enrichi ou appauvri en barrières nucléosomales. La **Figure 2.1** résume les densités en barrières dans les 10 génomes étudiés.

On observe que les densités génomiques peuvent être assez variables entre espèces. En effet, les valeurs vont de 0.39 NIEBs/kb chez la levure à 0.75 NIEBs/kb chez le poisson-zèbre, on a donc quasiment un facteur 2 de différence entre ces deux espèces. Cependant, les quatre mammifères, espèces proches phylogénétiquement présentent des valeurs de densités similaires, respectivement de 0.64, 0.64, 0.68 et 0.61 NIEBs/kb chez l'humain, le chimpanzé, le porc et la souris. On n'observe pas de corrélation entre la taille des génomes analysés et la densité en NIEBs. En effet, si l'on classe par exemple la drosophile, la poule, l'arabette et le poisson-zèbre selon leur nombre de NIEBs par kb, on obtient le classement suivant : poisson-zèbre (0.75) > drosophile (0.74) > arabette (0.52) > poule (0.49). Or, ce classement diffère largement de celui qu'on obtient avec la taille des génomes de ces organismes, à savoir : zebrafish (~ 1.4 Gb) > poule (~ 1.2 Gb) > drosophile (~ 142 Mb) > arabette (~ 119 Mb). Enfin, on observe une densité de barrières bien inférieure chez la levure (0.39 NIEBs/kb) par rapport aux autres organismes. Cette dernière observation pourrait être expliquée par la fonction des barrières nucléosomales chez la levure. En effet, dans cette espèce, les barrières ont été décrites principalement aux promoteurs des gènes, où elles facilitent la mise en place de la transcription en rendant ces séquences régulatrices plus accessibles que si elles étaient occupées par un nucléosome (Tsankov et al., 2010). En revanche, chez l'humain, l'inverse est observé, avec une absence de barrières nucléosomales aux promoteurs, voire même la présence de séquences favorisant la formation du nucléosome (Drillon et al., 2016; Tompitak

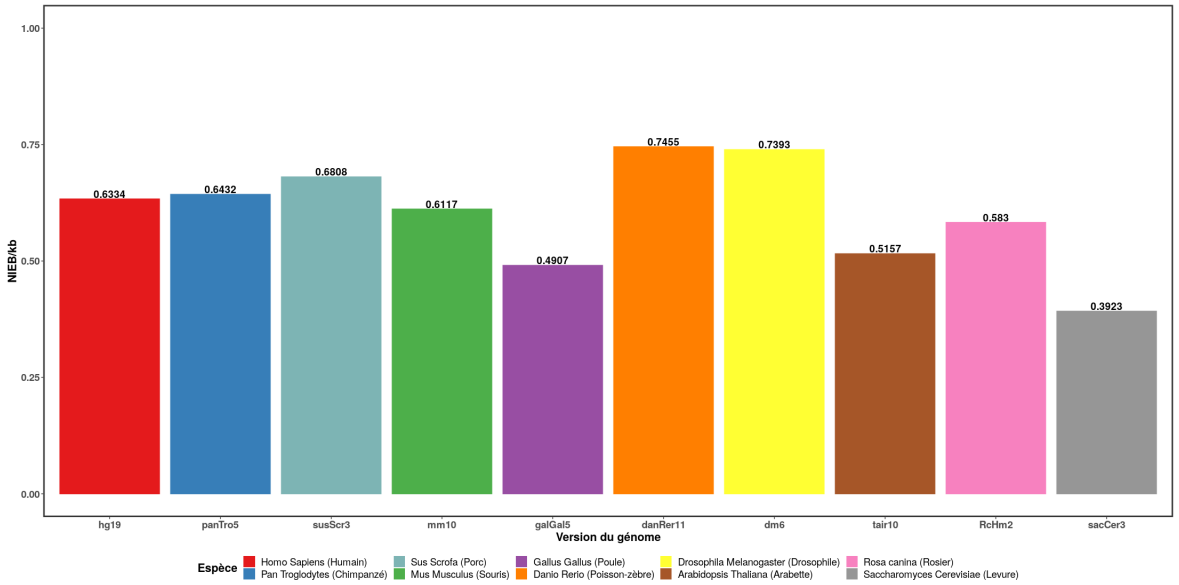


FIGURE 2.1 – **Densité en NIEBs dans les 10 génomes du Tableau 2.1**. Les couleurs associées aux espèces sont indiquées dans la légende. La densité en NIEBs de chaque génome est précisée au dessus de chaque barre (arrondie au dix-millième). La ligne horizontale grise correspond à la valeur obtenue chez l'humain.

et al., 2017). En fait, la présence/absence de barrière nucléosomale au promoteur a même été détectée comme corrélée à la complexité des organismes (définie par une estimation du nombre de types cellulaires différents). Plus un organisme est complexe, plus les promoteurs de ses gènes sont favorables à la formation d'un nucléosome (Tompitak et al., 2017). Concernant les espèces étudiées ici, seule la levure apparait comme un organisme "peu complexe", pour lequel les barrières nucléosomales sont particulièrement présentes aux promoteurs des gènes. Cela pourrait donc expliquer la densité plus faible qu'on observe dans cette espèce par rapport aux neuf autres. Pour finir, il est à noter que la densité de barrières semble, quelle que soit l'espèce, plutôt homogène sur le génome. En effet, les densités par chromosome dans chaque espèce restent globalement proche de la moyenne génomique de l'espèce (**Annexe A.1**). De plus, chez l'humain, l'étude de la densité en NIEBs dans des fenêtres de 100 kb a montré que cette densité est assez homogène tout le long du génome, variant entre 0.54 et 0.65 NIEBs/kb, et ce quel que soit le contexte génomique (régions riches ou pauvres en GC, génique ou intergénique) ou épigénétique (réplication précoce ou tardive, sensibilité à la DNase I) (Drillon et al., 2016). Dans cette espèce, les barrières semblent donc ubiquitaires, excepté aux promoteurs des gènes où l'on retrouve significativement moins de barrières que dans les autres régions (Drillon et al., 2016).

2.2.2 La taille des barrières est très conservée entre les espèces

Une autre caractéristique des barrières nucléosomales est leur taille. En effet, la taille de la barrière définit la longueur de la zone où la formation du nucléosome est inhibée, et donc où l'ADN est plus accessible que s'il est enroulé autour des histones. Il semble intéressant de comparer la distribution de la taille des barrières dans les différentes espèces pour voir si des spécificités apparaissent comme on a pu l'observer précédemment avec les densités génomiques en barrières. Pour ça, j'ai établi la distribution de la taille des barrières dans les dix génomes analysés. J'ai simplement effectué un comptage du nombre de barrières de chaque taille, et normalisé les comptages en les

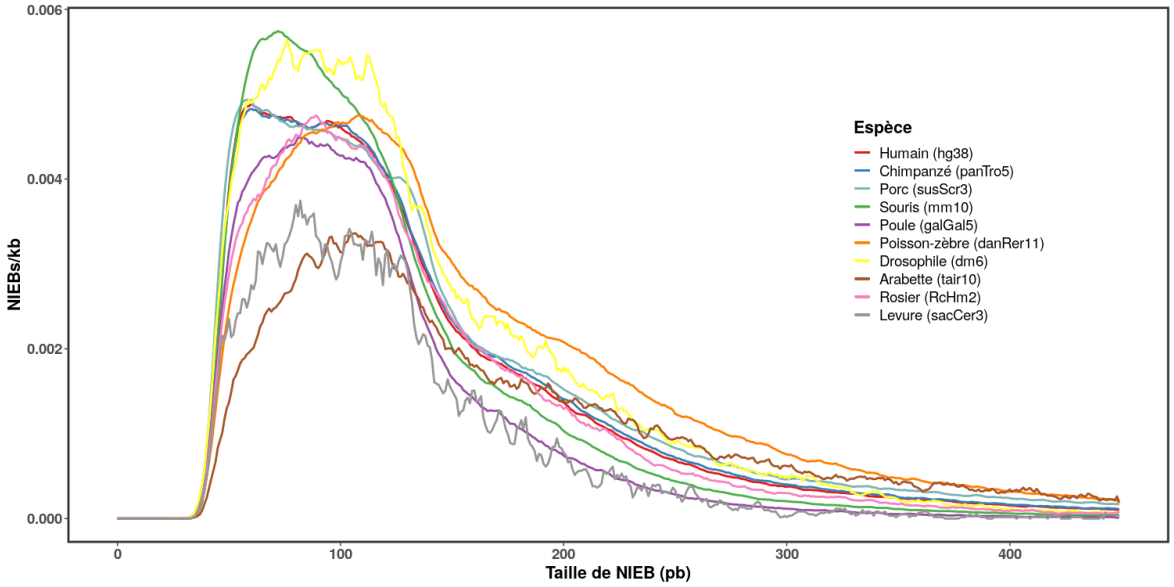


FIGURE 2.2 – **Distribution de la taille des NIEBs dans les 10 espèces analysées.** Les couleurs correspondent à celles utilisées dans la **Figure 2.1**. Les courbes présentées sont des moyennes glissantes sur 5 pb.

divisant par la taille du génome (en kb), afin de s'affranchir des tailles de génomes très différentes. L'ordonnée de la **Figure 2.2** correspond donc au nombre de barrières par kilobase. Enfin, il est à noter que chez l'humain il a été montré que les barrières de tailles supérieures à 450 pb ne sont que partiellement confirmées par les données expérimentales de positionnement du nucléosome (Drillon et al., 2016), ces barrières sont donc systématiquement exclues des analyses.

On voit que la distribution de la taille des barrières est très similaire entre les différents génomes (**Figure 2.2**). Elle est centrée autour de 100 pb, avec une courbe en cloche autour de cette abscisse, suivie d'une queue assez longue à droite indiquant qu'un certain nombre de barrières ont une taille supérieure à 200 pb. Très peu de barrières présentent une taille inférieure à 50 pb, et ce quelle que soit l'espèce étudiée. On note également que deux espèces très proches comme l'humain et le chimpanzé présentent une distribution quasiment identique. Enfin, on observe que chez l'arabette et la levure, la densité de barrières est nettement inférieure aux autres sur la zone 50-150 pb. Cependant, ces deux génomes présentent également des densités en NIEBs inférieures aux autres (**Figure 2.1**). La normalisation effectuée ici n'est pas indépendante de la densité en NIEBs au sein de l'espèce. En effet, on a ici un nombre de NIEBs par kb pour chaque taille de barrière, ainsi le nombre de NIEBs par kb global dans le génome a un effet direct sur cette mesure. En d'autres termes, si une espèce a une densité génomique en NIEBs faible, alors la courbe de sa distribution de taille sera mécaniquement en dessous de celle d'une espèce ayant les mêmes proportions de NIEBs de chaque taille mais une densité génomique en NIEBs plus importante. Lors des comparaisons entre deux espèces, il faut donc tenir compte de la densité globale en NIEBs au sein de chacune de ces espèces représentées dans la **Figure 2.1**. Cette influence de la densité génomique en NIEBs pourrait expliquer que les courbes de l'arabette et particulièrement de la levure sont plus basses que les autres. En effet, la levure présente également la densité génomique en NIEBs la plus faible du groupe d'espèces (0.39 NIEBs/kb). Il est donc logique de voir sa distribution de taille se positionner en dessous des espèces à densité génomique en NIEBs plus fortes. Concernant l'arabette, on note que sur la zone 50-150 pb, la courbe est nettement en dessous de celle de la poule, qui présente

pourtant une densité génomique en NIEBs similaire. Sur la zone 150-450 pb, on observe l'inverse. Les NIEBs de l'arabette semblent globalement de plus grandes tailles que ceux de la poule. La forme globale de la distribution de taille des NIEBs est néanmoins la même chez ces deux espèces. De manière générale, chez les 10 espèces analysées ici, on observe un enrichissement des NIEBs de tailles allant de 50 à 150 pb par rapport aux tailles plus grandes, et la majorité des barrières nucléosomales sont de tailles inférieures à 250 pb. Enfin, les tailles moyennes et médianes des NIEBs sont maximales pour l'arabette (168 pb et 140 pb, resp.) et le poisson-zèbre (159 pb et 134 pb, resp.) et minimales pour la souris (119 pb et 103 pb, resp.), la poule (116 pb et 103 pb, resp.) et la levure (120 pb et 108 pb, resp.) (**Table 2.2**).

2.2.3 La distribution des distances inter-barrières indique une contrainte sur leur positionnement, conservée entre espèces

La dernière caractéristique générale des NIEBs étudiée ici concerne la distribution spatiale relative des barrières. En d'autres termes, comment les barrières nucléosomales sont-elles positionnées les unes par rapport aux autres ? Existe-t-il des contraintes sur leur positionnement ? Et si oui, ces contraintes sont-elles les mêmes dans tous les génomes ? Pour répondre à ces questions, j'ai établi la distribution de la taille des inter-NIEBs dans les 10 génomes étudiés jusqu'ici. Un inter-NIEBs est identifiée ici par la zone séparant deux barrières consécutives. Un inter-NIEBs ne peut donc, par construction, pas contenir de barrière nucléosomale. En principe, pour chaque couple de barrières consécutives, on devrait avoir un inter-NIEBs associé. Ainsi, le nombre d'inter-NIEBs sur un chromosome devrait correspondre au nombre de barrières moins un. Dans la réalité, il y a un petit peu moins d'inter-NIEBs que ce chiffre, car sont retirés de l'analyse tous ceux qui contiennent des bases non-identifiées lors du séquençage des génomes (les bases N), pour se concentrer sur les zones entièrement séquencées, afin notamment de retirer de l'analyse des zones dont les séquences sont méconnues comme les régions centromériques. Ainsi, selon la qualité des génomes, le nombre d'inter-NIEBs peut légèrement différer du nombre attendu sachant le nombre de NIEBs. Cependant, pour nos génomes d'intérêt, ces différences sont faibles, comme on peut le voir dans le **Tableau 2.2**.

Les distributions de distances inter-NIEBs ont été réalisées de la même manière que les distributions de taille des NIEBs (**Partie 2.2.2**). Il faut donc ici encore tenir compte de la densité en NIEBs globale de chaque génome dans la comparaison des courbes les unes par rapport aux autres. S'il n'existe pas de contrainte sur le positionnement des barrières les unes par rapport aux autres, alors leur positionnement devrait être aléatoire dans les génomes. Dans le cas d'un positionnement aléatoire d'objet, la distribution de la distance entre ces objets devrait suivre une loi exponentielle, de paramètre 1/(Moyenne des distances entre objets). Ainsi, par exemple chez l'humain, on a une distance moyenne entre les barrière de 1436 pb (**Tableau 2.2**), donc si les barrières humaines sont distribuées aléatoirement, on s'attend à ce que la distribution des distances inter-barrières humaines suive une distribution exponentielle de paramètre 1/1436, soit 0.00070. Les courbes exponentielles attendues pour chaque espèce sont disponibles en **Annexe A.2**.

On voit sur la **Figure 2.3** que les distributions des distances inter-NIEBs sont en deux parties. La partie droite correspond à la queue d'une distribution exponentielle (**Annexe A.2**), ce qui indique que lorsque les barrières sont distantes de plus d'un kilobase, il ne semble pas y avoir de contrainte sur leur positionnement les unes par rapport aux autres. En revanche, lorsque la distance

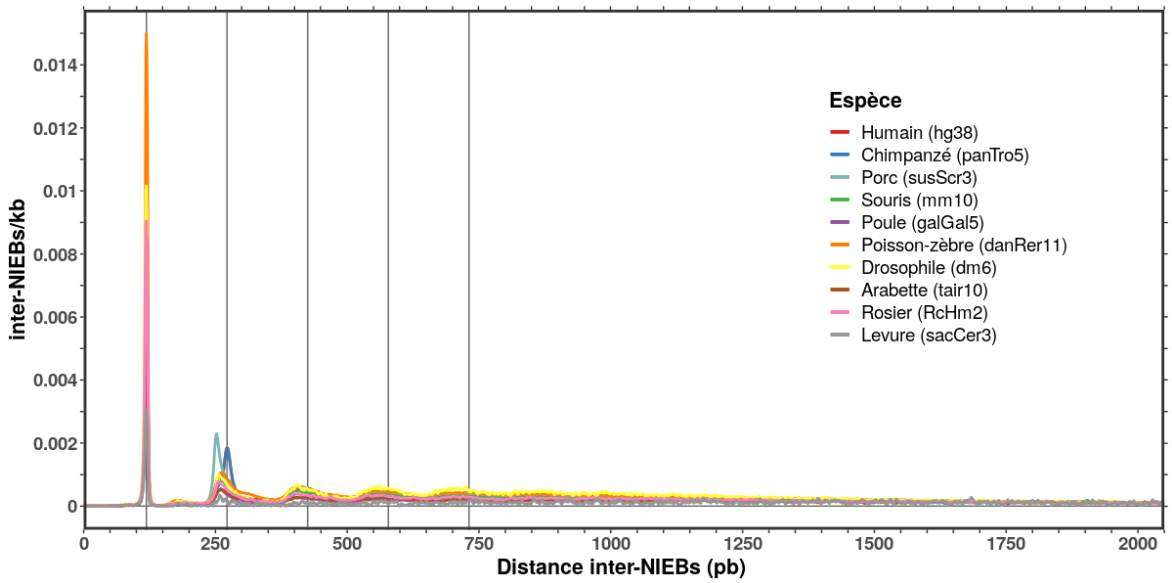


FIGURE 2.3 – **Distribution de la taille des inter-NIEBs dans les 10 espèces analysées** Les couleurs correspondent à celles utilisées dans la **Figure 2.1**. Les lignes verticales grises correspondent aux maxima de la courbe obtenue chez l'humain. Les abscisses sont 119, 272, 425, 578 et 731 pb. La courbe humaine (en rouge) n'apparaît pas ou peu car elle est parfaitement superposée avec celle du chimpanzé (en bleu).

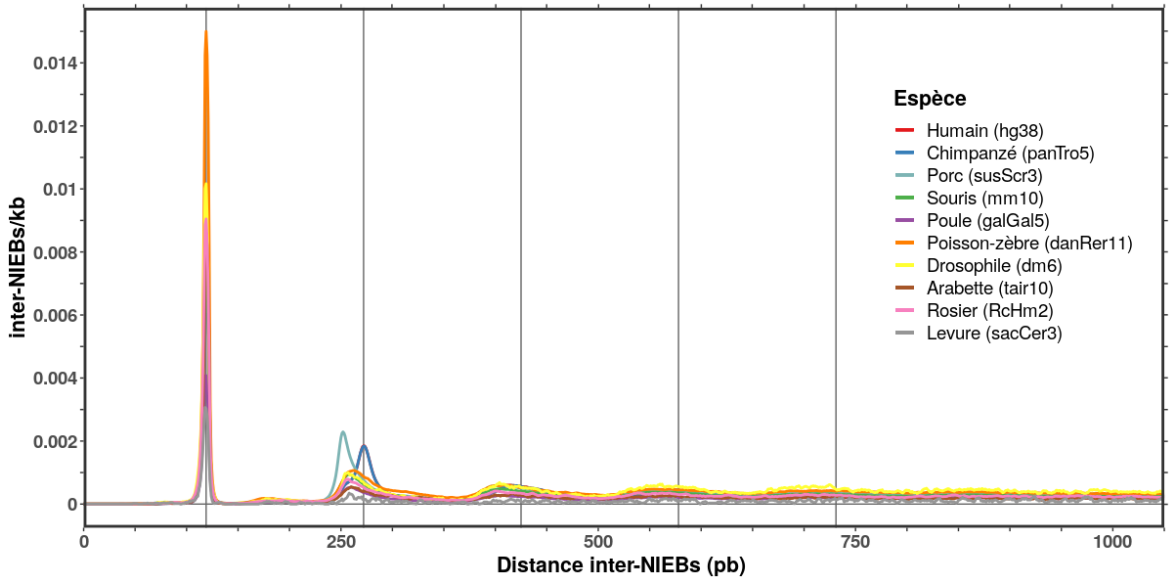


FIGURE 2.4 – **Distribution de la taille des inter-NIEBs pour les tailles 0 à 1000 pb**. Les lignes verticales grises correspondent aux maxima de la courbe obtenue chez l'humain. Les abscisses sont 119, 272, 425, 578 et 731 pb. La courbe humaine (en rouge) n'apparaît pas ou peu car elle est parfaitement superposée avec celle du chimpanzé (en bleu).

inter-NIEBs est inférieure à 1 kb, la distribution change complètement de forme. On observe des pics, régulièrement espacés d'environ 153 pb, une distance très proche des 147 pb constituant un nucléosome. La forme de la courbe, que ce soit pour les distances inférieures ou supérieures à 1 kb, est globalement conservée dans les différentes espèces. Il semble donc qu'il y ait une contrainte commune à toutes les espèces sur le positionnement des barrières lorsque celles-ci sont proches les unes des autres, mais pas lorsque les barrières sont plus éloignées. Pour mieux comprendre cette contrainte, on va se focaliser sur la zone 0-1000 pb (**Figure 2.4**).

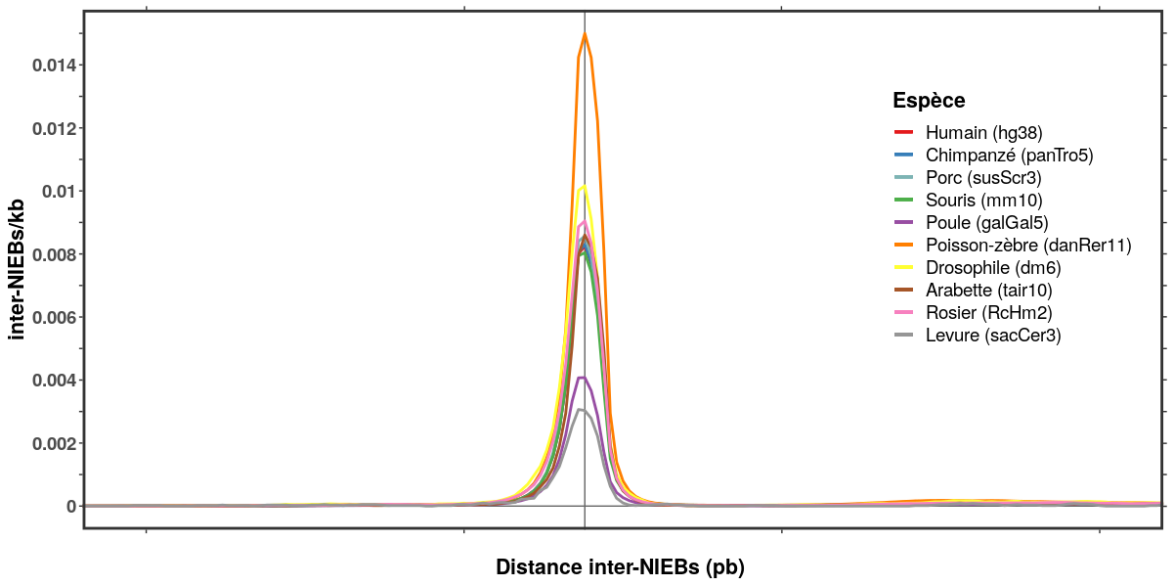


FIGURE 2.5 – **Distribution de la taille des inter-NIEBs pour les tailles 50 à 200 pb**. La ligne verticale grise correspond au maximum de la courbe obtenue chez l'humain, à l'abscisse 119. La courbe humaine (en rouge) n'apparaît pas ou peu car elle est parfaitement superposée avec celle du chimpanzé (en bleu).

Pour les inter-NIEBs de tailles inférieures à 1000 pb, la distribution de taille forme des pics, régulièrement espacés d'environ 153 pb. Lorsque les barrières sont proches les unes des autres, elles sont majoritairement distantes d'environ 119, 272, 424 ou encore 578 pb, avec les distances intermédiaires sous-représentées. Cette forme de courbe, très éloignée d'une forme exponentielle, traduit une contrainte sur le positionnement des barrières les unes par rapport aux autres. On remarque que les pics sont espacés d'environ 153 pb, ce qui est compatible avec la formation d'un nucléosome. Lorsque les barrières sont proches les unes des autres, celles-ci sont placées de manière à ce qu'un nombre entier de nucléosomes puissent se former entre elles, ni plus ni moins. Le premier pic à 119 correspondrait alors à un nucléosome entre deux barrières, celui à 272 à deux nucléosomes entre deux barrières, celui à 425 à trois nucléosomes, et ainsi de suite. Cette hypothèse est validée par l'analyse de données expérimentales de positionnement de nucléosomes *in vivo* chez l'humain, réalisée par Drillon et al. (Drillon et al., 2016). En effet, sur la figure 3 de cet article, on peut voir la heatmap de la densité en nucléosomes entre les barrières nucléosomales en fonction de la taille (Panneaux A et B pour les données *in vivo*, C pour les données *in vitro*). On voit que pour les inter-barrières correspondant à la taille "1 nucléosome", on a une augmentation de la densité au centre de l'inter-barrières, traduisant qu'un nucléosome est bien présent, positionné au centre de l'inter-NIEBs. Les inter-NIEBs de taille "2 nucléosomes" contiennent également bien 2 nucléosomes, ceux de taille "3 nucléosomes" en contiennent bien 3, etc. Ce résultat vient donc confirmer l'hypothèse d'une contrainte imposant un nombre entier de nucléosomes entre les barrières lorsque celles-ci sont proches les unes des autres. On note que la distance entre les pics de 153 pb correspond à un chapelet nucléosomal très compact avec des linkers de 6 pb pour des nucléosomes canoniques de 147 pb. Pour finir, cette contrainte semble partagée par tous les génomes analysés ici. Cependant, on observe quelques différences entre les espèces, principalement au niveau des deux premiers pics. Deux zooms sur chacun de ces pics sont présents en **Figure 2.5** et **Figure 2.6**, afin d'en détailler les spécificités.

Pour les inter-NIEBs de taille correspondant à un nucléosome (**Figure 2.5**), on observe une forte

conservation de la distance inter-NIEBs, avec toutes les espèces présentant le même pic à 119 pb. Les différences entre espèces concernent l'amplitude du pic, à savoir la densité en NIEBs de cette taille. On constate un facteur ~ 3 dans la densité de ces objets entre les génomes. Si l'on tient compte de la densité globale en barrières dans chaque génome lors des comparaisons, il est cohérent de voir la courbe de la drosophile (en jaune) monter bien plus haut que celle de la levure (en gris), car la densité en NIEBs de la drosophile est bien plus élevée que celle de la levure (**Figure 2.1**). On note néanmoins plusieurs différences qui ne sont pas expliquées par la variabilité de densité génomique en barrières. Par exemple, le premier pic est bien plus haut chez le poisson-zèbre que chez la drosophile, malgré des densités globales similaires (**Figure 2.1**). Il y a donc plus d'inter-NIEBs de taille "1 nucléosome" chez le poisson-zèbre que chez la drosophile. Si l'on compare aux autres génomes étudiés, on voit que la courbe de la drosophile est légèrement supérieure à celle des autres espèces, en accord avec les densités génomiques en NIEBs de chaque espèces. Quant à la courbe du poisson-zèbre, elle est largement supérieure aux autres courbes, beaucoup plus que ce à quoi on pourrait s'attendre selon les densités génomiques en NIEBs respectives (**Figure 2.1**). Ainsi, parmi les inter-NIEBs du poisson-zèbre, une proportion importante semble être de taille "1 nucléosome", plus importante que la proportion correspondante dans les autres génomes. Chez le rosier (courbe rose), on observe une courbe se plaçant légèrement au dessus de celle des mammifères (courbes rouge, bleue, verte clair et verte), alors que la densité génomique en NIEBs du rosier est légèrement inférieure à celle des mammifères (0.58/kb contre 0.61/kb à 0.68/kb). Comme celui du poisson-zèbre, le génome du rosier présente donc une proportion d'inter-NIEBs de taille "1 nucléosome" plus importante que les autres. Pour finir, on observe que les courbes des 4 mammifères sont très similaires. Étant donnée leur proximité en terme de densité génomique en NIEBs, on peut en déduire que la proportion d'inter-NIEBs de taille "1 nucléosome" dans ces génomes est proche. Pour les inter-NIEBs de taille "1 nucléosome", si on observe quelques spécificités en terme de proportion dans les génomes, la taille des inter-NIEBs semble très conservée. Des spécificités plus intéressantes sont observées pour les inter-NIEBs de tailles 2 et 3 nucléosomes (**Figure 2.6**).

Sur la **Figure 2.6**, on peut observer plus précisément le second et le troisième pic de la **Figure 2.3**. Tout d'abord, on remarque que la courbe de la levure présente un nombre d'inter-NIEBs/kb très proche de 0 pour ces tailles. Comme évoqué précédemment pour la densité génomique en NIEBs dans cette espèce, la fonction particulière des NIEBs dans cette espèce (à savoir une présence aux promoteurs des gènes pour augmenter leur accessibilité à la machinerie de transcription) par rapport aux autres étudiées ici pourrait être en cause. La courbe humaine (en rouge) et la courbe chimpanzé (en bleu) sont identiques, et forment deux pics, aux positions 272 et 425, avec les tailles intermédiaires largement sous-représentées. Les courbes des autres espèces sont décalées vers la gauche par rapport aux courbes de l'humain et du chimpanzé. Les courbes de la souris, la poule, le poisson-zèbre, la drosophile, l'arabette et la rose présentent des maxima autour de 260 pb et 410 pb, soit environ 15 pb en deçà des courbes de l'humain et du chimpanzé. Pour les inter-NIEBs de taille correspondant à deux nucléosomes (le premier pic de la **Figure 2.6**), on observe cependant que les courbes de l'humain et du chimpanzé présentent un épaulement au niveau du maximum des courbes des autres espèces (~ 260 pb). Il semble donc qu'il y ait deux types d'inter-NIEBs de taille 2 nucléosomes chez l'humain et le chimpanzé, un premier type pour lequel la distance inter-NIEBs est identique à celle des inter-NIEBs correspondants chez les autres espèces (sauf le porc), et un second type pour lequel la distance inter-NIEBs est spécifique de ces deux espèces. Le second

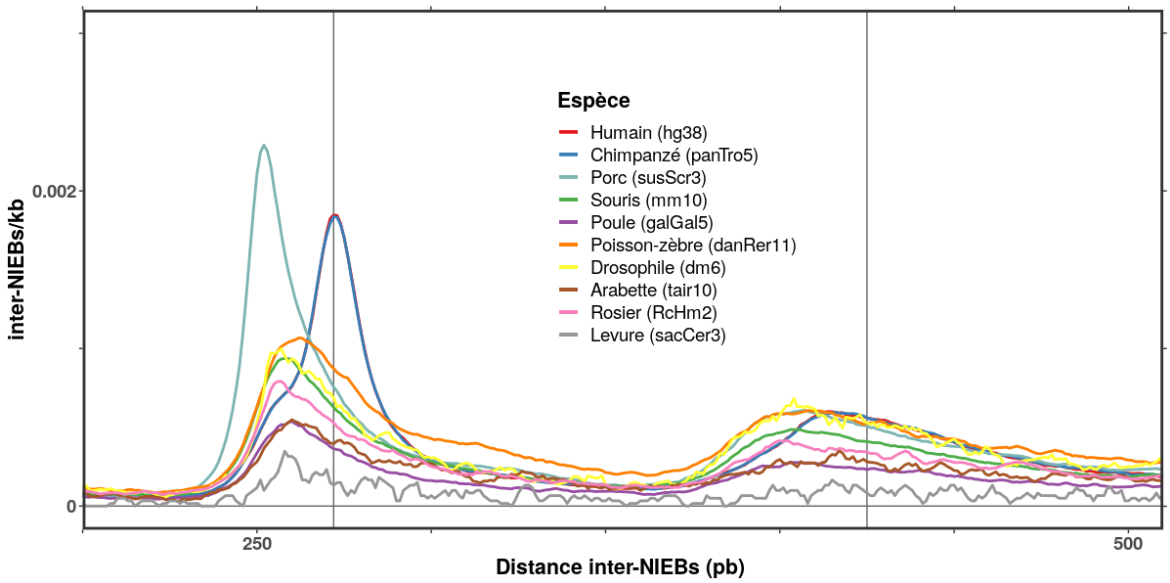


FIGURE 2.6 – **Distribution de la taille des inter-NIEBs pour les tailles 200 à 500 pb.** Les lignes verticales grises correspondent aux maxima de la courbe obtenue chez l'humain, soit aux abscisses 272 et 425. La courbe humaine (en rouge) n'apparaît pas ou peu car elle est parfaitement superposée avec celle du chimpanzé (en bleu).

type semble majoritaire dans ces génomes, avec un pic d'amplitude 0.002 inter-NIEBs/kb pour la distance 272 pb contre 0.00075 pour la distance 260 pb. Les spécificités de distances inter-NIEBs chez l'humain ont été mises en relation avec d'autres éléments de ce génome, particulièrement les éléments transposables. Cela a mis en évidence que les distances inter-NIEBs spécifiques de cette espèce sont associées à la présence d'éléments transposables Alu (Brunet et al., 2018). En effet, en l'absence de ces éléments, la distribution des distances inter-barrières chez l'humain est similaire à celle observée pour les autres espèces présentées ici (hors chimpanzé). Les éléments Alu étant spécifiques des primates, on suppose la même relation entre présence d'éléments Alu et distance inter-NIEBs chez le chimpanzé. La présence d'éléments Alu dans les inter-NIEBs de longueur 2 ou 3 nucléosomes semble donc associée à un léger allongement de ceux-ci. La relation entre NIEBs et éléments Alu a été longuement étudiée durant cette thèse, et les résultats obtenus sont présentés en détail dans le **Chapitre 4** de ce manuscrit.

Pour terminer, on observe que la distribution de tailles d'inter-NIEBs chez le porc montre également des spécificités. En effet, si le pic correspondant aux inter-NIEBs de taille 3 nucléosomes semble similaire à celui observé pour la majorité des espèces (et l'humain/le chimpanzé en l'absence d'Alu), celui correspondant aux inter-NIEBs de taille 2 nucléosomes est décalé vers la gauche, à 250 pb. Les inter-NIEBs du porc de taille 2 nucléosomes semblent donc plus petits que les inter-NIEBs correspondant dans les autres espèces (et a fortiori bien plus petits que ceux contenant des Alu chez l'humain et le chimpanzé). Il serait intéressant d'essayer d'associer ces inter-NIEBs à d'autres éléments du génome du porc, pour voir si l'on peut retrouver une relation similaire à celle entre NIEBs et Alu chez l'humain et le chimpanzé.

Les barrières nucléosomales semblent donc être des structures communes aux eucaryotes, dont les caractéristiques principales sont partagées entre les espèces. La densité en NIEBs dans les génomes peut varier, indépendamment de la phylogénie, même si les espèces proches phylogénétiquement semblent contenir des densités génomiques en NIEBs similaires. La taille des NIEBs

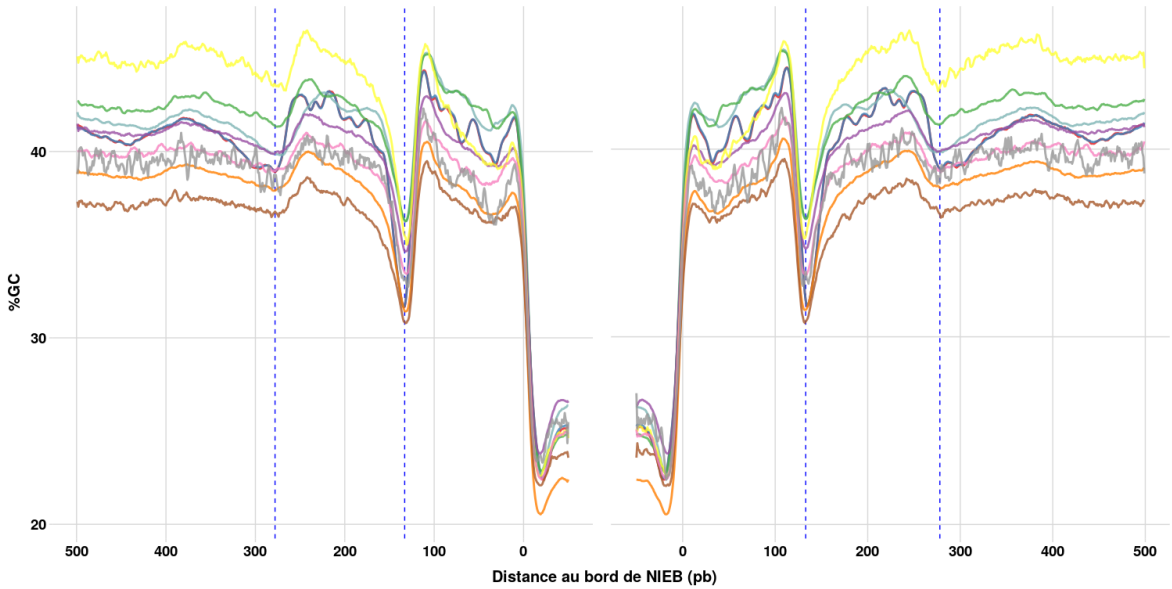
semble, elle, très conservée entre les espèces, avec une majorité de NIEBs de tailles 50 à 150 pb. Enfin, la distribution des distances entre deux barrières nucléosomales successives illustre une même contrainte de nombres entiers de nucléosomes entre deux NIEBs dans toutes les espèces analysées ici, ce qui suggère un encodage universel du positionnement nucléosomal dans les séquences génomiques.

2.3 Une écriture génomique ubiquitaire des barrières aux nucléosomes

2.3.1 Une oscillation conservée du profil G+C aux bords des barrières aux nucléosomes

Les résultats présentés en **Partie 2.2** suggèrent que l'écriture des nucléosomes dans la séquence génomique est universelle. Pour étudier cette hypothèse, on peut comparer le profil moyen de composition en GC aux bords des barrières. En effet, la composition en GC est une bonne approximation du positionnement nucléosomal (**Chapitre 1**). J'ai calculé les profils moyens de composition en GC le long de chaque barrière nucléosomale étendue de 500 pb en amont et en aval dans les 10 espèces étudiées ici. Pour cela, j'ai d'abord restreint le jeu de données aux NIEBs de tailles supérieures à 70 pb, et bordant un inter-NIEBs de taille supérieure à 1000 pb. En effectuant ce filtrage sur la taille d'inter-NIEBs, je m'assure que les observations sur les 500 pb au bord d'une barrière ne sont pas biaisées par la présence d'une autre barrière trop proche. Pour un inter-NIEBs de taille 1000 pb au minimum, les positions 1 à 499 dans l'inter-NIEBs sont plus proches de la barrière de gauche que de celle de droite. De la même manière, imposer une taille minimale de barrières de 70 pb permet d'avoir des résultats également aux bords internes des barrières, jusqu'à 35 pb à l'intérieur. Ces deux filtrages conservent plus d'un tiers des données dans chaque génome (**Tableau 2.2**). Après avoir filtré les NIEBs, j'ai, pour chaque NIEB, récupéré les séquences composées des 35 pb au bord interne de la barrière et des 500 pb au bord externe, et compté le nombre de nucléotides de chaque type (A, T, C ou G) à chaque distance du bord de la barrière. Le profil GC est alors calculé comme le nombre de nucléotides C ou G divisé par le nombre total de sites, et ce pour chaque distance entre −35 pb et 500 pb. Cette manipulation a été effectuée à la fois pour les barrières de gauche (pour calculer le profil moyen en aval des barrières) et celles de droite (pour calculer le profil moyen en amont) (**Figure 2.7**).

On observe une très forte conservation du profil GC aux bords des barrières. En effet, dans les 10 espèces étudiées, on voit une très nette oscillation du taux de GC, quel que soit le taux de GC moyen génomique de l'espèce. Aux bords internes des barrières, les valeurs sont particulièrement basses, aux alentours de 25 %. On observe ensuite un taux beaucoup plus haut (de 35 % à 45 %) entre les positions 0 pb et 130 pb, suivi d'une dizaine de paires de bases à faible taux de GC (de 30 % à 35%, avec un minimum en position 133 pb), puis à nouveau environ 140 pb à fort taux de GC, une seconde chute du taux sur quelques paires de bases (minimum à 278 pb), et à nouveau une remontée sur les 150 pb suivantes. Il est à noter que la forme de ce profil est remarquablement conservée entre les espèces, les différences de valeurs étant ici principalement expliquées par les différences de taux de GC moyen dans les différents génomes (**Tableau 2.1**). On observe cependant quelques différences dans les oscillations, particulièrement chez l'humain au niveau des deux première zones à fort taux de GC. En effet, les profils GC de l'humain et du chimpanzé présentent



Les profils présentés sont des moyennes glissantes sur 5 pb.

FIGURE 2.7 – **Profils GC moyens aux bords des NIEBs dans les 10 espèces analysées.** Les couleurs correspondent à celles utilisées dans la **Figure 2.1**. Le nombre d'inter-NIEBs utilisés pour chaque espèce est indiqué dans le **Tableau 2.2**. Les lignes verticales pointillées bleues correspondent aux minima du taux de GC chez l'humain, aux abscisses 133 pb et 278 pb.

des variations beaucoup plus marquées que les autres (exceptée pour la levure, pour laquelle la faible quantité de données explique les fluctuations observées, **Tableau 2.2**). Ces variations ont déjà été remarquées précédemment, et sont dues à la présence d'éléments transposables de type Alu (Brunet et al., 2018). En effet, en retirant les éléments Alu de l'analyse chez l'humain et le chimpanzé, les variations observées spécifiquement dans ces deux espèces disparaissent et le profil obtenu reproduit celui retrouvé dans les autres espèces (Brunet et al., 2018; Drillon et al., 2016). Pour finir, il a été observé qu'en plus d'un taux de GC très faible, les bords internes des NIEBs sont également enrichis en séquences de type polyA et polyT chez les vertébrés (Brunet et al., 2018). Chez l'humain, cet enrichissement est d'ailleurs asymétrique entre les deux types de séquences selon le bord du NIEB étudié (droit ou gauche), en raison de la présence d'éléments Alu sens d'un côté et antisens de l'autre (Brunet et al., 2018). La distribution des éléments Alu aux bords des NIEBs, qui explique ce résultat, sera présentée en détail dans le **Chapitre 4** de ce manuscrit.

2.3.2 Un positionnement intrinsèque de 2-3 nucléosomes aux bords des barrières nucléosomales commun aux 10 espèces

Le taux de GC est une bonne approximation de l'affinité d'une séquence pour le nucléosome. Aux bords des NIEBs humains, l'oscillation du taux de GC a été corrélée au positionnement nucléosomal, à la fois prédit et observé expérimentalement (Drillon et al., 2016). J'ai voulu étendre cette analyse aux autres espèces analysées ici, en produisant le profil d'occupation moyen en nucléosome prédits à partir de notre modèle aux bords des barrières de ces 10 espèces. À l'aide du modèle, j'ai d'abord calculé le profil d'occupation en nucléosomes paire de base par paire de base dans chaque génome. J'ai pu ensuite établir le profil moyen sur les 500 paires de base bordant les barrières nucléosomales dans chaque génome. Ici encore, comme en **Partie 2.3.1**, j'ai restreint l'analyse

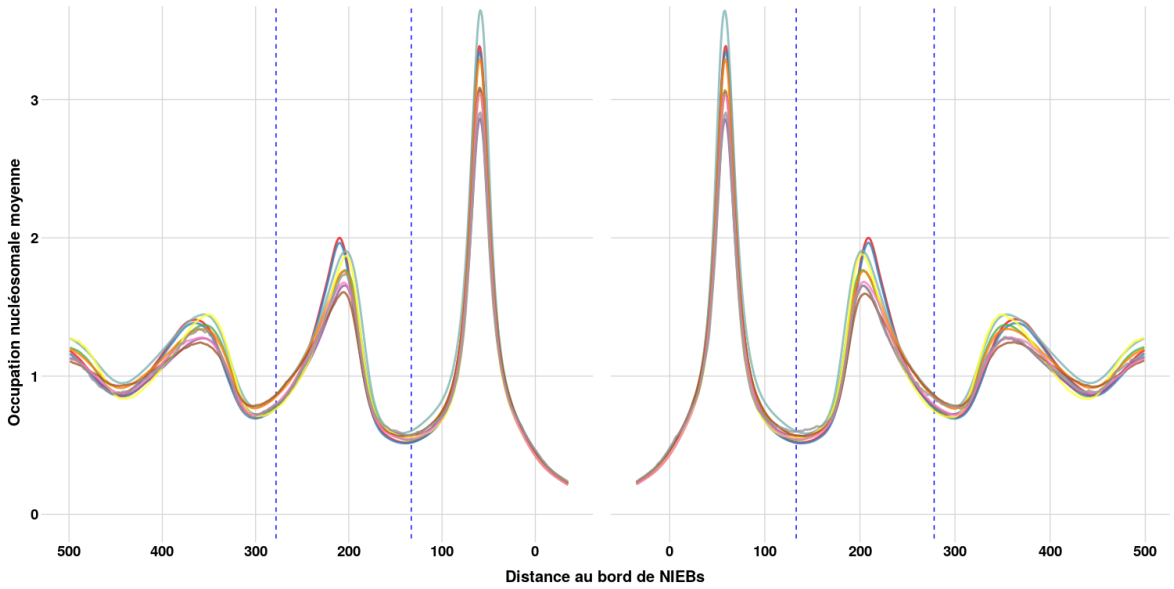


FIGURE 2.8 – **Profil moyen d'occupation nucléosomale prédite par le modèle aux bords des barrières dans les 10 espèces analysées.** Les couleurs correspondent à celles utilisées dans la **Figure 2.1**. L'analyse est restreinte aux inter-barrières de tailles supérieures à 1000 pb bordées par des barrières de tailles supérieures à 70 pb. Le nombre d'inter-NIEBs utilisés pour chaque espèce est indiqué dans le **Tableau 2.2**. Les lignes verticales pointillées bleues correspondent aux minima du taux de GC chez l'humain, aux abscisses 133 pb et 278 pb (**Figure 2.7**).

aux inter-barrières de tailles supérieures à 1000 pb bordées par des barrières de tailles supérieures à 70 pb, afin d'être sûr de contrôler la distance à la barrière la plus proche. Enfin, j'ai normalisé chaque profil moyen par la moyenne génomique d'occupation nucléosomale prédite par le modèle. Ainsi, dans la **Figure 2.8**, les valeurs d'occupation supérieures (resp. inférieures) à 1 indiquent que les positions correspondantes sont en moyenne plus (resp. moins) couvertes en nucléosomes que la moyenne du génome.

Sur la **Figure 2.8**, on observe une remarquable conservation de l'occupation en nucléosome aux bords des barrières nucléosomales. Dans chaque espèce, on observe la même forme de courbe, et quasiment les mêmes valeurs, à l'exception du premier pic pour lequel les valeurs peuvent légèrement différer selon les espèces). On remarque également une symétrie parfaite entre les parties gauche et droite de la figure. Le positionnement nucléosomal prédit est donc exactement le même qu'on se place en amont ou en aval d'une barrière. On observe trois pics dans l'occupation nucléosomale prédite. Au niveau de ces pics, l'occupation prédite est entre 1.5 fois (pour le troisième pic) et 3 fois (pour le premier pic) supérieure à l'occupation moyenne sur le génome. Les pics représentent donc des positions hautement préférentielles pour les nucléosomes. Entre ces positions préférentielles, on observe que l'occupation prédite est inférieure à la moyenne génomique (entre 0.5 et 0.75), les minima d'occupation nucléosomale correspondant aux positions des minima de la composition en GC. Enfin, aux bords internes des barrières, l'occupation prédite chute drastiquement, jusqu'à une valeur de 0.20, soit cinq fois inférieure à la moyenne génomique. Les courbes d'occupation nucléosomale indiquent que les barrières nucléosomales sont donc des zones fortement déplétées en nucléosomes. On observe également que les nucléosomes se formant autour de ces barrières ont un positionnement contraint, avec 3 positions préférentielles. On remarque que ces trois positions sont assez proches les unes des autres. En effet, elles sont

espacées d'environ 150 à 160 paires de base. Ces loci présentent donc une nucleosome repeat length (NRL) très courte, bien plus courte que la NRL moyenne du génome humain par exemple (qui est de 203 pb , Valouev et al. (2011)). Cette configuration nucléosomale très compacte, cohérente avec la distribution quantifiée des distances inter-NIEBs, et confirmée expérimentalement chez l'humain (Drillon et al., 2016) pourrait empêcher la condensation de la chromatine en fibre de 30 nm (Beshnova et al., 2014; Diesinger & Heermann, 2009; Everaers & Schiessel, 2015; Kepper et al., 2008; Lesne & Victor, 2006). L'analyse de données expérimentales pour confirmer les prédictions du positionnement nucléosomal présentées ici dans d'autres espèces a constitué une partie importante de cette thèse, dont les résultats sont présentés en détail dans le **Chapitre 3**.

Pour terminer, on observe que les courbes sont quasiment identiques d'une espèce à l'autre, ce qui est en accord avec les formes très similaires des oscillations des profils GC. Au niveau du premier pic, on observe néanmoins quelques différences dans la hauteur du pic, avec des valeurs allant d'environ 2.9 pour la levure ou la poule à près de 3.5 pour le porc. Mais la principale différence se situe au niveau des pics 2 et 3, où on observe un léger décalage du pic vers la droite, seulement chez l'humain et le chimpanzé. Ce décalage est très similaire à celui observé dans la **Partie 2.2.3** avec les tailles d'inter-barrières. On peut donc supposer qu'ici encore, la différence observée est due à la présence d'éléments Alu.

2.4 La conservation des barrières aux nucléosomes est très forte entre l'humain et le chimpanzé

Dans les **Parties 2.2** et **2.3**, on a pu voir que les caractéristiques des barrières nucléosomales étaient conservées entre les eucaryotes, que ce soit pour la taille des barrières et des inter-barrières, ou encore pour les oscillations du profil GC et pour l'occupation nucléosomale intrinsèque aux bords des barrières. La conservation de ces caractéristiques suggère une conservation globale des barrières nucléosomales entre les espèces. Les résultats obtenus chez l'humain et le chimpanzé, deux espèces très proches, sont même quasiment identiques. Les barrières nucléosomales étant encodées dans la séquence, si cette dernière est similaire entre deux espèces, on s'attend en effet à ce que les barrières le soient également. Mais cette supposition n'a encore jamais été vérifiée par la comparaison directe des barrières entre génomes. Pour étudier cette question, j'ai comparé les génomes de l'humain et du chimpanzé, afin de voir dans quelle mesure les barrières nucléosomales sont conservées entre les deux espèces. J'ai choisi ces espèces principalement pour leur proximité de séquence. En effet, les génomes de l'humain et du chimpanzé s'alignent particulièrement bien. De plus, le taux de mutations ponctuelles entre ces deux espèces est très faible (environ 1 %, Suntsova & Buzdin (2020)). La question est donc ici de savoir dans quelle mesure une conservation de la séquence entraîne une conservation des barrières nucléosomales ? Pour étudier cette question, je me suis basé sur les chaînes d'alignement des génomes humain et de chimpanzé disponibles dans la base de données de l'UCSC (https://hgdownload.soe.ucsc.edu/downloads.html), et qui sont souvent utilisées pour traduire des coordonnées génomiques d'un génome à l'autre avec le programme LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver). J'ai traduit ces chaînes sous la forme de couples d'intervalles alignés entre les deux espèces, sans insertion ou délétion. J'ai donc ici obtenu un jeu de données contenant les coordonnées des séquences alignées sans

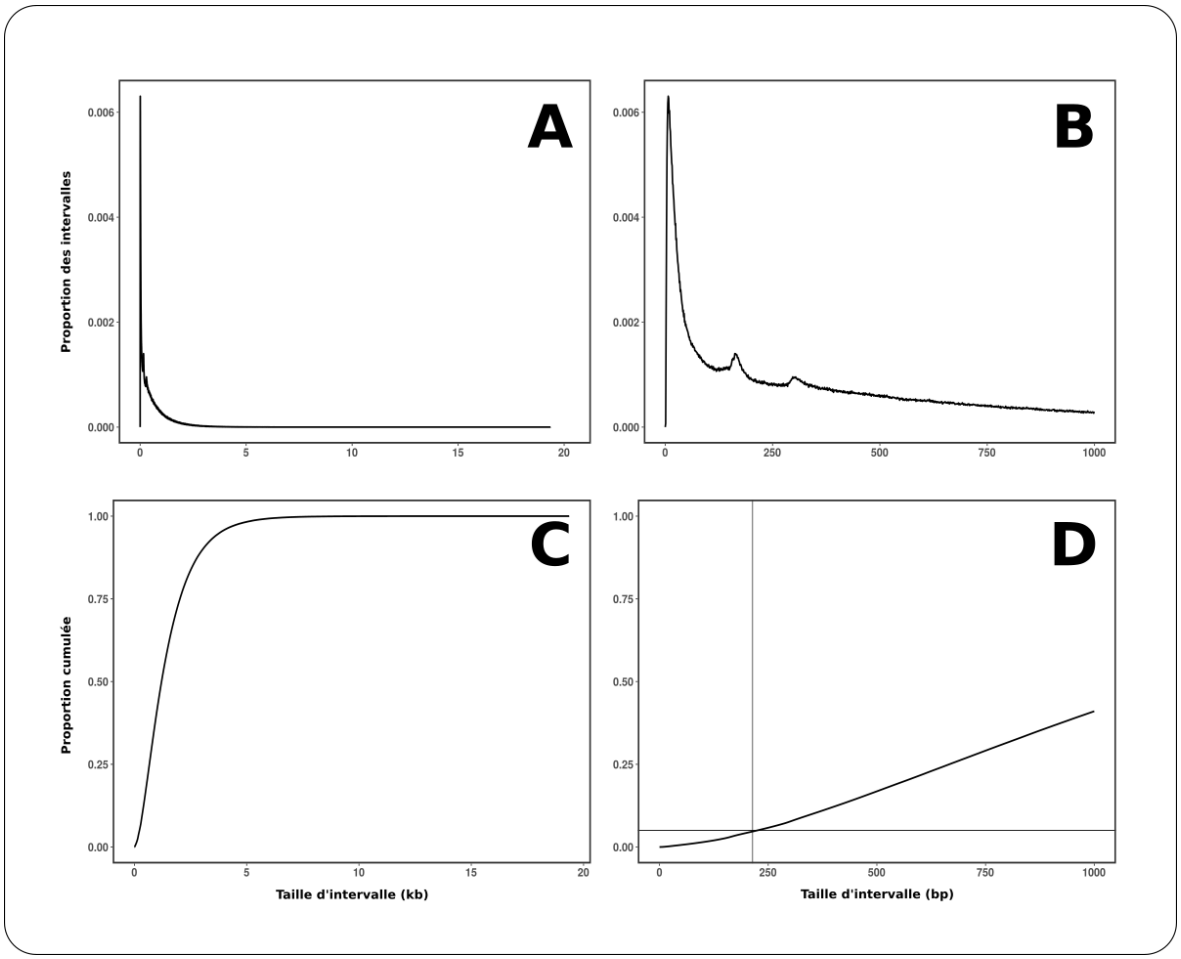


FIGURE 2.9 – **Distribution de la taille et proportion cumulée de la couverture totale des** 4 128 585 **intervalles alignés entre l'humain et le chimpanzé**. Le panel A représente la distribution de la taille. Le panel B est un zoom du panel A pour les tailles 0 à 1000 pb. Le panel C représente la proportion cumulée de la couverture de l'ensemble des intervalles plus petit qu'une certaine taille (en abscisse). Le panel D est un zoom du panel C pour les tailles 0 à 1000 pb. Les lignes horizontales et verticales du panel D correspondent respectivement à l'ordonnée 0.05 et à l'abscisse 214 pb.

gap entre l'humain et le chimpanzé. Ainsi, les seules différences entre les deux séquences d'un couple d'intervalles sont les mutations ponctuelles. Comme il était difficile de distinguer avec certitude le couple d'intervalles orthologue du (ou des) paralogue(s), j'ai pris la décision de retirer de l'analyse tous les intervalles concernés par une éventuelle duplication qui sont identifiés comme les intervalles, humain ou chimpanzé, présents dans plusieurs alignements. Les intervalles étant également utilisés pour la recherche de mutations ponctuelles entre l'humain et le chimpanzé (**Partie 2.5**), ce choix évite aussi l'introduction d'incohérences dans les patrons de mutations liées à la conservation d'intervalles dupliqués. Après ce retrait, les intervalles alignés couvrent encore respectivement 88 % et 90.9 % des génomes de l'humain et du chimpanzé. Si l'on se concentre sur les parties des génomes dont la séquence est connue (en retirant les bases "N" des génomes de référence), ces pourcentages grimpent respectivement à 91.8 % et 93.9 % pour l'humain et le chimpanzé. Le détail de ces couvertures chromosome par chromosome est présent en **Annexe A.4**. La **Figure 2.9** présente la distribution de la taille des intervalles considérés et la couverture associée. La grande majorité des intervalles sont de petites tailles, inférieures à 1000 pb (79.7 %) (**Figures 2.9 - A et B**). Cependant, ces intervalles de petites tailles ne représentent qu'une faible proportion de

$n_{1/1}$	$n_{1/0}$	$n_{0/1}$	**Couverture humain**	**Couverture chimpanzé**
164714785	37635323	39677213	81.4 %	80.59 %

TABLEAU 2.3 – **Conservation des bases appartenant aux barrières nucléosomales entre l'humain et le chimpanzé.** Les 3 premières colonnes correspondent respectivement aux nombres de bases appartenant à une barrière nucléosomale dans les deux espèces ($n_{1/1}$), chez l'humain uniquement ($n_{1/0}$) et chez le chimpanzé uniquement ($n_{0/1}$), comptés sur les intervalles communs à l'humain et au chimpanzé (Partie 2.4.1). La colonne "Couverture humain" indique la couverture des NIEBs humains par ceux du chimpanzé. La colonne "Couverture chimpanzé" indique la couverture des NIEBs du chimpanzé par ceux de l'humain.

la couverture (**Figures 2.9 - C et D**). En fait, moins de la moitié (41 %) des bases couvertes par les intervalles alignés sont présentes dans des intervalles de tailles inférieures à 1000 pb. On remarque cependant sur la **Figure 2.9 - C** que la proportion cumulée atteint quasiment 1 aux alentours de 5 kb, ce qui signifie qu'en terme de couverture, quasiment toutes les données (98.3 %) sont contenues dans les intervalles de tailles inférieures à 5000 pb. Seulement 5 % de la couverture totale est contenue dans les intervalles de tailles inférieures à 214 pb (**Figure 2.9 - D**). Ainsi, j'ai décidé de ne travailler qu'avec les intervalles de tailles supérieures à 200 pb, ce qui me permettait d'enlever tous les intervalles de trop petites tailles, tout en conservant plus de 95 % des données en terme de couverture (95.75 %). J'ai retiré les petits intervalles de l'analyse car on veut ici étudier la conservation des barrières dans le cas d'une conservation de séquence. Il ne paraissait donc pas pertinent de travailler avec des séquences conservées sur seulement quelques dizaines de paires de bases, d'autant plus que les barrières ont une taille moyenne de ~ 136 pb chez l'homme et le chimpanzé (**Tableau 2.2**). Le retrait de ces intervalles, même s'ils sont nombreux, n'affecte que très peu la couverture des génomes par les intervalles alignés restant. Il semblait donc raisonnable d'appliquer ce filtrage des données. Après filtrage, les couvertures des régions séquencées des génomes humain et chimpanzé par les intervalles sont respectivement de 87.9 % et 89.9 %.

2.4.1 20% des barrières aux nucléosomes ne sont pas partagées entre humain et chimpanzé

Tout d'abord, j'ai voulu voir quelle était la couverture globale des barrières de l'humain par celles du chimpanzé et vice-versa. Pour ça, j'ai établi, à partir de la taille des chromosomes et des coordonnées des barrières nucléosomales dans les deux espèces, des vecteurs binaires où 0 correspondait à une base n'appartenant pas à une barrière nucléosomale, et 1 à une base appartenant à une barrière nucléosomale. À partir de ces vecteurs, j'ai pu comparer très facilement les bases appartenant ou non à des barrières dans chaque espèce et dans chaque intervalle aligné, afin d'obtenir :

— Le nombre de bases «barrière» dans les deux espèces ($n_{1/1}$)

— Le nombre de bases «barrière» seulement chez l'humain ($n_{1/0}$)

— Le nombre de bases «barrière» seulement chez le chimpanzé ($n_{0/1}$)

À partir de ces chiffres, on peut calculer la couverture des barrières humaines par les barrières chimpanzé ($n_{1/1}/(n_{1/1}+n_{1/0})$) et vice-versa ($n_{1/1}/(n_{1/1}+n_{0/1})$). Les résultats obtenus avec cette méthode sont présentés dans le **Tableau 2.3**.

Tout d'abord, on peut, à partir du **Tableau 2.3**, déterminer quel pourcentage des barrières de

chaque espèce se situe dans les intervalles. En effet, on observe que 202 350 108 pb ($n_{1/1} + n_{1/0}$) appartenant à une barrière humaine sont situées dans un intervalle aligné entre les deux espèces. Sachant que la taille totale des barrières humaines est de 235 131 477, on a donc environ 86 % des bases appartenant à une barrière nucléosomale humaine qui sont situées dans les intervalles alignés entre l'humain et le chimpanzé. On remarque que ce chiffre est très proche des 87.9 % de couverture du génome par les intervalles. Chez le chimpanzé, les chiffres sont similaires, avec 204 391 998 pb «barrière» situées dans les intervalles alignés, soit environ 86.3 % des bases «barrière» totales, proche des 89.9 % de couverture. Le pourcentage de barrières couvrant les intervalles est néanmoins très légèrement inférieur à la couverture globale du génome. Cela pourrait suggérer que la présence de barrières nucléosomales favorise très légèrement la quantité d'insertions ou de délétions, car ce type d'évènement est exclu des intervalles. Dans le **Tableau 2.3**, on voit que la couverture des barrières d'une espèce par celle de l'autre espèce est légèrement supérieure à 80 % dans les deux cas (81.4 % pour l'humain et 80.6 % pour le chimpanzé). Étant donné que l'on travaille ici avec des séquences parfaitement alignées, dont les seules différences sont des mutations ponctuelles, et que la prédiction de la position des barrières nucléosomales se base uniquement sur la séquence, on pourrait s'attendre à une meilleure couverture que ces 80 %. Cependant, si les intervalles sont en effet parfaitement alignés, on n'a pas ici l'information sur la taille du gap bordant les intervalles dans l'une ou l'autre des deux espèces (ou dans les deux). Il est donc possible que la différence de ~ 20 % de barrières soit due à un "effet de bord". Dans ces données, chaque intervalle, par construction, est bordé de chaque côté par une mutation de type indel. Il est possible que ces insertions et délétions aient une influence sur la présence de barrières dans les séquences adjacentes des intervalles. Ainsi, on pourrait avoir, aux bords des intervalles, des différences dans la composition en barrières nucléosomales entre les deux espèces, et ce même si hormis les indels, les séquences sont parfaitement alignées, voire parfaitement identiques. Ce principe d'effet de bord est résumé dans la **Figure 2.10**.

Afin de voir si les différences observées précédemment s'expliquent par un effet de bord dû aux indels bordant les intervalles, j'ai voulu retirer les bords d'intervalles de l'analyse pour voir si cela modifiait le résultat. Pour ça, j'ai d'abord sélectionné les intervalles de tailles supérieures à 2500 pb. J'ai ensuite retiré de l'analyse les 500 pb (resp. 1000 pb) à chaque extrémité des intervalles. Cela permet de conserver au moins 1500 pb (reps. 500 pb) au centre de l'intervalle, bordés par au moins 500 pb (resp. 1000 pb) sans indel. Si l'on revient à la **Figure 2.10**, il s'agit donc de se concentrer sur la partie centrale des séquences alignées (en bleu). Si les différences observées dans le **Tableau 2.3** sont dues à un effet de bord lié aux indel, elles devraient être gommées par ces retraits, de façon progressive selon la distance jusqu'à laquelle une indel peut influer sur les séquences adjacentes. Ainsi, la couverture des NIEBs d'une espèce par ceux de l'autre espèce devrait être sensiblement plus importante avec le retrait de 1000 pb. Les résultats obtenus avec les deux retraits sont présentés en **Tableau 2.4** (retrait de 500 pb) et en **Tableau 2.5** (retrait de 1000 pb).

Que l'on retire 500 pb ou 1000 pb au bord de chaque intervalle, la couverture des barrières d'une espèce par celles de l'autre n'augmente que de 2.5 % à 3 % par rapport aux intervalles complets (**Tableaux 2.3, 2.4, 2.5**). On en déduit que l'effet de bord est faible. De plus, la distance jusqu'à laquelle une indel a une influence sur la composition en barrière nucléosomale est courte, inférieure à 500 pb, car l'augmentation de la couverture entre le retrait de 500 et 1000 pb est très légère. En somme, même si un effet de bord est effectivement présent (car la couverture augmente légère-

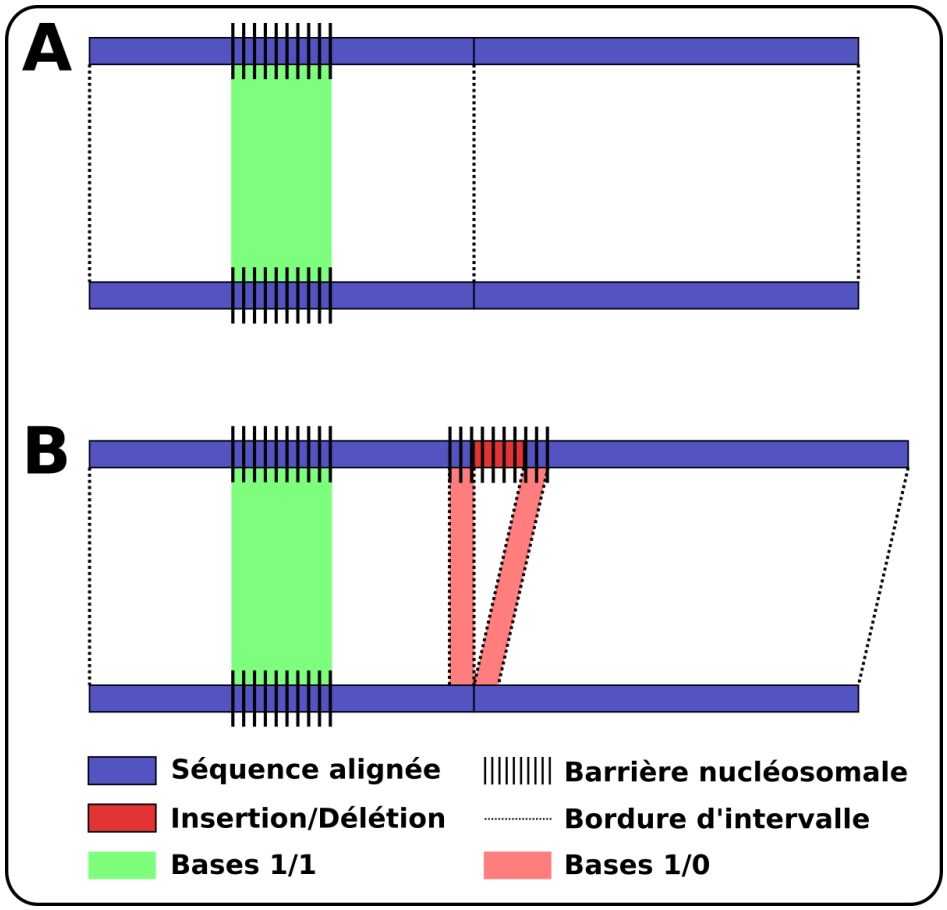


FIGURE 2.10 – **Schéma explicatif de l'effet de bord.** Le panel A illustre l'alignement des séquences sans l'insertion, avec un cas idéal d'une barrière au milieu d'un intervalle que l'on retrouve exactement dans l'intervalle correspondant dans l'autre génome. Dans le panel B est ajoutée une insertion entre les deux intervalles dans le premier génome. Cette insertion est à l'origine d'une barrière nucléosomale, qui s'étend également aux séquences alignées adjacentes à l'insertion. Ainsi, dans la comparaison des barrières entre les deux génomes, des bases de type 1/0 (identifiées ici en rouge) sont trouvées même là où les séquences sont parfaitement identiques. Ceci illustre un effet de bord résultant d'une insertion dans un des deux génomes.

$n_{1/1}$	$n_{1/0}$	$n_{0/1}$	**Couverture humain**	**Couverture chimpanzé**
15586001	2856668	3177355	84.5 %	83.1 %

TABLEAU 2.4 – **Conservation des bases appartenant aux barrières nucléosomales entre l'humain et le chimpanzé avec masquage de 500 pb aux bords des intervalles communs.** Les 3 premières colonnes correspondent respectivement aux nombres de bases appartenant à une barrière nucléosomales dans les deux espèces ($n_{1/1}$), chez l'humain uniquement ($n_{1/0}$) et chez le chimpanzé uniquement ($n_{0/1}$), comptés sur les intervalles communs à l'humain et au chimpanzé de tailles supérieures à 2500 pb après masquage de 500 pb à chaque extrémité (**Partie 2.4.1**). La colonne "Couverture humain" indique la couverture des NIEBs humain par ceux du chimpanzé. La colonne "Couverture chimpanzé" indique la couverture des NIEBs du chimpanzé par ceux de l'humain.

ment), la présence d'indel aux bords des intervalles n'explique pas les différences observées entre les barrières humaines et les barrières chimpanzé. Ces différences sont donc dues principalement aux mutations ponctuelles. Plusieurs hypothèses sont alors envisageables :

— Les bords de certaines des barrières ne sont pas aux mêmes coordonnées dans les deux espèces. Cela implique que quelques mutations ponctuelles suffiraient à déplacer significativement les bords de barrières.

1/1	1/0	0/1	Couverture humain	Couverture chimpanzé
8 877 461	1 601 664	1 809 967	84.7%	83.1%

TABLEAU 2.5 – **Conservation des bases appartenant aux barrières nucléosomales entre l'humain et le chimpanzé avec masquage de 1000 pb aux bords des intervalles communs.** Les 3 premières colonnes correspondent respectivement aux nombres de bases appartenant à une barrière nucléosomales dans les deux espèces ($n_{1/1}$), chez l'humain uniquement ($n_{1/0}$) et chez le chimpanzé uniquement ($n_{0/1}$), comptés sur les intervalles communs à l'humain et au chimpanzé de tailles supérieures à 2500 pb après masquage de 1000 pb à chaque extrémité (Partie 2.4.1). La colonne "Couverture humain" indique la couverture des NIEBs humain par ceux du chimpanzé. La colonne "Couverture chimpanzé" indique la couverture des NIEBs du chimpanzé par ceux de l'humain.

— Certaines barrières sont présentes dans une espèce et absentes dans l'autre. Cela implique que les mutations ponctuelles suffiraient à former (ou détruire) une barrière nucléosomale.

2.4.2 Des faux négatifs lors de la détection des barrières ?

Les résultats discutés dans la **Partie 2.4.1** amènent à une nouvelle question concernant les différences trouvées entre les barrières nucléosomales de l'humain et du chimpanzé. Ces différences sont-elles dues à des «mouvements» de barrières, c'est-à-dire à des barrières dont les bords seraient décalés dans l'un ou l'autre des génomes ? Ou bien sont-elles dues à des «pertes» (ou des «gains») de barrières dans l'une ou l'autre des deux espèces ? Pour répondre à cette question, j'ai cherché à associer chaque barrière du génome humain se trouvant dans un intervalle aligné avec le génome du chimpanzé à sa barrière homologue chez le chimpanzé, et vice-versa. Il est à noter qu'afin d'éviter d'éventuelles erreurs ou biais liés aux indels, je n'ai considéré ici que les barrières entièrement comprises dans un intervalle aligné entre les deux espèces. Ainsi, toutes les barrières chevauchant deux intervalles adjacents dans l'une ou l'autre des deux espèces, ce qui peut par exemple correspondre à une barrière dans laquelle il y a eu une insertion dans une des deux espèces, ont été retirées de l'analyse. Ce filtrage permet tout de même de travailler avec près d'un million de barrières dans chaque espèce (959 808 barrières humaines et 956 258 barrières chimpanzé).

J'ai produit un jeu de données de couples de barrières homologues dans les deux espèces. Je me suis servi des coordonnées des intervalles alignés entre les deux espèces pour traduire les positions des barrières humaines (resp. du chimpanzé) chez le chimpanzé (resp. chez l'humain), et comparer ces positions à celles des barrières du chimpanzé (resp. de l'humain). Cela m'a permis de calculer la couverture mutuelle de chaque couple de barrières, à savoir le pourcentage de la barrière humaine couverte par la barrière chimpanzé associée et vice-versa. La couverture mutuelle est donc représentée par deux chiffres, compris entre 0 et 1, correspondant aux deux pourcentages. Ainsi, une couverture mutuelle de (1, 1) représente une barrière identique chez l'humain et le chimpanzé. Une couverture mutuelle de (0.5, 1) correspond à un couple de barrières pour lequel la barrière humaine a été agrandie par rapport à celle du chimpanzé (ou celle du chimpanzé a été réduite). En effet, dans ce cas, seule la moitié de la barrière humaine est couverte par la barrière du chimpanzé alors que cette dernière est totalement englobée dans la barrière humaine, qui est donc plus grande. Enfin, une couverture mutuelle de (0.5, 0.5) correspond à un couple de barrières dont les bords se seraient décalés entre les deux espèces. L'analyse de couverture mutuelle ainsi que les différents cas de figures évoqués sont illustrés par les schémas de la **Figure 2.11**. Enfin, il

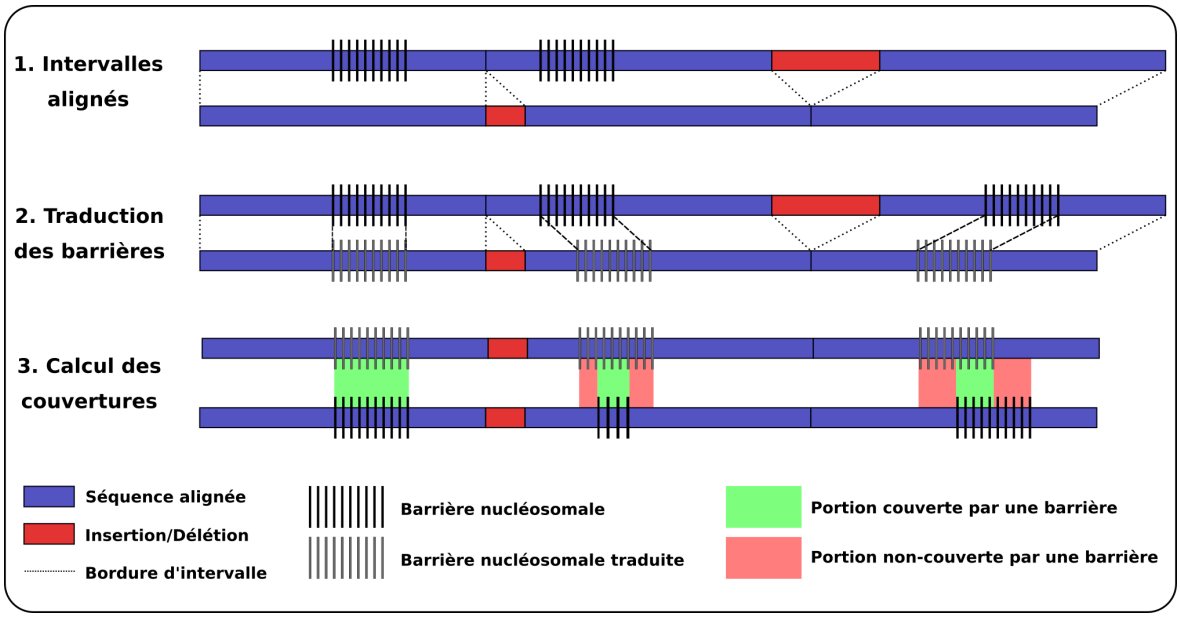


FIGURE 2.11 – **Schéma explicatif du calcul de la couverture mutuelle pour un couple de barrières.** Le panel 1 illustre l'alignement des génomes, avec les insertions (en rouge) séparant les intervalles alignés (en bleu). Les barrières nucléosomales du premier génome sont indiquées par les zones hachurées. Le panel 2 illustre la traduction des coordonnées des barrières du génome 1 dans le génome 2. Le panel 3 illustre la comparaison des coordonnées traduites avec les barrières du génome 2 afin de déterminer la couverture des barrières du génome 1 par celle du génome 2. Il est à noter que cette analyse est répétée en inversant les génomes 1 et 2, afin d'obtenir la couverture des barrières du génome 2 par celles du génome 1. À l'issue des deux analyses, une couverture mutuelle est associée à chaque couple de barrières, comme expliqué en **Partie 2.4.2**

est également possible qu'une barrière soit spécifique d'une des deux espèces, et donc totalement absente du génome de l'autre espèce. Ces barrières sont identifiées comme ayant une couverture mutuelle de (0,1) pour les barrières spécifiques à l'humain, et de (1,0) pour les barrières spécifiques au chimpanzé.

Une couverture mutuelle est donc associée à chaque couple de barrières présent dans un intervalle. Trois cas de figures sont possibles :

— Une barrière humaine et une barrière chimpanzé sont associées, avec une couverture mutuelle comprise dans les intervalles suivants : (]0, 1],]0, 1]).

— La barrière humaine n'est associée à aucune barrière chimpanzé et est donc identifiée comme spécifique à l'humain. La couverture mutuelle est donc de (0, 1).

— La barrière chimpanzé n'est associée à aucune barrière humaine, et est donc identifiée comme spécifique au chimpanzé. La couverture mutuelle est donc de (1,0).

Pour rappel, environ 20 % des bases «barrières» de chaque espèce sont identifiées comme des bases «non-barrières» dans l'autre espèce (**Tableau 2.3**). La question est ici de savoir si les mutations ponctuelles, qui semblent être responsables de ces 20 % de différence, le sont parce qu'elles déplacent les bords des NIEBs ou parce qu'elles suppriment/ajoutent des NIEBs ? Pour répondre à cette question, on peut étudier la répartition des différents cas de figure listés ci-avant. Dans l'hypothèse des «mouvements de barrières», on ne devrait observer que très peu de barrières totalement spécifiques de l'une ou l'autre des deux espèces, et beaucoup de barrières pour lesquelles le chevauchement est incomplet. Cela se traduirait en terme de couvertures mutuelles par très peu de couvertures à (0, 1) ou (1, 0), et beaucoup de couvertures dont les deux pourcentages

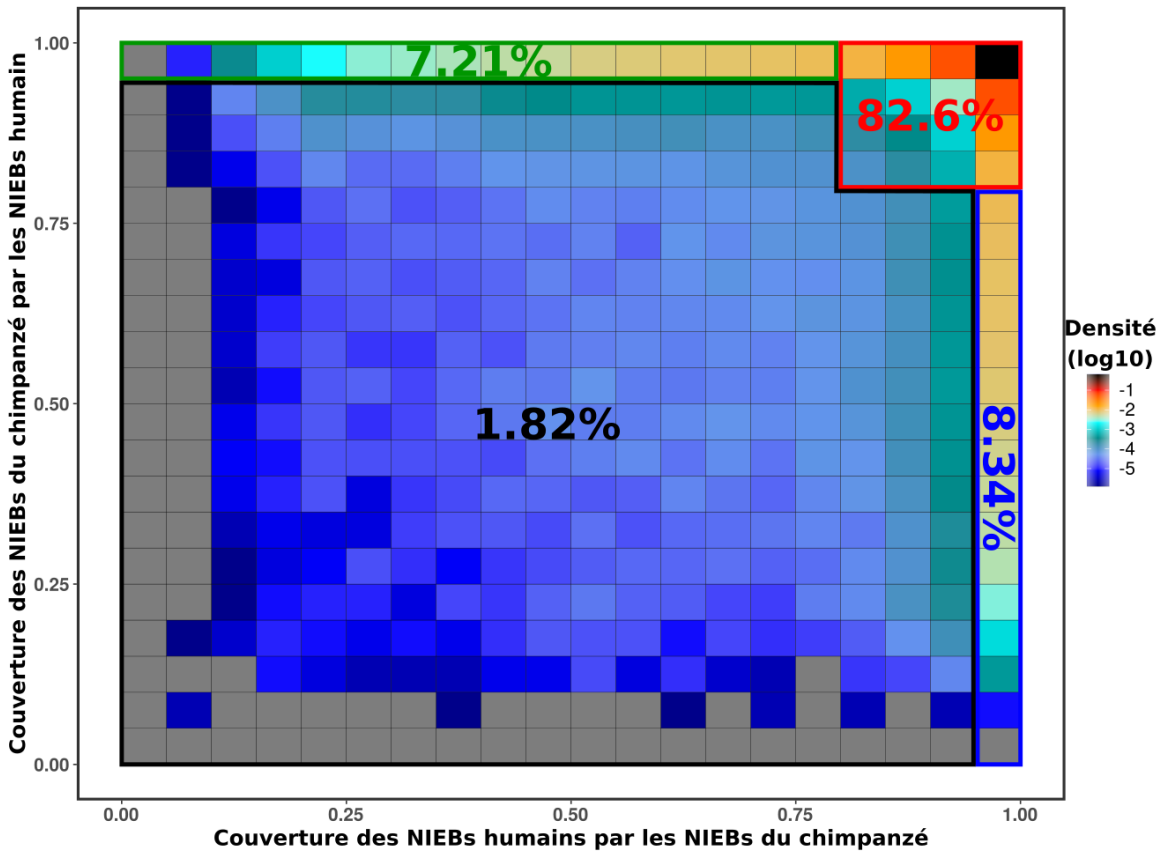


FIGURE 2.12 – **Distribution 2D des couvertures mutuelles des** 810 002 **couples de barrières identifiés entre l'humain et le chimpanzé**. La valeur en abcisse représente la couverture de la barrière humaine du couple par la barrière chimpanzé. La valeur en ordonnée représente la couverture de la barrière chimpanzé du couple par la barrière humaine.

seraient assez faibles, par exemple inférieurs à 0.8. Dans l'hypothèse de l'ajout ou de la suppression de NIEBs par les mutations ponctuelles, on devrait au contraire observer beaucoup de barrières spécifiques d'une des deux espèces, et peu de chevauchement incomplet entre couple de barrières. En terme de couvertures mutuelles, cela se traduirait par un nombre important de couvertures à (0, 1) et (1, 0) illustrant les barrières spécifiques, et également de (1,1) illustrant les barrières n'ayant pas été modifiées.

Le calcul des couvertures mutuelles des couples de barrières homologues chez l'humain et le chimpanzé indique que plus de 15 % des barrières de chaque espèce sont identifiées comme spécifiques à cette espèce. Comme expliqué dans le paragraphe précédent, ce chiffre est en faveur de l'hypothèse d'ajout/suppression de NIEBs par les mutations ponctuelles. En effet, il est très proche des 20 % de différences observées plus tôt, et quasi identique aux chiffres obtenus lorsqu'on s'est affranchi de l'effet de bord (**Tableaux 2.4** et **2.5**). Il semble donc que l'hypothèse de mouvements de barrières ne soit pas validées. Pour confirmer ce résultat, j'ai déterminé la distribution 2D des couvertures mutuelles (**Figure 2.12**). Il est à noter que ne sont considérés pour cette distribution que les couples de barrières, et pas les barrières identifiées comme spécifiques à chaque génome. On observe que parmi les couples de barrières, la très grande majorité correspond à des barrières quasi identiques entre l'humain et le chimpanzé (**Figure 2.12**). En effet, pour plus de 80% des couples de barrières, la couverture de la barrière humaine par la barrière chimpanzé comme la couverture inverse sont toutes deux supérieures à 80 %. Ces deux couvertures sont même supérieures à 90 %

pour plus des trois quarts des couples de barrières (76.3 %), et supérieures à 95 % pour 64.58 % des couples de barrières. La majorité du reste des données se situe sur la ligne Y >= 0.95 et la colonne X >= 0.95 (avec respectivement 7.21 % et 8.34 % des couples de barrières). Comme expliqué dans le second cas du schéma de la **Figure 2.11**, ces cas correspondent à des barrières s'étant agrandies dans une espèce ou s'étant réduites dans l'autre. Enfin, seulement 1.82 % des couples correspondent au dernier cas du schéma de la **Figure 2.11**, à savoir celui d'un mouvement de barrière entre les deux espèces. Cela confirme que l'hypothèse de mouvement de barrières n'explique pas les différences observées dans la **Partie 2.4.1**. Les barrières nucléosomales semblent donc être ajoutées (ou supprimées) par les mutations ponctuelles. Cet ajout/suppression peut concerner l'entièreté de la barrière (comme le montrent les plus de 15 % de NIEBs identifiés comme spécifiques à l'humain ou au chimpanzé), ou seulement une partie (comme le montrent les 7 à 8 % de couples de NIEBs dont la couverture mutuelle est de (]0, 0.80[,]0.95,1]) ou (]0.95,1],]0, 0.80[).

2.4.3 Les profils d'énergie et d'occupation intrinsèques confirment la conservation des barrières aux nuclésomes entre les deux espèces

Dans la **Partie 2.4.1**, nous avons vu que dans les intervalles alignés entre l'humain et le chimpanzé, environ 20 % des bases appartenant à une barrière chez l'humain ne sont pas identifiées comme appartenant à une barrière chez le chimpanzé, et vice-versa. Nous avons également vu que cette différence ne s'explique pas par un effet de bord. Les résultats discutés dans la **Partie 2.4.2** indiquent également que des «mouvements» de barrières n'expliquent pas non plus le résultat. On pourrait donc avoir soit l'apparition ou la disparition de barrières provoquée par les mutations ponctuelles, soit un problème de détection lié par exemple à un effet de seuil. Lors de la détection des barrières, le logiciel de détection des NIEBs calcule, à partir de la séquence, l'énergie nécessaire pour former un nucléosome à chaque locus du génome. Il en résulte un profil énergétique, à partir duquel on peut prédire le positionnement nucléosomal et donc la position des barrières nucléosomales. Les barrières nucléosomales sont identifiées comme des zones où la formation du nucléosome est inhibée par la séquence. La question est alors : « A partir de quand considère t-on une séquence comme inhibitrice ?». Concrètement, il s'agit de définir un seuil, à partir duquel une séquence sera identifiée comme une barrière nucléosomale. Ce seuil est arbitraire. Dans notre cas, il a été ajusté lors du dévelopement du lociciel par la confrontation des prédictions aux données expérimentales, pour que le seuil choisi soit celui qui permette au modèle de reproduire le plus fidèlement possible les données expérimentales de positionnement nucléosomal chez la levure (Chevereau et al., 2011). Ici, on suspecte que le choix du seuil soit à l'origine des ~ 15 % de barrières identifiées comme spécifiques à chaque espèces (**Partie 2.4.2**). J'ai donc comparé directement les profils énergétiques de formation du nucléosome et d'occupation nucléosomale prédits par le modèle dans les deux espèces au niveau des barrières conservées entre l'humain et le chimpanzé et des barrières spécifiques à l'un des deux génomes. Aux barrières identifiées comme identiques entre les deux espèces, les profils devraient être très proches, voire identiques. Dans le cas d'un effet de seuil, les profils aux barrières spécifiques devraient également être très proches entre les deux espèces, mais avec les profils de l'espèce possédant la barrière plus marqués que ceux de l'espèce ne possédant pas la barrière. S'il n'y a pas d'effet de seuil, alors on devrait observer des profils différents pour les zones correspondant aux barrières spécifiques, avec des profils de type

«barrière» pour l'espèce possédant une barrière et un autre type de profil (a priori plutôt plat) pour l'espèce ne possédant pas la barrière. Pour chaque barrière humaine, j'ai récupéré les profils d'énergie de formation du nucléosome et d'occupation nucléosomame de la zone contenant la barrière chez l'humain et de la zone correspondante chez le chimpanzé. J'ai aligné ces profils entre eux sur les positions des bords de barrières (qu'on retrouve en position 0 et 450 sur la **Figure 2.13**) et j'ai calculé à chaque position la moyenne sur toutes les barrières alignées. J'ai effectué ce calcul pour les trois groupes de barrières identifiés, à savoir les couples de barrières homologues, les barrières spécifiques à l'humain et les barrières spécifiques au chimpanzé. L'analyse aux couples de barrières homologues est restreinte aux couples dont les couvertures mutuelles sont supérieures à 0.95 (]0.95, 1],]0.95, 1]). J'ai effectué ce filtrage afin d'avoir comme référence des couples de barrières quasiment identiques. En restreignant l'analyse aux couvertures mutuelles supérieures à 0.95, je m'assure d'éviter des biais liés à des barrières légèrement différentes, tout en gardant une quantité importante de données (plus de 500 000 couples de barrières). Les résultats obtenus en prenant en compte tous les couples de barrières sont présentés en **Annexe A.3**. On peut y voir qu'ils sont très proches de ceux présentés dans la **Figure 2.13**. Ainsi, ce filtrage des couples de barrières homologues ne change en rien l'interprétation de la **Figure 2.13**.

Les profils d'énergie de formation du nucléosome présentés dans les panneaux A et B de la **Figure 2.13** confirment l'existence d'un effet de seuil dans le logiciel de détection des NIEBs, expliquant les différences de barrières nucléosomales entre l'humain et le chimpanzé observées dans la **Partie 2.4.1**. Pour les barrières homologues , on obtient exactement le même profil énergétique dans les deux espèces, comme le montre la superposition parfaite des courbes bleue et rouge (panel A) pour les barrières humaines et violette et verte (panel B) pour les barrières du chimpanzé. On observe également que pour les barrières identifiées comme spécifiques à l'humain, le profil énergétique "barrière" calculé sur le génome humain (panel A, courbe orange) a la même forme que le profil de la zone correspondante identifiée "non-barrière" chez le chimpanzé (panel A, courbe marron). Cependant, on observe une légère différence de valeurs entre les deux courbes, le profil énergétique de la situation "non-barrière" est légèrement en dessous de celui de la situation "barrière". De plus, ces deux courbes présentent des valeurs bien inférieures aux courbes correspondant aux couples de barrières conservées. On est donc précisément dans le cas décrit précédemment d'un effet de seuil de détection. Les zones chez le chimpanzé qui correspondent aux barrières identifiées comme spécifiques à l'humain semblent donc se comporter comme des barrières nucléosomales elles aussi, mais légèrement plus faibles, ce qui explique qu'elles n'aient pas été détectées comme telles par le modèle. Ces zones semblent donc être des faux-négatifs de la détection de barrières nucléosomales. Elles auraient dû être identifiées comme des barrières nucléosomales, comme chez l'humain, mais quelques mutations ponctuelles ont fait diminuer leur niveau énergétique, le faisant passer sous le seuil de détection. On observe exactement le même phénomène avec les barrières spécifiques au chimpanzé (courbes rose et grise du panel B). Les différences d'énergie de formation du nucléosome se retrouvent également dans les profils d'occupation nucléosomale prédits par le modèle (panneaux C et D). L'occupation prédite est quasiment identique chez les deux espèces pour les couples de barrières homologues (courbe bleue et rouge du panel C, verte et violette du panel D). Mais là où l'énergie était plus forte pour la situation barrière que pour la situation non-barrière, l'occupation est plus faible (courbes orange et marron du panel C, rose et grise du panel D). Comme attendu, l'occupation varie à l'inverse de

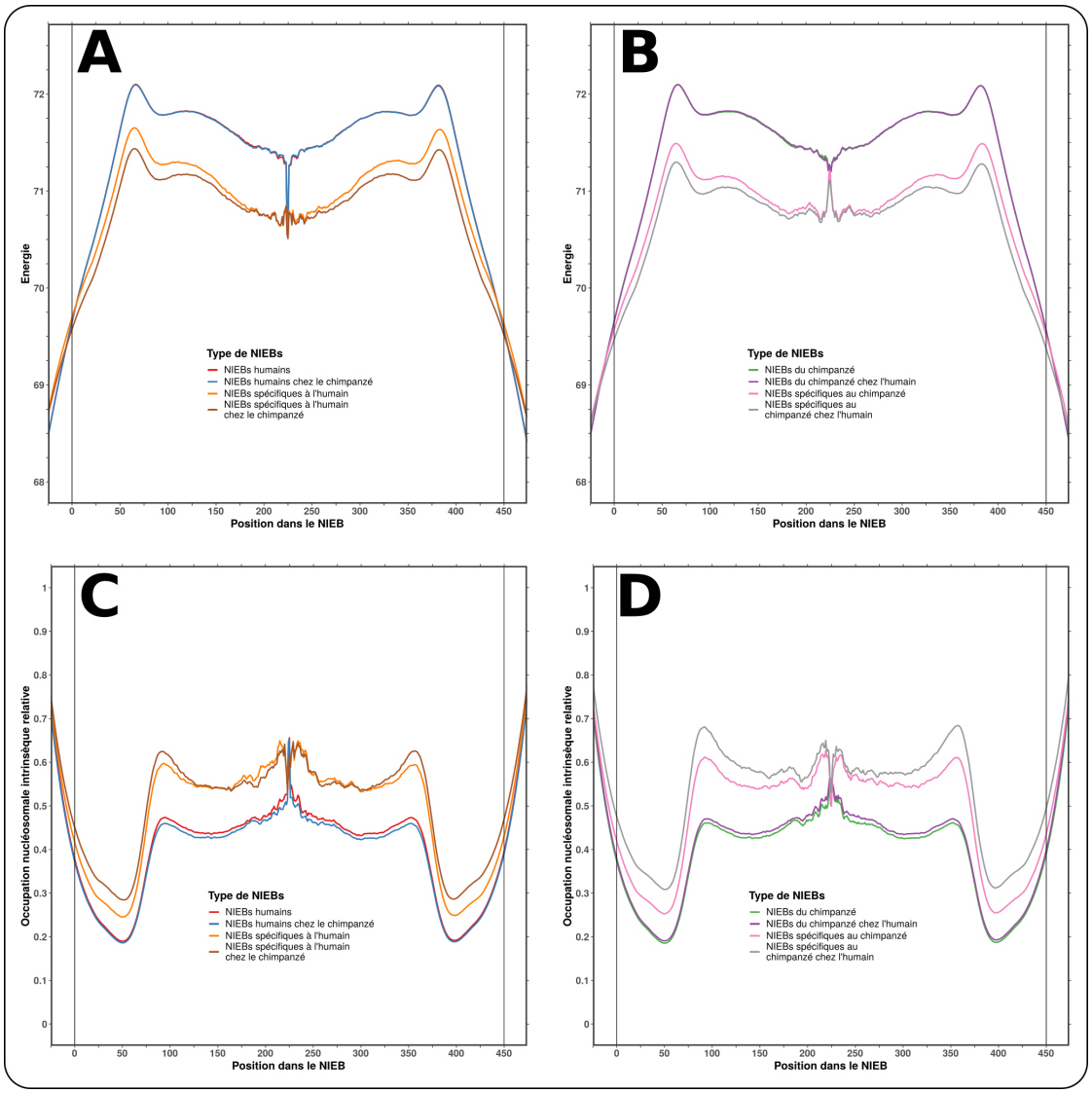


FIGURE 2.13 – **Profils d'énergie de formation du nucléosome et d'occupation nucléosomale prédits pour les couples de NIEBs homologues et les NIEBs spécifiques à chaque espèce.** Les panneaux A et B représentent l'énergie de formation du nucléosome calculée par le modèle physique de positionnement du nucléosome, les panneaux C et D représentent l'occupation nucléosomale intrinsèque qui en résulte. Les courbes rouge et bleue des panneaux A et C, ainsi que les courbes violette et verte des panels B et D correspondent aux profils pour les 523 084 couples de NIEBs quasi-identiques avec des couvertures mutuelles dans (]0.95, 1],]0.95, 1]). En rouge, les profils obtenus avec les NIEBs humains chez l'humain. En bleu, les profils obtenus sur la séquence chimpanzé en traduisant les coordonnées des NIEBs humains chez le chimpanzé. En vert, les profils obtenus avec les NIEBs du chimpanzé chez le chimpanzé. En violet, les profils obtenus en traduisant les coordonnées des NIEBs du chimpanzé chez l'humain. Les courbes orange et marron correspondent aux 149 806 NIEBs identifiés comme spécifiques à l'humain, avec les profils obtenus sur la séquence humaine (orange) et du chimpanzé (marron). De la même manière, les courbes rose et grise correspondent aux 146 256 NIEBs identifiés comme spécifiques au chimpanzé, avec les profils obtenus chez le chimpanzé (rose) et chez l'humain (gris). Toutes les courbes d'occupation nucléosomale sont normalisées par la moyenne génomique. Ainsi, les valeurs des courbes d'occupation sont directement comparables à celles de la **Figure 2.8**. L'axe des abscisses est la position dans les NIEBs. Les NIEBs étant de taille variables entre 23 et 450 pb, ils sont alignés sur leurs deux bords. L'abscisse 0 pb correspond au bord gauche des NIEBs et l'abscisse 450 pb au bord droit. Pour une abscisse x à gauche de la figure (et son pendant à droite : $450 - x$), seuls les NIEBs ayant une taille d'au moins $2 \times x$ sont considérés. Ainsi, plus on s'éloigne de ces bords, plus le nombre de NIEBs utilisés diminue. La taille médiane des NIEBs étant de 113 pb et 115 pb chez l'humain et le chimpanzé (**Tableau 2.2**), l'augmentation de l'occupation nucléosomale prédite à l'intérieur des NIEBs ne concerne donc qu'une petite partie des NIEBs. On remarque qu'elle reste près de deux fois inférieure à la moyenne génomique. Cette observation spécifique aux NIEBs les plus grands, qui n'avait pas été faite auparavant, ne sera pas analysée ici.

l'énergie de formation. Ce résultat montre donc que la conservation des barrières nucléosomale est bien meilleure que suggéré par les ~ 80 % de couvertures mutuelles des régions identifiées comme NIEBs (**Partie 2.4.1**) ; le long des régions alignées, quasiment toutes les barrières nucléosomales sont donc conservées entre ces deux espèces, les différences observées venant principalement d'un effet de seuil de détection. Pour la plupart des barrières, le pouvoir inhibiteur de la formation des nucléosomes (représenté par la hauteur du profil énergétique) est assez important pour ne pas être perturbé par les quelques mutations ponctuelles entre l'humain et le chimpanzé. Dans ce cas, deux séquences qui s'alignent entre les deux espèces s'alignent également en terme de barrières nucléosomales. Cependant, pour certaines barrières, le pouvoir inhibiteur est proche du seuil de détection, et quelques mutations ponctuelles peuvent faire basculer une zone d'un état barrière détectée à un état barrière non-détectée et vice-versa. C'est principalement ce mécanisme qui explique les différences de barrières entre l'humain et le chimpanzé. On peut donc conclure que lorsque les séquences sont conservées et s'alignent (hors indel), les NIEBs le sont également.

2.5 Positionnement intrinsèque des nucléosomes et patrons de mutations

Les exemples de sélection de séquences pour leurs effets sur le positionnement nucléosomal sont nombreux dans la littérature (Chen et al., 2017a,b; Field et al., 2009; Tsankov et al., 2010). L'effet inverse a également été mis en évidence, à savoir une influence du nucléosome sur les patrons de mutations (Chen et al., 2012; Morganella et al., 2016; Wu et al., 2018). L'ensemble de ces influences mutuelles a fait l'objet d'une revue bibliographique publiée durant cette thèse (Barbier et al., 2021). Aux bords des barrières nucléosomales, l'évolution des séquences semble sous pression de selection en relation avec le positionnement nucléosomal. En effet, des signes de sélection de certaines mutations ont été détectés aux bords des NIEBs, selon qu'on se place dans les séquences nucléosomales ou inter-nucléosomales. Ces effets de sélection ont été étudiés en détail par Drillon et al. (Drillon et al., 2016), et décrits dans l'introduction de ce manuscrit (**Partie 1.1.4**). Chez l'humain les mutations vers les nucléotides G et C sont favorisées dans les séquences nucléosomales, alors que celles vers les nucléotides A et T sont favorisées dans les séquences inter-nucléosomales (donc les linker et les NIEBs). Ces patrons de mutations semblent donc renforcer l'oscillation du taux de GC que l'on observe aux bords des NIEBs (**Partie 2.3.1**). Ainsi, l'écriture du nucléosome dans la séquence génomique au niveau des NIEBs aurait été, chez l'humain, sélectionnée durant l'évolution. Dans cette Partie, l'étude des mutations ponctuelles sera étendue à une autre espèce proche de l'humain, le chimpanzé, dont les barrières nucléosomales sont très conservées par rapport à celles de l'humain (**Partie 2.4**). Il s'agira de voir si les patrons de mutations renforçant l'oscillation de la composition en GC aux bords des NIEBs chez l'humain sont observés chez le chimpanzé. Dans un second temps, nous étudierons les mutations des dinucléotides CpG, qui n'ont jusqu'ici jamais été étudiés aux bords des NIEBs.

2.5.1 Les mutations ponctuelles renforcent l'oscillation de la composition en GC chez le chimpanzé

Pour obtenir les taux de mutations ponctuelles, il est nécessaire de comparer les séquences de plusieurs espèces base par base afin de déterminer quelles bases sont identiques dans toutes les espèces (les bases non-mutées), et quelles bases ont été modifiées dans une ou plusieurs espèces. Dans certains cas, il est possible de conclure à une mutation spécifique à une espèce ou à un groupe d'espèce. Les différents cas de figures seront précisés dans le paragraphe suivant. Pour étudier les mutations aux bords des NIEBs de l'humain et du chimpanzé, nous avons comparé les séquences de ces deux espèces entre elles et à la séquence déterminée comme ancestrale, obtenue en utilisant deux groupes externes : le gorille (*Gorilla Gorilla*, version gorGor4) et l'orang-outan (*Pongo Abelii*, version ponAbe2). Pour effectuer ces comparaisons, il est donc nécessaire d'aligner les quatres génomes entre eux, afin d'obtenir des régions où les séquences sont comparables base par base. Cela a été réalisé en utilisant le même type de données que dans la Partie précédente sur la conservation des NIEBs entre l'humain et le chimpanzé (**Partie 2.4**). À partir des chaînes d'alignement disponibles dans la base de données de l'UCSC, j'ai pu produire des jeux de données d'intervalles alignés entre l'humain et le gorille et entre l'humain et l'orang-outan. J'ai, pour ça, utilisé le même protocole que celui décrit pour les intervalles humain-chimpanzé (**Partie 2.4**). A l'issue de cette étape, on a donc trois jeux de données d'intervalles correspondant aux zones alignées entre le génome humain et les trois autres génomes. La comparaison de ces jeux de données entre eux a permis d'extraire les intervalles communs aux trois alignements, pour mettre au point un nouveau jeu de données, contenant les zones alignées dans les quatre espèces considérées ici.

À partir de ces alignements, la comparaison des séquences base par base a pu être effectuée pour déterminer les mutations ponctuelles spécifiques à l'humain et au chimpanzé. Pour détecter ces mutations, il est nécessaire d'identifier, pour chaque locus, la base ancestrale de l'humain et du chimpanzé, pour la comparer à la base correspondante dans les deux espèces. Pour un locus donné, deux cas sont alors possibles :

A) Il n'y a pas de correspondance entre les bases des deux groupes externes. On considère alors qu'on ne peut pas déterminer la base ancestrale. Dans ce cas, il est donc impossible d'étudier les mutations.

B) Les bases des deux groupes externes sont identiques. On identifie alors la base en question comme étant la base ancestrale, que l'on compare à la base correspondante chez l'humain et le chimpanzé. Quatre sous-cas sont alors possibles :

   1 Les bases de l'humain et du chimpanzé correspondent toutes deux à la base ancestrale. On considère alors qu'il n'y a eu de mutation ni chez l'un ni chez l'autre.

   2 La base de l'humain correspond à la base ancestrale mais pas celle du chimpanzé. On considère alors qu'il y a eu une mutation chez le chimpanzé.

   3 La base du chimpanzé correspond à la base ancestrale mais pas celle de l'humain. On considère alors qu'il y a eu une mutation chez l'humain.

   4 La base de l'humain et celle du chimpanzé sont différentes de la base ancestrale. On considère alors qu'on ne peut pas conclure à une mutation spécifique de l'une ou l'autre des deux espèces. On note qu'il s'agit ici d'un choix assez conservatif, car on pourrait,

également conclure soit à une mutation commune aux deux espèces (si la base humaine est identique à celle du chimpanzé) soit à deux mutations spécifiques différentes (si les bases humaine et du chimpanzé sont différentes).

Déterminer les bases ancestrales (lorsque c'était possible) a permis de mettre au point un "génome humain ancestral" et un "génome chimpanzé ancestral", de mêmes dimensions que les deux génomes correspondant, en remplaçant, au sein de ces génomes, les bases de l'humain (et du chimpanzé) par les bases ancestrales lorsqu'elles étaient déterminées. Les bases dont l'état ancestral n'était pas connu (que ce soit car les groupes externes ne se correspondaient pas à ce locus ou parce que le locus ne faisait pas partie des zones alignées entre les quatre espèces) ont été remplacées par des "N". Le mise au point des génomes ancestraux a alors permis d'obtenir facilement les taux de mutations ponctuelles aux bords des NIEBs, à la fois le taux global de mutations mais également les taux de chaque type de mutations. Ces taux sont calculés en évaluant la distance entre chaque mutation et son bord de NIEB le plus proche. Cela permet d'obtenir un comptage, pour chaque distance aux bords des NIEBs, de chacun des 12 types de mutations possibles. Pour transformer ces comptages en taux, il suffit alors de diviser ces comptages par le comptage des bases ancestrales correspondant au type de mutation. Par exemple, pour les mutations de A vers C (ou G ou T), le taux de mutation à une certaine distance du NIEB correspond au nombre de mutations A vers C (ou G ou T) observées à cette distance, divisé par le nombre de A ancestraux observés à la même distance d'un NIEB. Ainsi, les taux de mutations en fonction de la distance au NIEB ont pu être calculés pour tous les types de mutations. Le taux de mutations global en fonction de la distance au NIEB a également été déterminé en prenant le nombre total de mutations observées à chaque distances (indépendamment du type de mutation) et en le divisant par le nombre de sites totaux considérés (correspondant au nombre de fois où une base ancestrale a pu être déterminée à cette distance d'un NIEB).

Pour cette analyse, on a également choisi de séparer les séquences en les sites ancestraux de type CpG et non-CpG, puis de focaliser dans un premier temps l'analyse sur les sites non-CpG. En effet, les sites CpG sont associés à des taux de mutations très supérieurs à la moyenne génomique, notamment à cause de la désamination de la cytosine méthylées (Bird, 1980; Coulondre et al., 1978; Moore et al., 2013). En raison du caractère spécifique du patron de mutation de ces CpG, ils sont habituellement analysés indépendamment du reste des sites, à l'image de ce qui a été fait dans la publication de Drillon et al. qui ne les considèrent pas (Drillon et al., 2016). L'analyse des substitutions en contexte CpG est présentée en **Partie 2.5.2**. Les taux de chaque type de mutation non-CpG en fonction de la distance aux bords des NIEBs ont ainsi été calculé chez l'humain et chez le chimpanzé. Le calcul chez l'humain a été fait pour valider notre méthode en comparant ces résultats à ceux déjà publiés par Drillon et al. (Drillon et al., 2016), dans lesquels des signes de sélection de certains patrons de mutations ont été retrouvés aux bords des NIEBs. Les calculs chez le chimpanzé n'avaient jamais été fait jusqu'ici, et il s'agissait alors de voir si les observations effectuées chez l'humain sont également valables dans une autre espèce proche. Les résultats obtenus chez l'humain sont présentés dans la **Figure 2.14**, ceux chez le chimpanzé sont présentés dans la **Figure 2.15**.

Les taux de substitutions aux bords des NIEBs obtenus chez l'humain reproduisent fidèlement ceux précédemment publiés dans cette même espèce (Drillon et al., 2016, Figure 4) (**Figure 2.14**), ce qui valide notre protocole d'analyse. On retrouve bien un taux de substitutions global oscillant

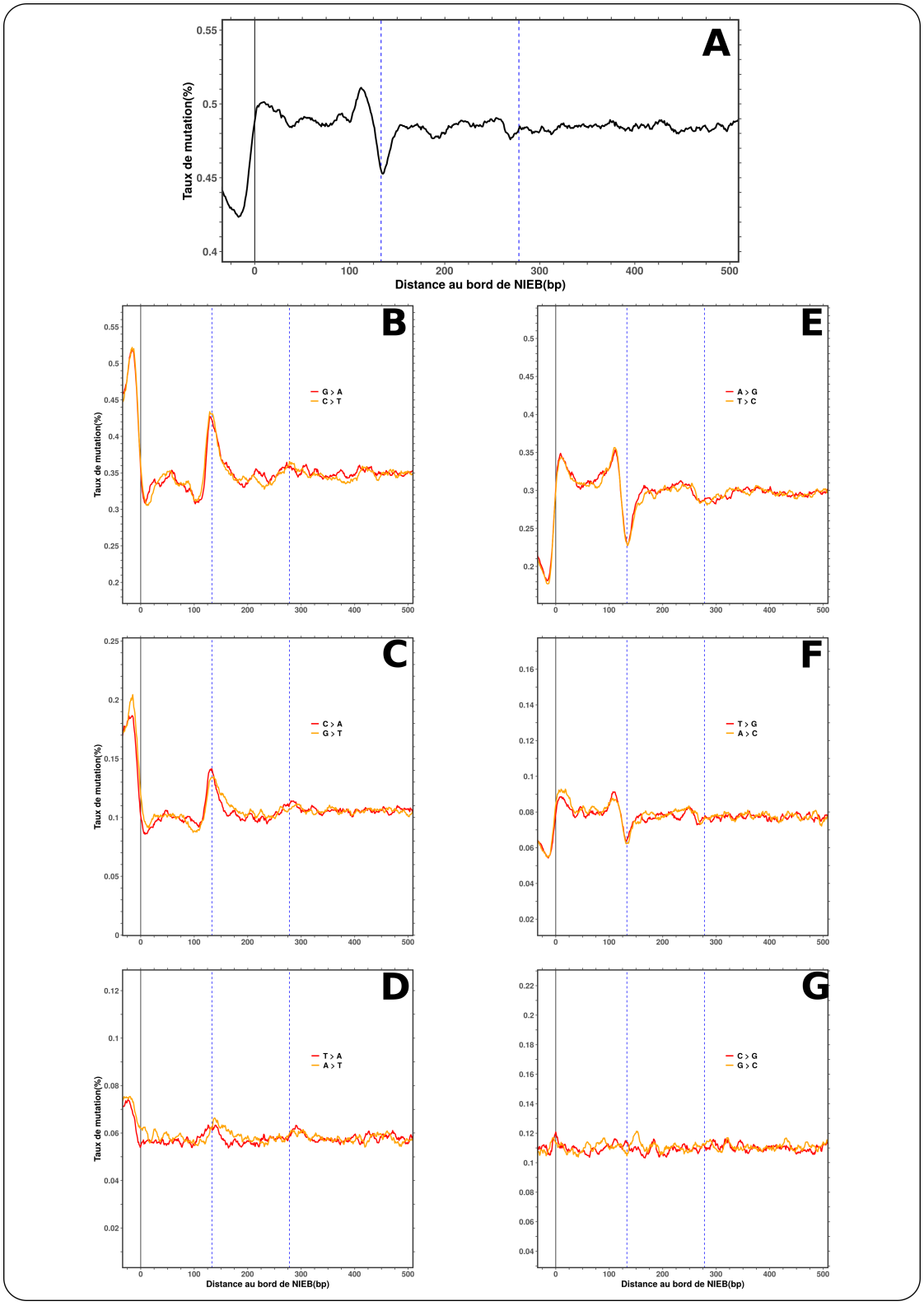


FIGURE 2.14 – **Taux de mutations aux bords des barrières nucléosomales chez l'humain.** Les substitutions complémentaires sont représentées sur un même graphique. **A** Taux de mutations total. **B** Taux de substitutions G>A (rouge) et C>T (orange). **C** Taux de substitutions C>A (rouge) et G>T (orange). **D** Taux de substitutions T>A (rouge) et A>T (orange). **E** Taux de substitutions A>G (rouge) et T>C (orange). **F** Taux de substitutions T>G (rouge) et A>C (orange). **G** Taux de substitutions C>G (rouge) et G>C (orange). La ligne verticale noire correspond à l'abscisse 0 pb. Les lignes verticales pointillées bleues correspondent aux abscisses 133 pb et 278 pb (minima du %GC aux bords des NIEBs de l'humain). Les courbes sont lissées sur 10 pb, et représentent les moyennes des taux obtenus à droite des NIEBs (bord 3') et du complément inverse des substitutions à gauche des NIEBs (bord 5') pour les 1 745 801 NIEBs humains.

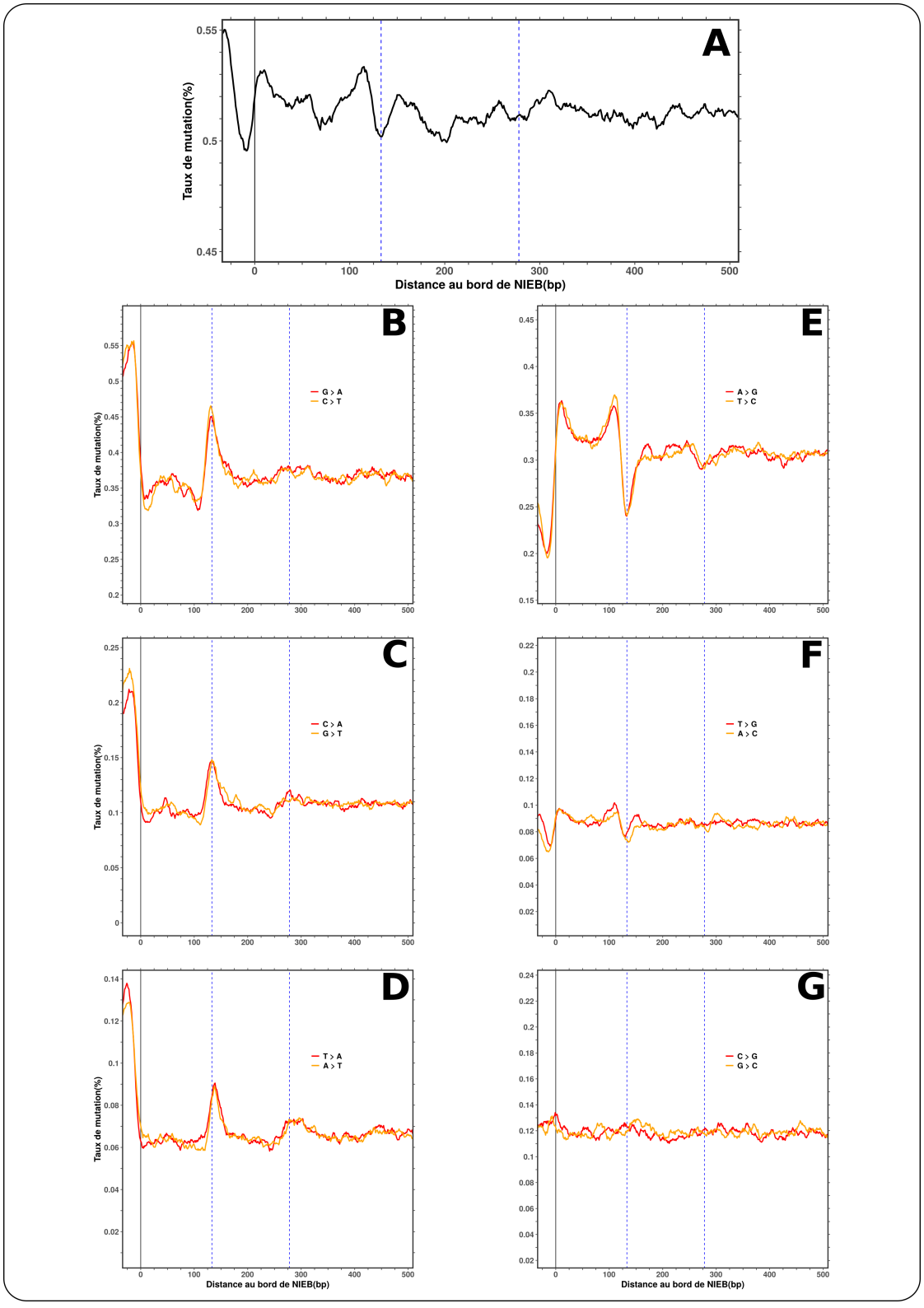


FIGURE 2.15 – **Taux de mutations aux bords des barrières nucléosomales chez le chimpanzé.** Les substitutions complémentaires sont représentées sur un même graphique. **A** Taux de mutations total. **B** Taux de substitutions G>A (rouge) et C>T (orange). **C** Taux de substitutions C>A (rouge) et G>T (orange). **D** Taux de substitutions T>A (rouge) et A>T (orange). **E** Taux de substitutions A>G (rouge) et T>C (orange). **F** Taux de substitutions T>G (rouge) et A>C (orange). **G** Taux de substitutions C>G (rouge) et G>C (orange). La ligne verticale noire correspond à l'abscisse 0 pb. Les lignes verticales pointillées bleues correspondent aux abscisses 133 pb et 278 pb (minima du %GC aux bords des NIEBs de l'humain). Les courbes sont lissées à 10 pb, et représentent les moyennes des taux obtenus à droite des NIEBs (bord 3') et du complément inverse des substitutions à gauche des NIEBs (bord 5') pour les 1 733 364 NIEBs chimpanzé.

entre valeurs basses aux séquences inter-nucléosomales (jusqu'à 0.42 % dans les NIEBs et 0.46 % dans le premier linker) et valeurs hautes dans les séquences nucléosomales (un plateau à 0.49 % au niveau du premier nucléosome avec un pic à 0.51 % et un plateau à 0.48 % au niveau du second nucléosome) (**Figure 2.14 - A**). De plus, on voit que les taux de substitutions vers les bases A et T sont plus importants dans les séquences inter-nucléosomales que dans les séquences nucléosomales (**Figure 2.14 - B à D**), et qu'à l'inverse, ceux vers les bases C et G sont plus importants dans les séquences nucléosomales qu'entre les nucléosomes (**Figure 2.14 - E à G**). On note cependant que les substitutions de types C vers G et G vers C ne montrent pas de préférences pour les nucléosomes ou les linkers/NIEBs. Cela peut s'expliquer par le fait que ce type de substitution ne modifie pas le taux de GC, et n'aurait donc pas de réel impact sur l'encodage du nucléosome dans la séquence. Les mutations T vers A et A vers T, dont on pourrait alors s'attendre à ce qu'elle ne montrent pas de préférences non plus, semblent elles au contraire légèrement favorisées dans les NIEBs et linkers. Une substitution de T vers A ou de A vers T qui permettrait la formation d'un polyA ou d'un polyT, ou l'allongement d'une telle séquence pourrait, elle impacter plus fortement le positionnement nucléosomal, en raison du fort pouvoir inhibiteur des séquences de type polyA et polyT (Drillon et al., 2016).

Le taux de substitutions moyen est légèrement supérieur dans la lignée chimpanzé que dans la lignée humaine. En effet, à 500 pb des NIEBs, où on suppose que le taux est représentatif de la moyenne génomique, il est de 0.51 chez le chimpanzé contre 0.49 chez l'humain. Néanmoins, la dépendance des taux de substitutions avec la distance aux bords des NIEBs observée chez l'humain est également retrouvée chez le chimpanzé, même si l'on note une différence principale au niveau du bord interne des NIEBs (**Figure 2.15**). On retrouve chez le chimpanzé l'augmentation du taux de substitution global au niveau des séquences nucléosomales par rapport au linker (avec une oscillation autour de 0.52 % au niveau du premier nucléosome contre un taux de 0.5 % au niveau du premier linker) (**Figure 2.15 - A**). En revanche, à l'intérieur des NIEBs, les taux de substitutions sont différents entre les deux espèces, la forte diminution observée chez l'humain étant remplacée par une diminution plus faible (jusqu'à 0.49 %) suivie d'une forte augmentation chez le chimpanzé, avec même un taux de substitution maximal de 0.55 % retrouvé à l'intérieur des NIEBs dans cette espèce. La forme des courbes pour les taux obtenus substitutions par substitutions sont conservées pour chacun des types de substitutions (**Figure 2.15 - B à G**). On remarque cependant une claire différence au niveau des mutations de type T vers A et A vers T, qui explique la différence de taux de substitutions global. En effet, chez l'humain, ce type de substitutions est légèrement favorisé à l'intérieur des NIEBs et dans le premier linker. Cependant, la préférence est faible, avec des valeurs passant de 0.055 % à 0.07 % entre l'intérieur des NIEBs et le premier nucléosome (**Figure 2.15 - D**). Chez le chimpanzé, la préférence est beaucoup plus forte, avec des valeurs de 0.06 % dans le premier nucléosome contre ~ 0.14 % au bord interne des NIEBs. Il semble donc que ce type de substitutions soit beaucoup plus favorisé au niveau des séquences inter-nucléosomales du chimpanzé par rapport aux mêmes séquences chez l'humain, au point d'avoir un impact important sur le taux de substitutions global.

Avant de chercher une explication biologique à cette différence, il est cependant nécessaire de se pencher sur les différences entre les deux génomes de référence. En effet, le génome de référence humain, dans sa version hg38, a été obtenu à partir du séquençage de plusieurs individus. De plus, des projets de séquençage de plusieurs milliers de génomes humains (comme le 1000

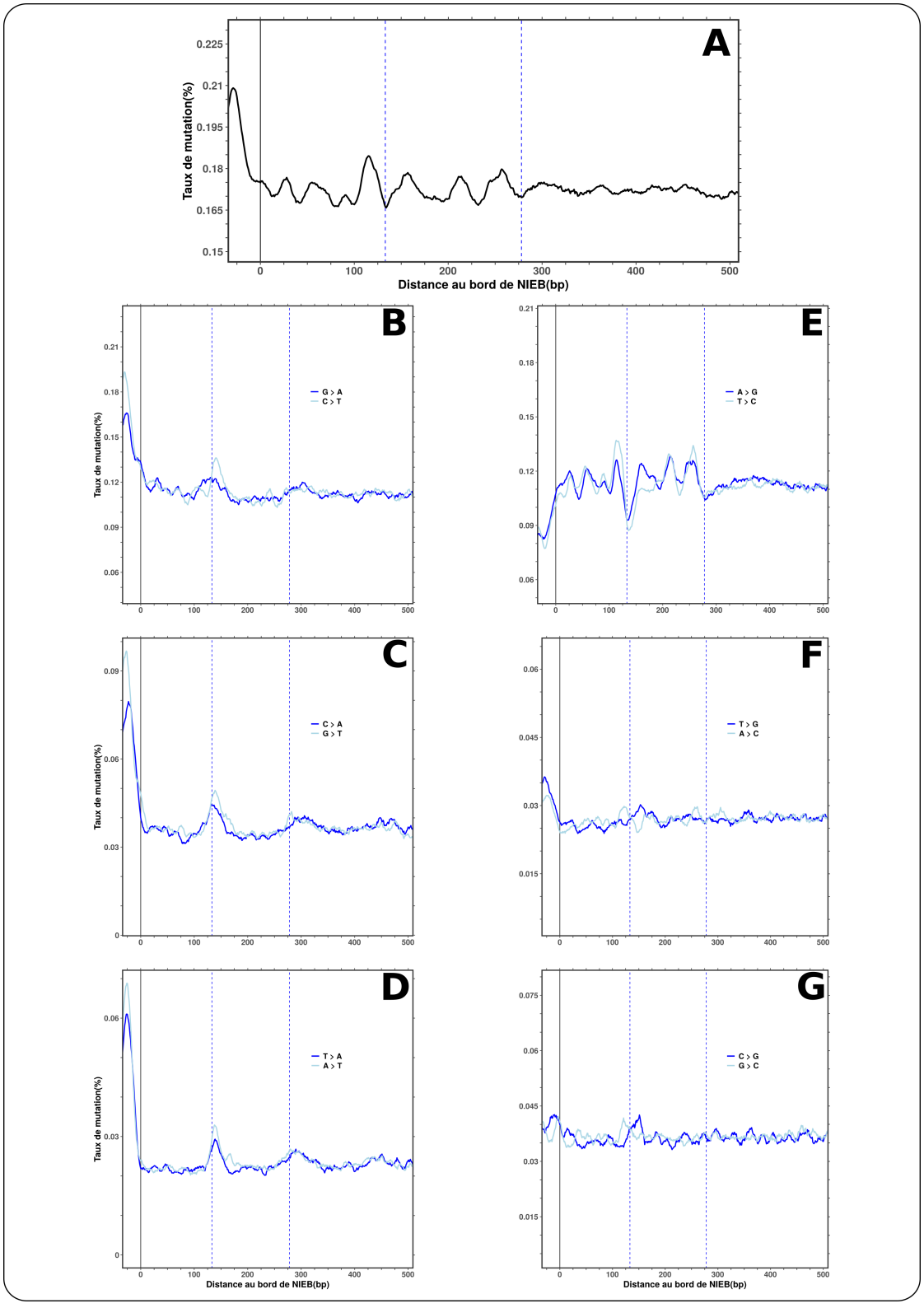


FIGURE 2.16 – **Taux de SNPs aux bords des barrières nucléosomales chez l'humain.** Les substitutions complémentaires sont représentées sur un même graphique. **A** Taux de mutations total. **B** Taux de substitutions G>A (bleu) et C>T (bleu clair). **C** Taux de substitutions C>A (bleu) et G>T (bleu clair). **D** Taux de substitutions T>A (bleu) et A>T (bleu clair). **E** Taux de substitutions A>G (bleu) et T>C (bleu clair). **F** Taux de substitutions T>G (bleu) et A>C (bleu clair). **G** Taux de substitutions C>G (bleu) et G>C (bleu clair). La ligne verticale noire correspond à l'abscisse 0 pb. Les lignes verticales pointillées bleues correspondent aux abscisses 133 pb et 278 pb (minima du %GC aux bords des NIEBs de l'humain). Les courbes sont lissées sur 10 pb, et représentent les moyennes des taux obtenus à droite des NIEBs (bord 3') et du complément inverse des substitutions à gauche des NIEBs (bord 5') pour les 1 745 801 NIEBs humains.

Genome Project, Durbin et al. (2010)) ont permis de mettre en évidence un polymorphisme de substitutions élevé (SNP pour Single Nucleotide Polymorphism). Ce fort polymorphisme a été pris en compte dans l'établissement de la version hg38 du génome de référence humain, qui en contient beaucoup moins que les versions précédentes (Church et al., 2015). A l'inverse, le génome de référence du chimpanzé a été obtenu à partir du séquençage d'un seul individu. Ainsi, dans ce génome de référence, on retrouve les mutations communes à l'ensemble des individus de l'espèce *Pan troglodytes*, mais également celles spécifiques à l'individu séquencé ou à la sous-population à laquelle appartient cet individu. Pour simplifier, on peut donc dire que le génome de référence humain contient assez peu de SNPs, celui-ci ayant été nettoyé à l'aide de projets de séquençage massif. A l'inverse, le génome du chimpanzé contient lui potentiellement beaucoup de SNPs, ce qui peut perturber l'analyse des taux de substitutions dans ce génome.

Lors de la comparaison des taux de substitutions aux bords des NIEBs obtenus chez l'humain et le chimpanzé, il est donc nécessaire de prendre en compte l'éventuel effet des SNPs, particulièrement chez le chimpanzé. Malheureusement, à ma connaissance, pour cette espèce, aucun projet de séquençage de centaines voire de milliers d'individus n'a été mis en place. Il n'est donc pas possible de détecter et prendre en compte les SNPs dans notre analyse. En revanche, il est possible d'analyser les taux de substitutions de type SNP chez l'humain grâce aux données du projet 1000 Genomes. Cela a été réalisé par Drillon et al. (Drillon et al., 2016, Figure 4), et reproduit ici (**Figure 2.16**).On observe que le taux des différents types de SNPs aux bords des NIEBs pourrait expliquer les différences observées entre l'humain et le chimpanzé. En effet, le taux global de SNPs humain augmente à l'intérieur des NIEBs, tout comme le taux de substitutions global observé chez le chimpanzé (comprenant donc les SNPs et les variations inhérentes à l'espèce). De plus, cette augmentation semble, à l'image de celle observée chez le chimpanzé, être provoquée par les substitutions de type A vers T et T vers A, dont l'augmentation à l'intérieur des NIEBs est bien plus importante pour les SNPs (de 0.02 % à 0.07 %) que pour les mutations humain-chimpanzé (0.055 % à 0.07 %). Les autres taux de substitutions sont similaires entre les mutations humain-chimpanzé et les SNPs chez l'humain. On retrouve donc la même différence entre les taux de substitution lorsqu'on compare les SNPs humains aux mutations spécifiques de l'espèce humaine que lorsqu'on compare les mutations du chimpanzé à celles de l'espèce humaine. La différence dans les taux de substitutions aux bords des NIEBs dans ces deux espèces semble donc s'expliquer par la présence de SNPs dans le génome de référence du chimpanzé. Néanmoins, pour confirmer cette hypothèse, il faudrait pouvoir accéder à un jeu de données de SNPs chez le chimpanzé. Cela permettrait de voir si on retrouve bien les mêmes taux de substitutions que dans les SNPs humains, et également de "nettoyer" le génome de référence du chimpanzé de ses SNPs, pour voir si cela supprime la différence observée du taux de substitutions à l'intérieur des NIEBs dans cette espèce par rapport au taux humain.

De manière générale, les résultats présentés dans cette Partie indiquent que les patrons de mutations observés aux bords des NIEBs chez l'humain sont également présents chez le chimpanzé. En effet, la seule différence se situe à l'intérieur des NIEBs, et peut s'expliquer par la présence de SNPs dans le génome de référence du chimpanzé là où ils ont été retirés du génome de référence humain. Ainsi, l'écriture du positionnement nucléosomal dans la séquence génomique aux bords des NIEBs du chimpanzé semble être renforcée par les substitutions. Chez l'humain, la comparaison des taux de substitutions de type SNP et non-SNP avait mis en évidence que des effets de sélection

étaient en cause, favorisant les mutations vers G et C au niveau des séquences nucléosomales et les mutations vers A et T au niveau des séquences inter-nucléosomales. On peut supposer qu'un mécanisme similaire est à l'œuvre chez le chimpanzé, même si l'absence de données de SNP dans cette espèce ne nous permet pas de confirmer cette hypothèse.

2.5.2 Les mutations en contexte CpG confirment le positionnement de 2-3 nucléosomes aux bords des barrières

L'étude des mutations ponctuelles aux bords des NIEBs chez l'humain a mis en évidence des effets de sélection renforçant le positionnement nucléosomal à ces loci (Drillon et al., 2016). Les résultats présentés dans la Partie précédente suggèrent fortement que des effets similaires sont présents chez le chimpanzé. Dans ces analyses, les substitutions de type C vers T/A/G survenues dans un contexte CpG n'ont jusqu'ici pas été prises en compte, alors que leur taux est bien plus élevé que les autres substitutions (Bird, 1980; Coulondre et al., 1978; Moore et al., 2013). En effet, en contexte CpG, la méthylation des cytosines en 5-méthylcytosines suivie de leur désamination spontanée amène à une substitution de la cytosine par une thymine. La fréquence de cette substitution domine largement celles des autres substitutions possibles dans ce contexte (**Figure A.5**). Elle est si fréquente qu'on observe, dans le génome humain, un nombre de dinucléotides CpG largement en deçà de celui attendu selon la composition génomique en GC (~ 5 fois moins de CpG observés qu'attendus). Cette sous-représentation des dinucléotide CpG est d'ailleurs commune aux mammifères (Moore et al., 2013). Les substitutions C vers T en contexte CpG ne sont pas indépendantes de la présence/absence de nucléosomes. En effet, la désamination spontanée de la 5-méthylcytosine est plus fréquente dans les séquences inter-nucléosomales que dans les séquences nucléosomales (Gonzalez-Perez et al., 2019). Cela est dû au fait que cette désamination nécessite une ouverture locale de la double-hélice d'ADN, qui est plus fréquente pour l'ADN libre que pour l'ADN nucléosomal, car ce dernier est plus contraint structurellement par son enroulement autour des protéines histones (Makova & Hardison, 2015). Ainsi, les mutations de type C vers T en contexte CpG sont sur-représentées dans l'ADN inter-nucléosomal par rapport à l'ADN nucléosomal. Aux bords des NIEBs, où l'on a prédit un positionnement fort de deux à trois nucléosomes, on s'attend donc à ce que ces substitutions ne soient pas distribuées de manière homogène, mais soient plutôt favorisées dans les séquences inter-nucléosomales (à savoir le NIEB et le linker). La distribution du taux de mutations C vers T aux bords des NIEBs chez l'humain et le chimpanzé est présentée en **Figure 2.17**. Elle a été obtenue de la même manière que les taux de substitutions présentés dans la Partie précédente, simplement en se concentrant ici sur les mutations de la cytosine des sites CpG ancestraux. Également, n'est présentée dans la **Figure 2.17** que la moyenne des taux de substitutions complémentaires C vers T et G vers A. En effet, en contexte CpG, ce type de substitutions est largement majoritaire. Les taux des autres types de substitutions ont été calculés et l'ensemble des taux de substitutions sont présentés en **Figure A.5**. On peut y voir que les mutations C vers T et G vers A sont en effet largement majoritaires, avec des taux dix à vingt fois plus importants que les autres types de substitutions (4 à 8 % pour C vers T et G vers A contre 0.2 à 0.6 % pour les deux autres types).

Les taux de substitutions C vers T et G vers A en contexte CpG sont en accord avec le positionnement nucléosomal intrinsèque observé aux bords des NIEBs (**Figure 2.17**). En effet, on observe

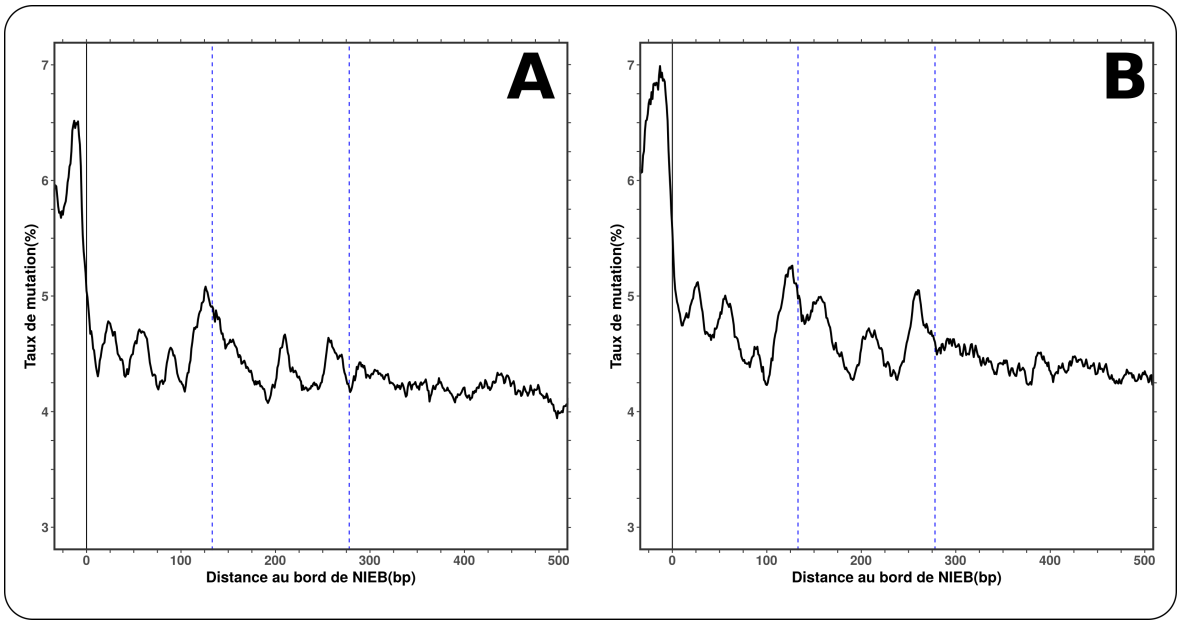


FIGURE 2.17 – **Taux de substitutions C vers T et G vers A en contexte CpG aux bords des barrières nucléosomales chez l'humain et le chimpanzé.** La courbe correspond à la moyenne des deux taux de substitutions complémentaires (voir **Figure A.5 - A et D** pour les taux individuels) **A** Taux de substitutions chez l'humain. **B** Taux de substitutions chez le chimpanzé. La ligne verticale noire correspond à l'abscisse 0 pb. Les lignes verticales pointillées bleues correspondent aux abscisses 133 pb et 278 pb (minima du %GC aux bords des NIEBs de l'humain). Les courbes sont lissées sur 10 pb, et représentent les moyennes des taux obtenus à droite des NIEBs (bord 3') et du complémentaire inverse des substitutions à gauche des NIEBs (bord 5') pour les 1 745 801 NIEBs humains et les 1 733 364 NIEBs chimpanzé.

un taux de substitution bien plus élevé au niveau des séquences inter-nucléosomales que dans les séquences nucléosomales. Aux bords internes des NIEBs les taux montent jusqu'à 6.5 % et 7 % respectivement chez l'humain et le chimpanzé. On observe ensuite une diminution au niveau du premier nucléosome aux bords des NIEBs (à 4.25 et 4.5 % chez l'humain/le chimpanzé). Dans le premier linker (première ligne pointillée bleue), le taux remonte (à 5 et 5.5 %). Ensuite, les résultats diffèrent légèrement entre les deux espèces. Chez le chimpanzé, on observe une nouvelle diminution au niveau du second nucléosome (à 4.75 %) et un pic assez élevé (à 5.25 %) très légèrement en amont du second linker (deuxième ligne pointillée bleue). Après le second linker, la courbe forme un plateau bas à 4.5 %. Chez l'humain, la diminution au niveau du second nucléosome n'est pas aussi claire, avec un pic secondaire à 200 pb aussi haut que celui légèrement en amont du second linker (tous deux à 4.75 %). On retrouve néanmoins, en aval du second linker, le plateau bas aux alentours de 4.25 %. On remarque également, dans les deux espèces, plusieurs pics intermédiaires au niveau des séquences nucléosomales (aux abscisses 20 pb, 50 pb, 80 pb et 200 pb). Ces pics sont dus à la présence d'éléments transposables Alu, qui contiennent des sites CpG. En effet, si on retire de l'analyse les NIEBs bordés par ces éléments, ces pics intermédiaires disparaissent (**Figure A.6**).

Les substitutions C vers T et G vers A montrent donc une préférence pour les séquences ayant été identifiées comme inter-nucléosomales. Deux hypothèses peuvent expliquer cette préférence. La première est liée au taux de mutations hétérogène des CpG selon la présence/absence de nucléosome, à cause du mécanisme mutationnel gêné par les contraintes structurelles imposées par l'enroulement de l'ADN autour des histones (Makova & Hardison, 2015). Dans cette hypothèse, c'est donc la présence de nucléosomes bien positionnés aux bords des NIEBs qui influence le

taux de substitution des cytosines en contexte CpG. On note que la présence de nucléosome peut également influencer le taux de méthylation (Collings & Anderson, 2017) ainsi que l'accessibilité à l'ADN pour les protéines de réparation (Barbier et al., 2021), ce qui pourrait également avoir une influence sur le taux de mutations. Les effets du positionnement nucléosomal sur la méthylation aux bords des NIEBs pourraient d'ailleurs être étudiés spécifiquement en analysant des données de séquençage de type BS-seq, MethylC-seq ou NOMe-seq (Yong et al., 2016). La seconde hypothèse est liée à ce qui a été montré dans la Partie précédente (**Partie 2.5.1**), où on a pu voir des patrons de mutations biaisés vers A et T dans les séquences inter-nucléosomales, que Drillon et al. ont identifiés comme des signes de sélection de ces mutations à ces loci (Drillon et al., 2016). En effet, à l'image des autres types de substitutions, les mutations C vers T en contexte CpG pourraient être sélectionnées au niveau des séquences inter-nucléosomales pour renforcer l'oscillation du taux de GC aux bords des NIEBs. On note que ces deux hypothèses ne sont pas mutuellement exclusives. En effet, il est tout à fait possible que le taux de mutations des bases C en contexte CpG soit biaisé pré-sélection aux bords des NIEBs en lien avec le positionnement nucléosomal, et que ce biais soit amplifié au cours de l'évolution par sélection des mutations vers T dans les séquences inter-nucléosomales. Il reste donc à déterminer dans quelle mesure chacun des deux effets contribue au résultat observé pour les taux de mutations des cytosines en contexte CpG. Pour ça, il sera nécessaire d'évaluer les effets de sélection notamment en comparant les résultats obtenus avec les mutations humain-chimpanzé à ceux obtenus avec les SNP humains, à l'image de ce qui a été fait pour les mutations non-CpG (Drillon et al., 2016). L'influence du nucléosome sur le taux de mutations pré-sélection pourrait aussi être évalué en analysant des jeux de données de mutations somatiques, ou mieux encore, *de novo*. Cela pourrait permettre de s'affranchir (au moins en partie) des effets de sélection, pour déterminer dans quelle mesure les mutations C vers T en contexte CpG sont influencées par la présence de nucléosomes.

Quelle que soit la (ou les) raison(s) du biais de mutations C vers T en contexte CpG aux bords des NIEBs, celui-ci est inégalement réparti entre séquences nucléosomales et inter-nucléosomales. Cette information pourrait être utilisée pour confirmer les prédictions de positionnement des nucléosomes aux bords des NIEBs, particulièrement dans les espèces où les données expérimentales de positionnement nucléosomal ne sont pas disponibles, comme ici chez le chimpanzé. Ainsi, les résultats présentés dans cette Partie confirment la présence et le positionnement prédit des nucléosomes aux bords des NIEBs du chimpanzé.

3 Les barrières nucléosomales, un point d'entrée pour les modifications épigénétiques ?

Sommaire

3.1 Introduction

Dans le **Chapitre 2**, j'ai montré que les barrières nucléosomales ont été retrouvées par notre modèle physique du positionnement de nucléosomes dans diverses espèces représentatives des eucaryotes. Un certain nombre de caractéristiques sont partagées entre les barrières nucléosomales des différentes espèces, et les barrières elles-mêmes sont très conservées entre espèces proches, comme en témoignent les comparaisons entre l'humain et le chimpanzé. Toutes ces analyses reposent cependant sur les prédictions d'un modèle de positionnement des nucléosomes, calibré à partir de données expérimentales chez la levure (Vaillant et al., 2007). Ce modèle a été validé chez l'humain par l'analyse de données expérimentales de type MNase-seq (Drillon et al., 2016). Cependant, les validations expérimentales des prédictions du modèle dans les autres espèces étudiées dans le **Chapitre 2** n'ont pas encore été effectuées, principalement par manque des jeux de données expérimentales de positionnement de nucléosomes correspondant. Un des travaux de ma thèse a consisté à faire un recensement des jeux de données disponibles, et essayer d'étendre la validation du modèle à d'autres espèces. J'ai pu, pour ce faire, m'appuyer sur l'inventaire des données de type MNase-seq disponibles effectué par V. Teif (Teif, 2016). Ce chapitre présente les résultats obtenus par l'analyse des données expérimentales de positionnement de nucléosomes dans diverses espèces, et ce que ces analyses ont apporté en terme de compréhension de l'organisation de la chromatine au niveau des barrières nucléosomales.

3.2 Les profils expérimentaux de positionnement du nucléosome corroborent l'existence de barrières nucléosomales dans plusieurs génomes eucaryotes

Pour retrouver expérimentalement la position des nucléosomes dans les génomes, on utilise classiquement des données de MNase-seq (Schones et al., 2008; Tolstorukov et al., 2010; Valouev et al., 2011; Yuan et al., 2005). Dans l'expérience de MNase-seq, la chromatine est digérée par une enzyme, la micrococcal nucléase (MNase), dont l'activité consiste à cliver et digérer l'ADN. Ce clivage et cette digestion ne sont possibles que si l'ADN est nu, aussi les fragments enroulés autour des protéines histones pour former les nucléosomes sont protégés de la digestion. Ainsi, la digestion de la chromatine par la MNase suivie d'une élimination des protéines histones permet de récupérer les fragments d'ADN formant des nucléosomes. Le séquençage de ces fragments et leur alignement sur le génome correspondant permet alors d'obtenir la position de ces nucléosomes dans le génome. L'ADN peut être protégé de la digestion par la fixation d'autres protéines comme des facteurs de transcription, ce qui peut être confondu avec une protection par les histones. Les protocoles expérimentaux peuvent ainsi passer par une étape de sélection des fragments ne conservant que ceux ayant une taille compatible avec la formation d'un nucléosome (soit environ 150 pb).

Généralement, l'expérience de MNase-seq est effectuée sur une population de cellules, ce qui permet d'obtenir un signal moyen d'occupation des nucléosomes. Un pic à une position dans ce signal indique que pour un grand nombre de cellules, un nucléosome se trouve à cette position. C'est

ce qu'on identifie comme un nucléosome "positionné", dans le sens où les nucléosomes se formant dans cette zone dans différentes cellules occupent souvent la même position. A l'inverse, certaines régions présentent un signal plus faible que les autres, illustrant une déplétion en nucléosome. Ces régions sont identifiées comme des NDRs (pour Nucleosome Depleted Regions). Typiquement, dans notre cas, où l'on prédit des barrières inhibitrices de la formation du nucléosome, on observe que chez l'humain, les NIEBs correspondent à des NDRs dans les données expérimentales, *in vitro* et *in vivo* (Drillon et al., 2015, 2016). Aux bords de ces barrières nucléosomales, les prédictions obtenues à partir de notre modèle indiquent des pics dans l'occupation correspondant à des nucléosomes très positionnés, que l'on retrouve également dans les données expérimentales (Drillon et al., 2015, 2016). Cela indique qu'aux bords des barrières nucléosomales, en moyenne, les nucléosomes ont des positions préférentielles bien définies. L'analyse des données de type MNase-seq permet donc de valider notre modèle physique de positionnement de nucléosomes, à la fois pour confirmer que les NIEBs correspondent bien à des NDRs *in vivo*, mais également le positionnement des nucléosomes aux bords des NIEBs. Dans cette partie, je vais présenter les résultats obtenus par l'analyse de données MNase-seq dans 4 nouvelles espèces (souris, drosophile, poisson-zèbre, arabette), et voir dans quelle mesure ces résultats confirment la présence de NIEBs et le positionnement des nucléosomes aux bords de ces barrières intrinsèques. Je décrirai d'abord le pipeline mis au point pour faciliter l'analyse des données, puis les résultats obtenus avec ce pipeline dans les 4 espèces étudiées.

3.2.1 Un pipeline d'analyse de données expérimentales de positionnement de nucléosomes

Pour faciliter l'analyse des données expérimentales de positionnement de nucléosomes, j'ai mis au point un pipeline d'analyse de données MNase-seq. Le but de ce pipeline est d'automatiser autant que possible l'analyse des données, et également d'homogénéiser les paramètres des analyses pour qu'elles soient aussi comparables que possible. Le principe général du pipeline est résumé dans la **Figure 3.1**. Il peut être succinctement résumé à trois étapes :

1. L'alignement des lectures de séquençage sur le génome de référence.
2. L'extraction de cet alignement du signal de positionnement des nucléosomes.
3. L'analyse de ce signal au niveau des barrières nucléosomales.

3.2.1.1 Données expérimentales et alignements

Les données expérimentales utilisées ici correspondent à des lectures de séquençage, la plupart du temps de type Illumina (Bentley et al., 2008), parfois de type SOLiD [1] (McKernan et al., 2009; Valouev et al., 2008). Ces données de séquençage peuvent être de type single-end, c'est-à-dire que pour chaque fragment d'ADN purifié, seule une extrémité du fragment est séquencée, ou paired-end, c'est-à-dire que pour chaque fragment, les deux extrémités du fragment sont séquencées. Enfin, la taille des lectures est généralement courte, de quelques dizaines de bases. Un résumé des

1. Brochure de présentation du séquençage SOLiD : http://www.columbia.edu/cu/biology/courses/w3034/Dan/readings/SOLiD_System_Brochure.pdf

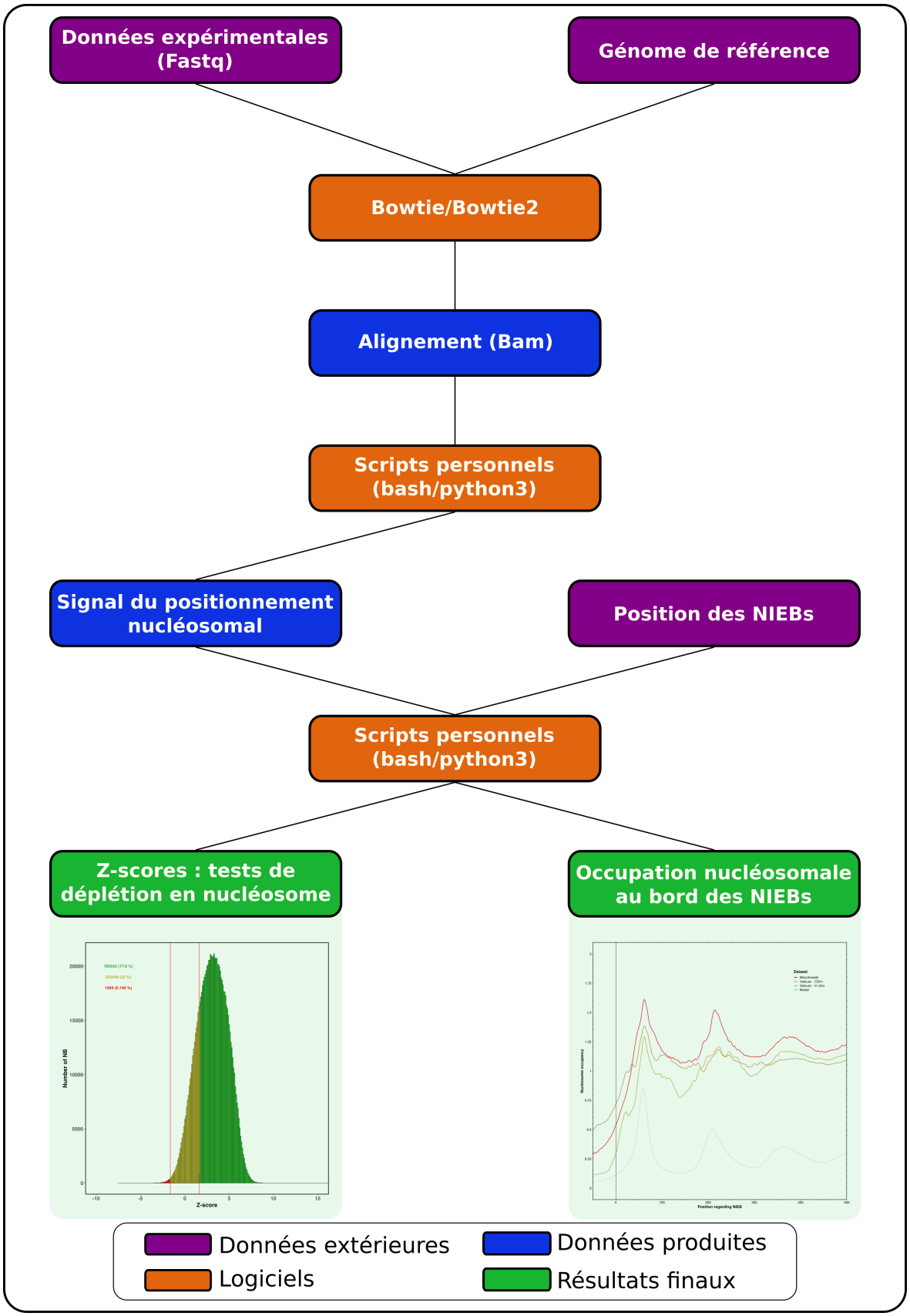


FIGURE 3.1 – **Schéma récapitulatif du pipeline d'analyse de données expérimentales de positionnement de nucléosomes.** Voir **Partie 3.2.1**.

Publication	Nom de l'expérience	Type de séquençage	Type de digestion	Nombre de lectures	$\overline{C}$
(Valouev et al., 2011)	Valouev *in vivo*	Single-end	Complète	388 350 928	26.4
(Valouev et al., 2011)	Valouev *in vitro*	Single-end	Complète	786 265 751	53.5
(Mieczkowski et al., 2016)	Mieczkowski	Paired-end	Complète	49 113 709	3.3

TABLEAU 3.1 – **Données expérimentales utilisées pour la validation du pipeline d'analyse de données de positionnement de nucléosome.** Les jeux de données Valouev *in vivo* et *in vitro* correspondent aux données obtenues sur les cellules CD4+ et par reconstitution *in vitro* de chromatine par Valouev et al. (Valouev et al., 2011). Le jeu de données Miezckowski correspond aux données obtenues avec les cellules K562 humaines, avec la plus grande quantité de MNase (304U). Le nombre de lectures représente le nombre de fragments restant après les étapes de nettoyage des données détaillées en 3.2.1.1. $\overline{C}$ représente le nombre moyen de lecture par nucléosome. Cette couverture est calculée en estimant le nombre de nucléosomes dans le génome à partir de la taille du génome et de la NRL (pour Nucleosome Repeat Length) moyenne rapportée dans la bibliographie (ici pour l'humain 200 pb, Valouev et al. (2011)). Le calcul est le suivant : $\overline{C} = \frac{n_{lectures}}{n_{nucl}}$ avec $n_{nucl} = \frac{size_{genome}}{\text{NRL}_{genome}}$.

jeux de données utilisées dans cette partie, avec le type de séquençage et le nombre de lectures alignées est disponible dans les **Tableaux 3.1** et **3.2**.

Concernant l'alignement, plusieurs logiciels sont capables d'aligner les lectures de séquençage sur un génome de référence. J'ai choisi d'utiliser les logiciels Bowtie (Langmead et al., 2009) et Bowtie2 (Langmead & Salzberg, 2012). J'aurais préféré utiliser le même logiciel d'alignement pour tous les jeux de données, en l'occurence Bowtie2, qui gère mieux les lectures pairées que Bowtie, mais ce dernier est le seul capable d'aligner des séquençages de type colorspace obtenus avec la technologie SOLiD. J'ai donc utilisé Bowtie2 pour tous les jeux de données de type Illumina, et Bowtie pour tous les jeux de données de type SOLiD. Dans les deux cas, j'ai utilisé ces logiciels en conservant les paramètres par défaut. Les génomes de références sont les mêmes que ceux utilisés dans le **Chapitre 2** (**Tableau 2.1**).

3.2.1.2 Traitement des alignements et extraction du signal de positionnement des nucléosomes

Les données de séquençage ont été téléchargées depuis l'archive publique SRA (Sequence Read Archive du NCBI, https ://www.ncbi.nlm.nih.gov/sra) sous forme de fichier fastq en utilisant les numéros d'accession donnés dans les publications originales (**Tableaux 3.1 et 3.2**). L'alignement des lectures produit des fichiers au format sam, que j'ai transformé en fichier au format bam avec le logiciel Samtools (Li et al., 2009), qui me permet également de produire un unique fichier d'alignements même lorsque les lectures sont réparties dans plusieurs fichiers fastq, ainsi que d'ordonner les alignements selon leur position sur le génome de référence. On applique alors un traitement aux fichiers bam, visant à "nettoyer" les alignements. Tout d'abord, lors d'un séquençage de type paired-end, les lectures qui ne sont pas correctement pairées (properly paired) sont éliminées. Cela arrive lorsque l'alignement des deux lectures associées à un fragment n'est pas compatible avec leur provenance d'un même fragment, par exemple, si la première lecture du fragment est alignée sur un chromosome différent de la seconde. Cela peut provenir d'une erreur d'alignement, d'une erreur de séquençage, d'une différence entre le génome de référence et le

génome la lignée cellulaire étudiée. Dans tous les cas, il est clair que l'alignement de ces paires de lecture est peu fiable, on choisit donc de les retirer de l'analyse. La seconde opération de nettoyage des alignements concerne l'unicité de l'alignement des lectures. En effet, lors de l'alignement d'une lecture de séquençage, il est possible que celle-ci s'aligne à plusieurs endroits dans le génome de référence, sans qu'il soit possible de trancher en faveur d'un endroit par rapport à un autre. Les logiciels d'alignements peuvent adopter différentes stratégies vis-à-vis de ce cas de figure, comme par exemple retourner tous les alignements possibles, ou bien choisir un alignement au hasard parmi ceux possibles. On préfère généralement se débarrasser de ces lectures, car il est impossible d'être certain de leur bon alignement sur le génome. Une façon de le faire est de se baser sur la métrique de qualité d'alignement associée à chaque alignement. Le logiciel calcule cette métrique à partir du score d'alignement de chaque alignement possible. Si un des alignements possibles a un score beaucoup plus élevé que tous les autres, alors il y a de grandes chances que ce soit le bon, et la qualité d'alignement augmente. A l'inverse, si plusieurs alignements d'une même lecture ont des scores similaires, alors il est difficile de déterminer quel alignement est le bon, et donc la qualité diminue. De manière générale, la qualité est rapportée comme : $Q = -10\log_{10}(p)$, où p est une estimation de la probabilité qu'il y ait une erreur d'alignement. Ainsi, plus la qualité de l'alignement est élevée, plus il est probable que la position identifiée pour la lecture soit la bonne. C'est une pratique courante que d'utiliser le seuil $Q = 10$, moins d'une chance sur 10 qu'il y ait une erreur d'alignement. J'ai donc retiré de l'analyse toutes les lectures dont la qualité d'alignement était inférieure à 10. Cette métrique étant indépendante du type de séquençage (paired-end ou single-end), j'ai appliqué ce filtre sur toutes les données utilisées.

Après alignement et nettoyage de ces alignements, il s'agit de transformer les données d'alignement en un signal correspondant au positionnement nucléosomal. La question est donc : Quelle est la contribution de chaque alignement au signal de positionnement nucléosomal ? Selon les publications, les méthodes peuvent différer. Plusieurs possibilités s'offrent donc à nous :

— On peut compter la totalité de chaque fragment. C'est à dire que lorsqu'un fragment est aligné, on ajoute 1 à toutes les positions concernées par l'alignement. L'idée est de représenter une forme de "couverture nucléosomale". Il est à noter que cette méthode n'est possible qu'avec un séquençage paired-end, pour lequel on connait donc à la fois la position du début et de la fin du fragment.

— On peut compter 60 paires de bases au milieu de chaque fragment (Ishii et al., 2015). C'est à dire que pour chaque fragment aligné, on ajoute 1 aux 60 paires de bases autour du milieu du fragment. L'idée est ici de représenter la couverture du centre des nucléosomes. Ici encore, cela n'est possible qu'avec un séquençage paired-end, car on a besoin des positions de début et de fin de fragment.

— On peut compter une seule position par fragment. L'idée est ici de représenter un point précis par nucléosome, qui peut être la dyade (le milieu du nucléosome), l'entrée, ou la sortie du nucléosome. Pour un séquençage paired-end, on ajoute donc 1 à la base correspondant au milieu, au début ou à la fin du fragment. L'avantage de cette méthode est qu'elle est adaptable au séquençage de type single-end. Dans le cas d'une lecture alignée sur le brin sens, le nucléosome associé est en aval, la position de début de l'alignement d est donc l'entrée du nucléosome et la position de la dyade est estimé 70 bp en aval. De même, dans le cas d'une lecture alignée sur le brin anti-sens, le nucléosome associé est en amont, la

position de fin de l'alignement f est donc la sortie du nucléosome et la position de la dyade est estimé 70 bp en amont. Il suffit donc de prendre l'ensemble des positions $d + 70$ pb des alignement sens et $f - 70$ pb des alignements anti-sens pour obtenir le signal correspondant aux positionnements des dyades de nucléosomes (Drillon et al., 2016; Schones et al., 2008; Valouev et al., 2011).

Pour ne pas se limiter aux jeux de données de type paired-end (ce qui restreindrait encore plus la disponibilité des données expérimentales) et permettre la comparaison entre les différents jeux de données malgré des techniques de séquençage différentes, j'ai opté pour la troisième option, à savoir le comptage d'une position de dyade par fragment, le milieu des fragments pour le séquençage paired-end, et les positions $d + 70$ pb ou $f - 70$ pb des alignements pour le séquençage single-end.

3.2.1.3 Analyser le positionnement des nucléosomes aux bords des barrières nucléosomales

Une fois les signaux produits, il reste à effectuer à proprement parler leur analyse aux bords des barrières nucléosomales. Deux analyses sont alors menées en suivant le protocole proposé par Drillon et al. (Drillon et al., 2016). La première consiste à vérifier que les barrières nucléosomales correspondent bien à des régions déplétées en nucléosomes dans les données expérimentales. Pour ça, on peut effectuer un Z-test, qui consiste à calculer une statistique (z-score), caractérisant l'enrichissement en nucléosomes dans les deux fenêtres de 300 pb flanquants chaque NIEB par rapport au NIEB. Cette statistique dépend de la moyenne et l'écart-type du nombre de lectures par position dans le NIEB (m_1 et σ_1) et dans les deux régions flaquantes (m_2 et σ_2). Le calcul est alors le suivant :

$$z = \frac{m_2 - m_1}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

où n_1 et n_2 sont les nombres de sites dans la barrière et dans les régions flanquantes ($n_2 = 2 * 300 =$ 600 pb), respectivement. Pour pouvoir calculer ces statistiques avec des fenêtres de 300 pb aux bords de chaque barrière sans être perturbé par une autre barrière trop proche, j'ai restreint le jeu de données de barrières nucléosomales à celles distantes d'au moins 300 pb d'une autre barrière nucléosomale, de chaque côtés de la barrière. J'ai également imposé un total de lectures minimum de 10 pb, car en deçà de cette valeur, le calcul de la statistique est peu fiable et le test a une puissance limitée.

Ce Z-test est effectué pour chaque barrière individuellement. Sous l'hypothèse nulle (qui stipule qu'il n'y a pas de déplétion en nucléosome dans la barrière, donc que la densité de lecture dans et hors de la barrière sont identiques), on approxime la distribution des z-scores par une distribution Normale. On a donc 5 % de chances d'avoir un z-score supérieur à 1.645, et 5 % de chance d'avoir un z-score inférieur à -1.645. On alors peut classer les barrières en trois groupes, celles déplétées en nucléosomes (G1, z-score > 1.645), celles pour lesquelles on ne peut pas conclure sur une différence d'occupation nucléosomale dans le NIEB par rapport à ses régions flanquantes (G2, 1.645 > z-score > -1.645), et celles enrichies en nucléosome (G3, z-score < 1.645). Pour représenter graphiquement les résultats obtenus, j'ai établi la distribution des z-scores, en représentant en vert les zones correspondant au G1, en jaune celles correspondant au G2, et en rouge celles

correspondant au G3. Il en résulte des figures dont un exemple est montré en bas à gauche de la **Figure 3.1**. Sous l'hypothèse nulle de même couverture en nucléosome dans les NIEBs et à leurs bords, la majorité de la distribution doit être jaune, avec 5 % de régions vertes et 5 % de régions rouges. Nous verrons, dans les figures suivantes, que ce n'est jamais le cas.

L'autre analyse à mener avec ces signaux consiste à étudier le positionnement nucléosomal aux bords des barrières. Bien souvent, la couverture en lectures pour chaque nucléosome ne permet pas cette analyse NIEB par NIEB (**Tableaux 3.1** et **3.2 - colonnes** $\overline{C}$). On considère alors le profil d'occupation en nucléosome moyenné sur une population de NIEBs, obtenue en les alignant sur leurs bords. Pour chaque distance au bord (comptée négativement à l'intérieur des barrières), on calcule la moyenne du nombre de lectures alignées pour chaque NIEB. J'ai établi ce signal pour la zone [−35 pb, 500 pb]. Pour contrôler la distance au bord de barrière le plus proche, j'ai restreint l'analyse aux NIEBs de tailles supérieures à 70 pb, et aux inter-NIEBs de tailles supérieures à 1000 pb. Afin que les courbes obtenues soient comparables entre les différents jeux de données, j'ai systématiquement normalisé chaque courbe par la moyenne génomique du signal dans le jeu de données correspondant. Ainsi, observer une valeur de 2 pour une position données signifie qu'à cette distance des bords de barrière, l'occupation en nucléosome est deux fois supérieure à la moyenne génomique. Cette normalisation permet de s'affranchir des larges différences de couverture observées entre les différents jeux de données (**Tableaux 3.1** et **3.2**). Un exemple de courbes ainsi obtenues est présent en bas à droite de la **Figure 3.1**.

3.2.1.4 Validation du pipeline d'analyse avec les données expérimentales humaines

Avant d'utiliser les méthodes décrites précédemment pour étendre la validation des prédictions de notre modèle de positionnement des nucléosomes à de nouvelles espèces, il était nécessaire de vérifier que ces méthodes produisent bien les résultats escomptés. Dans cette optique, j'ai reproduit l'analyse déjà menée antérieurement qui a permis de confirmer les prédictions du modèle chez l'humain (Drillon et al., 2016), en utilisant les données MNase-seq obtenues sur des cellules CD4+ humaines ainsi que sur de la chromatine reconstituée *in vitro* par Valouev et al. (Valouev et al., 2011). Ces données sont de type single-end, elles permettent donc de valider la partie single-end du pipeline. Concernant la partie paired-end, comme aucun jeu de données de ce type n'a été analysé précédemment, j'ai opté pour des données chez l'humain, publiées par Mieczkowski et al. (Mieczkowski et al., 2016), avec l'a priori qu'en travaillant dans la même espèce, et en appliquant un protocole censé retourner des résultats très similaires comme expliqué dans la **Partie 3.2.1.2**, je devrais retrouver des résultats compatibles avec les données single-end. J'ai donc appliqué les méthodes décrites précédemment aux différents jeux de données (**Tableau 3.1**), et comparé les résultats obtenus avec ceux déjà publiés (Drillon et al., 2016). La distribution des z-scores et celle de l'occupation nucléosomale moyenne aux bords des barrières sont présentées respectivement en **Figure 3.2** et **Figure 3.3**.

Les distributions des z-scores obtenues avec les données de Valouev, que ce soit pour les données *in vivo* ou *in vitro* (**Figure 3.2 - A et B**), sont en accord avec celles publiées précédemment (Drillon et al., 2016). La très grande majorité des barrières nucléosomales sont identifiées ici comme des régions déplétées en nucléosomes (respectivement, 96.2 % et 92.4 % des barrières pour les données *in vivo* et *in vitro*). On note que ces chiffres sont supérieurs à ceux publiés

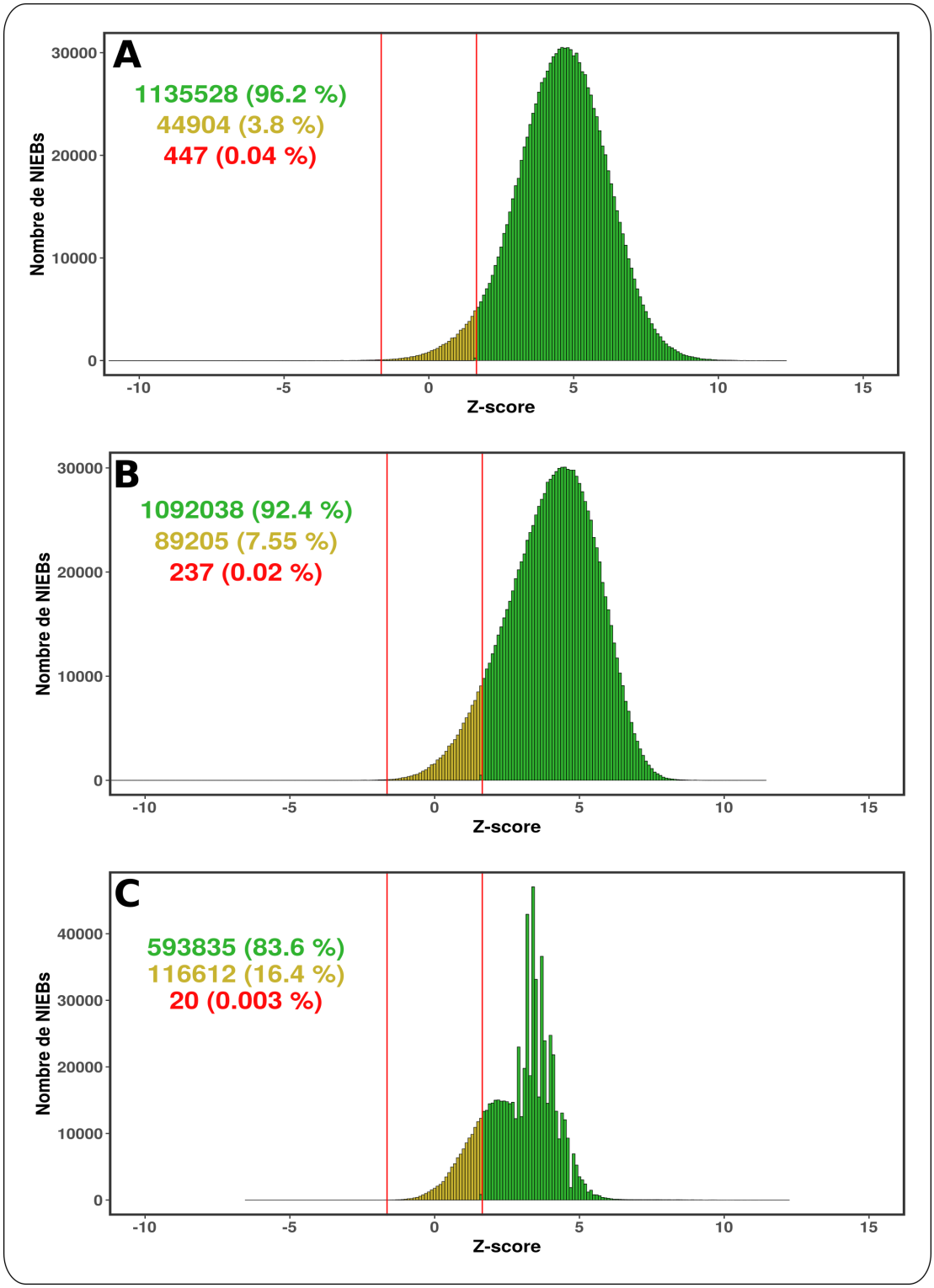


FIGURE 3.2 – **Distributions des z-scores des barrières nucléosomales chez l'humain obtenues à partir des données expérimentales de Valouev et Mieczkowski.** Les z-scores ont été obtenus selon la méthode détaillée en **Partie 3.2.1**. Les figure A, B et C représentent respectivement les résultats obtenus avec les données de Valouev *in vivo*, Valouev *in vitro* et Mieczkowski. La couleur verte représente les z-scores traduisant une déplétion significative en nucléosomes à l'intérieur des barrières par rapport à leurs bords. La couleur rouge représente les z-scores traduisant un enrichissement significatif des nucléosomes à l'intérieur des barrières. La couleur jaune représente les z-scores ne traduisant pas d'enrichissement significatif ni à l'intérieur ni à l'extérieur des barrières. Les barres verticales rouges représentent les seuils de significativité au risque 5 % (−1.645 et 1.645). Les NIEBs utilisés pour le calcul sont distant d'au moins 300 pb de leurs deux voisins et ne sont considérées que les régions (NIEB + régions flanquantes) avec au moins 10 lectures alignées. Le nombre et la proportion de barrières dans chaque catégorie sont indiqués en suivant le même code de couleur.

(respectivement 76.7 % et 54 % pour les données *in vivo* et *in vitro*). Cela peut s'expliquer par le fait qu'on a restreint ici le jeu de données aux barrières pour lesquelles on avait au moins 10 lectures alignées dans la zone analysée, ce que l'analyse précédente ne faisait pas. On observe d'ailleurs qu'on ne trouve pas forcément plus de barrières nucléosomales significativement déplétées en nucléosomes dans mon analyse que dans la précédente. En fait, pour les données *in vitro*, on en retrouve même moins (1.135 millions dans mon analyse contre 1.180 millions dans la précédente). Là où les chiffres sont sensiblement différents, c'est pour les barrières identifiées comme non-significatives. On en retrouve entre 350 000 et 700 000 selon les jeux de données dans l'analyse précédente, contre seulement entre 45 000 et 90 000 dans mon analyse. Il semble donc qu'imposer un nombre minimal de lectures fasse drastiquement baisser le nombre de barrières pour lesquelles le seuil de significativité n'est pas atteint, ce qui est attendu dans la mesure où avec un faible nombre de lectures la puissance du test est limitée. Une autre différence entre les deux protocoles peut aussi expliquer la plus grande proportion de barrières ni déplétées ni enrichies dans la première analyse par rapport à la mienne. Dans mon analyse, on se concentre sur les barrières distantes d'au moins 300 pb d'une autre barrière afin d'être sûr que les deux zones analysées dans le Z-test soient bien respectivement une zone "barrière" et deux zones flanquantes "non-barrière". Dans l'analyse précédente, il semble que cette distinction n'ait pas été effectuée, et que les Z-tests ait été appliqués quel que soit la distance entre la barrière analysée et les barrières adjacentes. Ainsi, dans les fenêtres de 300 pb flanquant la barrière analysée, d'autres barrières peuvent être présentes. Cela revient donc à comparer une zone "barrière" avec deux zones possiblement constituées à la fois de parties "barrière" et "non-barrière". Donc de potentiellement comparer une zone déplétée avec une zone "mi-déplétée/mi-enrichie", ce qui peut amener à des tests non-significatifs non pas parce que les zones déplétées ne le sont pas, mais parce qu'elles sont comparées entre elles plutôt que d'être comparées aux zones enrichies. Les deux différences de protocoles exposées ci-dessus peuvent donc expliquer les différences de proportions observées entre les deux analyses. On voit néanmoins que la conclusion à tirer de ces résultats est inchangée, à savoir que les barrières sont, dans tous les cas, identifiées comme des zones déplétées en nucléosomes. Cela valide la partie de la méthode présentée précédemment concernant les Z-tests et corrobore que les NIEBs correspondent à des NDRs *in vitro* et *in vivo*.

Les résultats obtenus avec les données paired-end (**Figure 3.2 - C**) permettent de valider la méthode mise au point pour les données de type paired-end. On retrouve dans ces données des résultats très similaires à ceux obtenus avec les données single-end, à savoir une très grande majorité de barrières identifiées comme des NDRs (83.6 %, en vert), le reste étant principalement des barrières identifiées comme ni déplétées, ni enrichies en nucléosomes (16.4 %, en jaune), avec seule une infime part des barrières identifiées comme enrichies en nucléosomes (0,003 %, en rouge). Les forts pics observés dans ces données sont un artefact lié au manque de couverture, qui est inférieure à celle des deux autres jeux de données utilisés ici (**Tableau 3.1**). Il en résulte notamment qu'après le filtrage imposant au moins 10 lectures alignées dans chaque zone analysées, le nombre de barrière considérées est bien inférieur pour ces données paired-end que pour les données single-end (~ 710 000 pour les premières contre près d'1.2 million pour les autres). L'analyse n'en est pas moins pertinente, et le résultat reste tout à fait interprétable, à savoir qu'avec ce jeu de données, les barrières sont également identifiées comme des régions déplétées en nucléosomes, ce qui valide la partie "paired-end" de mon protocole d'analyse.

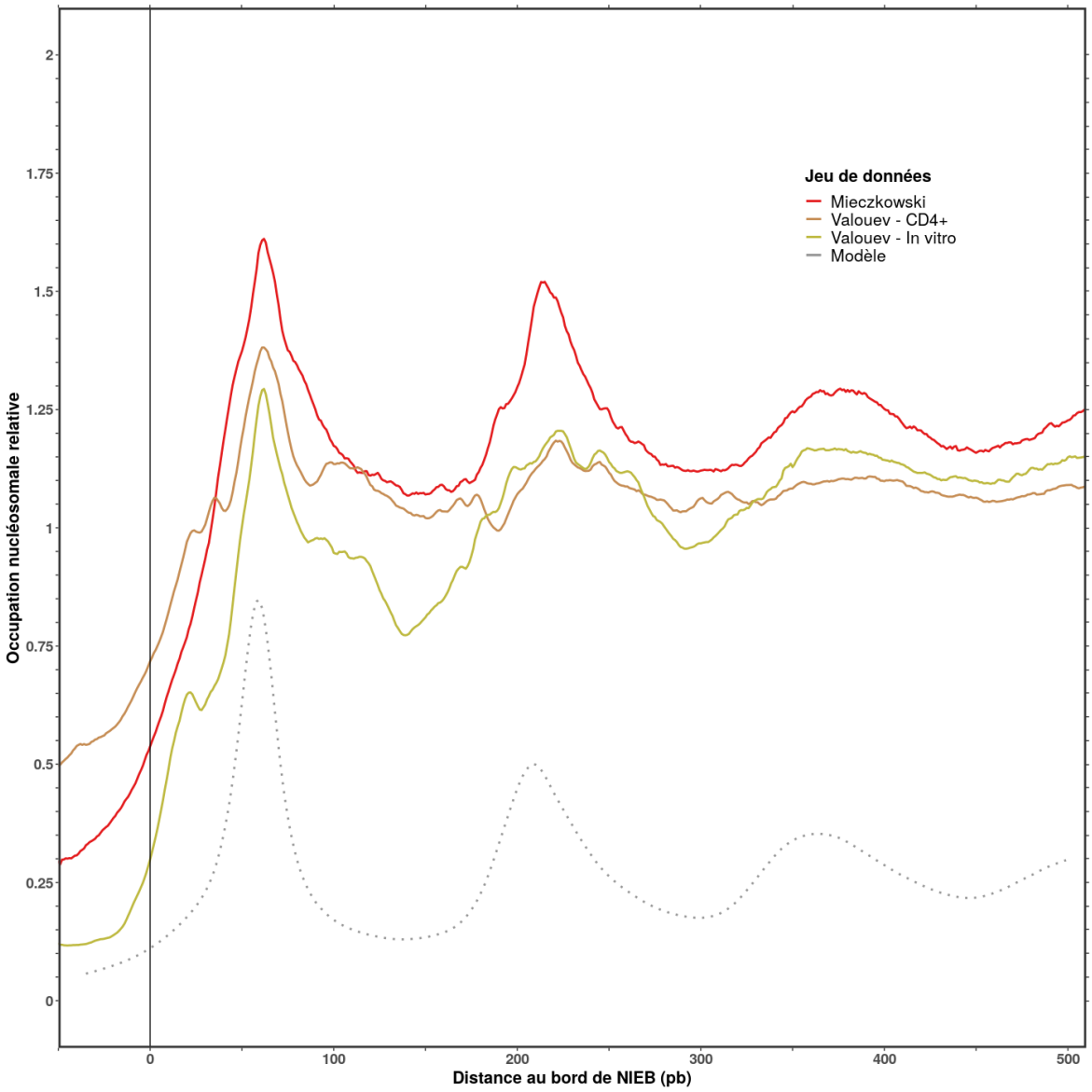


FIGURE 3.3 – **Distributions des nucléosomes aux bords des barrières nucléosomales chez l'humain obtenues à partir des données expérimentales de Valouev et Mieczkowski.** Les profils d'occupation relative en nucléosome ont été obtenus selon la méthode détaillée en **Partie 3.2.1** pour les données de Mieczkowski (courbe rouge), les données de Valouev *in vivo* (courbe marron) et les données de Valouev *in vitro* (courbe jaune). La courbe grise pointillée représente le positionnement prédit par le modèle physique de positionnement des nucléosome déjà présenté en **Figure 2.8** (ici, les valeurs prédites ont été divisées par 4 pour ne pas écraser le reste des courbes).

Concernant la seconde partie du pipeline d'analyse, on observe que celle-ci est aussi validée par les résultats obtenus avec les jeux de données chez l'humain (**Figure 3.3**). En effet, l'analyse des deux jeux de données single-end issus de la publication de Valouev et al. (Valouev et al., 2011) retourne les mêmes résultats que ceux publiés par Drillon et al. (Drillon et al., 2016). On observe bien une diminution drastique de l'occupation nucléosomale au bord interne de la barrière nucléosomale, illustrant l'absence de nucléosome dans les barrières déjà mise en évidence précédemment par la distribution des z-score de chaque barrière (**Figure 3.2**). De plus, le profil expérimental d'occupation nucléosomale est bien en accord avec celui prédit par le modèle (**Figure 3.3**, courbe pointillée). On observe des pics dans l'occupation nucléosomale, traduisant de positions préférentielles pour le nucléosome aux bords des barrières, autour des positions 70 pb, 220 pb et 370 pb.

Entre ces pics, on observe une diminution de l'occupation. On note que cette diminution est plus prononcée dans les donnée *in vitro* que *in vivo*, comme précédemment observé ((Drillon et al., 2016) - Figure 1.A et 1.A'). Ainsi, les résultats obtenus avec mon pipeline reproduisent précisément ceux obtenus lors de la première analyse de ces mêmes données, tant pour les données *in vitro* qu'*in vivo*. La seconde partie du pipeline d'analyse est donc validée pour les données single-end. La correspondance des prédictions du modèle physique de positionnemnet des nucléosome avec les données expérimentales *in vitro* indique que le modèle a bien capturé les propriétés de séquences de l'interaction entre histones et ADN ; la correspondance avec les résultats *in vivo* indique que ces propriétés de séquences sont pertinentes aux NIEBs même en présence des facteurs agissant sur le positionnement nucléosomal, en d'autre termes que les régions NIEBs ne sont pas significativement sujettes aux remodelages de la chromatine (Drillon et al., 2016). Les résultats obtenus avec les données paired-end valident quant à eux la seconde partie du pipeline d'analyse pour ce type de données. On observe les mêmes résultats avec les données de Mieczkowski (paired-end) qu'avec celles de Valouev (single-end), à savoir une forte diminution de l'occupation nucléosomale à l'intérieur des barrières, et un positionnement très fort de deux à trois nucléosomes aux bords de ces barrières, aux mêmes positions que celles précédemment observées et surtout que celles prédites par le modèle.

L'analyse de deux jeux de données de type single-end, pour lesquels les résultats étaient déjà publiés (Drillon et al., 2016), a donc permis de valider le pipeline d'analyse de données expérimentales single-end de positionnement de nucléosomes présenté précédemment. De plus, l'obtention de résultats en tout point similaires avec un jeu de données paired-end chez la même espèce a permis de valider la méthode pour l'analyse des données de type paired-end. Cette seconde validation est intéressante dans la mesure où ce type de données est de plus en plus utilisé dans les études du positionnement nucléosomal. Pour ce type de données on a accès à la longueur des fragments séquencés et donc à la longueur d'ADN engagé dans la formation des nucléosomes. On a ainsi en principe la possibilité d'identifier les positions nucléosomales pour lesquelles l'enroulement de l'ADN autour des histones n'est pas complet. Il apparaissait donc utile de mettre au point une méthode capable de tenir compte de cette information, ce qui a été fait dans la mise au point du pipeline. Cette information a d'ailleurs été prise en compte dans mes premières analyses de données MNase-seq paired-end. Les résultats obtenus ont révélé un besoin, pour leur compréhension complète, d'analyser la dynamique de notre modèle. La mise au point de telles simulations sort complètement du cadre de cette thèse, aussi l'information de taille des fragments n'a pas été exploitée dans mon travail. Nous avons donc ici une méthode automatisée qui nous permet d'analyser des jeux de données expérimentaux de positionnement de nucléosomes, afin de valider les prédictions de notre modèle physique de positionnement nucléosomal, en répondant à deux questions :

— Les barrières nucléosomales correspondent-elles à des zones déplétées en nucléosomes dans les données expérimentales ? La réponse à cette question est à retrouver dans la distribution des z-scores comme celle présentée en **Figure 3.2**.

— Le positionnement nucléosomal aux bords des barrières est-il en accord avec celui qui est prédit par le modèle ? La réponse à cette question est à retrouver dans les courbes d'occupation moyennes des nucléosomes aux bords des barrières comme celles présentées en **Figure 3.3**.

Avec ce nouvel outil, j'ai donc pu étendre la validation du modèle à de nouvelles espèces pour

lesquelles des données de type MNase-seq étaient disponibles, à savoir la souris (Carone et al., 2014), la drosophile (Chereji et al., 2019; Mieczkowski et al., 2016), le poisson-zèbre (Zhang et al., 2014) et l'arabette (Pass et al., 2017). Les résultats obtenus dans ces espèces sont présentés dans la **Partie 3.2.2**.

3.2.2 Les barrières nucléosomales correspondent à des régions déplétées en nucléosomes chez quatre nouvelles espèces

L'outil mis au point pour l'analyse de données expérimentales de positionnement de nucléosomes décrit dans la **Partie 3.2.1**, et validé avec les données expérimentales chez l'humain (**Partie 3.2.1.4**), a ensuite été utilisé pour analyser une série de jeux de données, afin d'étendre la validation des prédictions du modèle de positionnement à de nouvelles espèces. Malheureusement, les données expérimentales de positionnement de nucléosomes ne sont pas disponibles pour toutes les espèces étudiées dans le **Chapitre 2**, ce qui limite le potentiel de validation. En fait, des données de MNase-seq ne sont disponibles que pour très peu d'espèces, comme on peut le constater dans le recensement de ce type de données effectué par Vladimir Teif (Teif, 2016). Grâce à cet inventaire des données disponibles, et à un travail bibliographique, j'ai tout de même pu identifier des jeux de données pour quatre des espèces étudiées dans le **Chapitre 2**, à savoir la souris, le poisson-zèbre, la drosophile et l'arabette. Le détail de ces jeux de données est présenté dans le **Tableau 3.2**. J'ai analysé ces données expérimentales de positionnement de nucléosomes en utilisant le protocole mis au point précédemment, afin de voir dans quelle mesure les prédictions de notre modèle sont pertinentes pour ces espèces, tant en terme de correspondance entre barrières nucléosomales et NDRs que de positionnement des nucléosomes aux bords des barrières. Les résultats obtenus dans chacune des espèces sont présentés dans les **Figures 3.4**, **3.5** et **A.7**.

3.2.2.1 Les barrières nucléosomales correspondent à des régions déplétées en nucléosomes *in vivo* dans les quatre espèces analysées

La grande majorité des barrières nucléosomales des quatre espèces étudiées ici sont identifiées comme des régions déplétées en nucléosome (**Figure 3.4**). En effet, on voit que le z-score calculé pour chaque barrière dépasse la valeur seuil pour 77.8 %, 89 %, 88.4 % et 98.4 % des cas respectivement chez la souris, la drosophile, le poisson-zèbre et l'arabette. De plus, parmi les barrières non-identifiées comme significativement déplétées en nucléosome, seule une infime partie sont identifiées comme significativement enrichies en nucléosome (moins de 0.5 % dans chaque espèce). Le reste des barrières nucléosomales (respectivement 22 %, 10.6 %, 11.2 % et 1.38 % chez la souris, la drosophile, le poisson-zèbre et l'arabette) sont identifiées comme ni déplétées, ni enrichies en nucléosomes. On voit donc ici assez clairement que les barrières nucléosomales sont bien identifiées, *in vivo*, comme des régions déplétées en nucléosomes. Les résultats obtenus sont également cohérents entre jeux de données appartenant à la même espèce (**Figure A.7**). Cependant, si les tendances générales sont conservées, à savoir une grande majorité de barrières significativement déplétées en nucléosomes, très peu de barrières significativement enrichies et le reste des barrières ni déplétées, ni enrichies, on observe quelques différences dans le pourcentage de barrière appartenant à chaque catégories. Chez le poisson-zèbre notamment, où l'analyse a été menée sur

Espèce	Publication	Identifiant	Type de séquençage	Nombre de lectures	$\overline{C}$
Souris	(Carone et al., 2014)	Carone	Paired-end	228 033 166	17.2
Souris	(Mieczkowski et al., 2016)	Mieczkowski	Paired-end	39 963 129	3.0
Drosophile	(Mieczkowski et al., 2016)	Mieczkowski	Paired-end	23 220 773	32.2
Drosophile	(Chereji et al., 2019)	Chereji	Paired-end	67 800 376	94.0
Poisson-zèbre	(Zhang et al., 2014)	Zhang - 256c - 1	Single-end	83 346 946	11.8
Poisson-zèbre	(Zhang et al., 2014)	Zhang - 256c - 2	Single-end	311 958 220	44.2
Poisson-zèbre	(Zhang et al., 2014)	Zhang - dome - 3	Single-end	98 299 306	13.9
Poisson-zèbre	(Zhang et al., 2014)	Zhang - dome - 4	Single-end	338 216 752	47.9
Arabette	(Pass et al., 2017)	Pass 1	Paired-end	89 003 154	134.6
Arabette	(Pass et al., 2017)	Pass 2	Paired-end	138 974 511	210.2

TABLEAU 3.2 – **Données expérimentales utilisées pour étendre la validation du modèle physique de positionnement de nucléosomes.** Le jeu de données Carone *in vivo* correspond aux données obtenues sur le sperme de souris, avec centrifugation (dataset spun). Les jeux de données Miezckowski chez la souris et la drosophile correspondent respectivement aux données obtenues avec les cellules J1 ESC et S2, avec la plus grande quantité de MNase (64U et 100U). Le jeu de données Chereji correspond à un pool des quatre réplicats où le temps de digestion était de 15 min (pour être autant que possible dans des conditions semblables aux autres jeux de données). Les deux jeux de données chez l'arabette correspondent aux conditions de digestion complètes (120 U). L'identifiant correspond à celui utilisé dans la **Figure 3.5**. Le nombre de lectures et la valeur $\overline{C}$ représentent la même chose que dans le **Tableau 3.1**. Pour le calcul des couvertures moyennes par nucléosome, le nombre de nucléosome par génome a été estimé d'après les NRL suivantes : respectivement 180, 190, 190, 200 et 200 pb pour l'arabette (Choi et al., 2020), la drosophile (Cartwright & Elgin, 1986), le poisson-zèbre (Wu et al., 2011), la souris (Popova et al., 2013) et l'humain (Valouev et al., 2011).

quatre jeux de données correspondant à deux expérience pour chacune desquelles deux réplicats ont été réalisés (**Tableau 3.2**, on voit que le pourcentage de barrières significativement déplétées en nucléosomes peut varier entre 70 % et 90 % selon les réplicats (**Figure 3.4 - C** vs. **Figure A.7 - C** et **Figure A.7 - D** vs. **Figure A.7 - E**). Cependant, on remarque également une disparité de la valeur $\overline{C}$ du nombre moyen de lecture par nucléosome entre ces jeux de données. En effet, cette valeur est autour de 12 pour les deux jeux de données présentant ~ 70 % de NIEBs significativement déplétés en nucléosomes, contre ~ 45 pour les deux autres jeux de données présentant ~ 90 % de NIEBs significativement déplétés en nucléosomes (**Tableau 3.2**). De manière générale, on observe une corrélation entre la valeur $\overline{C}$ des jeux de données et le pourcentage de barrières identifiées comme significativement déplétées en nucléosomes. En effet, pour les jeux de données dont la valeur $\overline{C}$ est inférieure à 20, la proportion de NIEBs identifiés comme significativement déplétés en nucléosome est de 70 à 80 %. Lorsque la valeur $\overline{C}$ est plus importante, entre 30 et 50, cette proportion monte à 90 %. Enfin, les valeurs $\overline{C}$ proches ou dépassant 100 sont associées aux proportions supérieures à 95 %. Comme attendu, la puissance du test effectué ici est sensible à la quantité de données

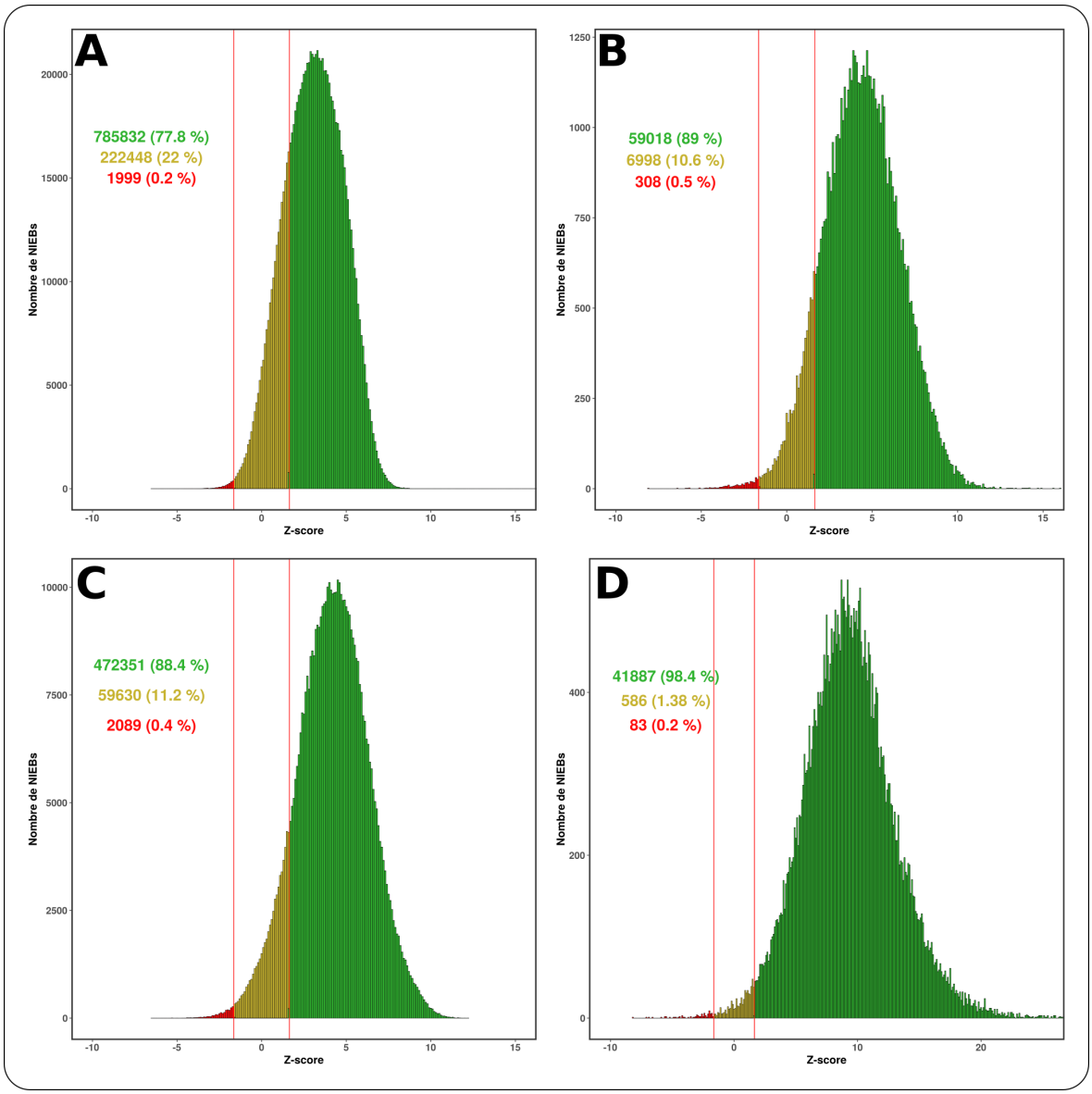


FIGURE 3.4 – **Distributions des z-scores des barrières nucléosomales chez la souris, la drosophile, le poisson-zèbre et l'arabette.** Les figures A, B C et D représentent respectivement les résultats obtenus chez la souris (Carone), la drosophile (Mieczkowski), le poisson-zèbre (Zhang 2) et l'arabette (Pass 1). Les données utilisées pour chaque espèce sont résumées dans le **Tableau 3.2**. Les z-scores ont été obtenus en utilisant le pipeline décrit en **Partie 3.2.1**.

disponibles, représentée ici par la valeur $\overline{C}$. Ainsi, les proportions variables de NIEBs pour lesquels le test s'est révélé non-significatif observées dans la **Figure 3.4** peuvent s'expliquer par un manque de puissance statistique. On observe néanmoins que quel que soit le jeu de données ou l'espèce analysée, on retrouve bien une forte correspondance entre barrières nucléosomales prédites par le modèle et régions déplétées en nucléosome *in vivo*, avec 70 à 98 % des NIEBs identifiées comme des régions significativement déplétées en nucléosomes selon les jeux de données. Le fait que plus la puissance statistique du test est importante, plus la proportion de NIEBs identifiés comme significativement déplétés en nucléosome augmente corrobore qu'une très large majorité des NIEBs sont des NDRs *in vivo*.

3.2.2.2 Le positionnement nucléosomal aux bords des barrières présente une variabilité entre les jeux de données.

La seconde partie de notre analyse des données expérimentales de positionnement nucléosomal consiste à établir la distribution de l'occupation nucléosomale aux bords des barrières, afin de voir si celle-ci est en accord avec celle prédite à l'aide de notre modèle physique de positionnement nucléosomal. L'analyse des jeux de données cités précédemment a permis de mettre en évidence que ce n'est pas toujours le cas (**Figure 3.5**). En effet, si les prédictions obtenues à partir de notre modèle (en pointillée sur chaque figure) sont très cohérentes entre les espèces, ce que l'on avait déjà observé dans le **Chapitre 2** (**Figure 2.8**), l'occupation nucléosomale expérimentale est, elle, plus variable.

Tout d'abord, on voit que chez la souris, les courbes obtenues expérimentalement reproduisent celles prédites par le modèle (**Figure 3.5 - A**). D'abord, on observe une forte diminution de l'occupation à l'intérieur des barrières nucléosomales, confirmant ce qui a été mis en évidence dans la Partie précédente à propos de la correspondance entre NIEBs et NDRs. Ensuite, on retrouve dans les données expérimentales les positions préférentielles pour les nucléosomes prédites par le modèle, ce qui est illustré par la correspondance entre les pics des différentes courbes (**Figure 3.5 - A**). Enfin, on voit que les deux courbes correspondant aux deux jeux de données expérimentales sont en accord l'une avec l'autre, même si l'on observe des différences dans les valeurs d'occupation. En somme, chez la souris, l'ensemble des prédictions du modèle physique de positionnement sont validées expérimentalement. Chez la drosophile, on observe également une cohérence assez bonne entre les données expérimentales et les prédictions du modèle (**Figure 3.5 - B**). On retrouve bien la diminution de l'occupation à l'intérieur des barrières. Concernant les positions préférentielles pour le nucléosome aux bords de ces barrières, on retrouve principalement la première position (à ~70 pb), les autres étant clairement moins marquées chez cette espèce que chez la souris ou l'humain. Également, les deux jeux de données expérimentales utilisés semblent globalement cohérents entre eux. On observe cependant une différence entre ces deux jeux de données entre les positions 100 et 150. Sur la courbe de prédiction du positionnement (**Figure 3.5 - B, courbe en pointillé court**), cette zone présente une occupation basse, que l'on retrouve également dans les données de Chereji (**Figure 3.5 - B, courbe en pointillé long**), dans une moindre mesure. En revanche, dans les données de Mieczkowski (**Figure 3.5 - B, courbe pleine**), on observe à cet endroit une légère augmentation de l'occupation au lieu d'une diminution. Cela pourrait illustrer une potentielle autre position préférentielle pour le nucléosome, différente de celles prédites par le modèle. En somme, chez la drosophile, seule une partie des prédictions du modèle sont confirmées, à savoir la correspondance entre NIEBs et NDRs *in vivo*, et le positionnement du premier nucléosome contre la barrière nucléosomale. En revanche, pour le reste du positionnement, les résultats obtenus avec les données expérimentales ne sont pas toujours cohérents entre les deux jeux de données, ou entre prédictions et données expérimentales. On pourrait dans cette espèce avoir aux régions NIEBs du remodelage variable entre lignée cellulaires.

C'est chez le poisson-zèbre et l'arabette que l'on observe les plus grosses différences entre les données expérimentales et les prédictions à partir du modèle (**Figure 3.5 - C et D**). En effet, dans ces deux espèces, quel que soit le jeu de données utilisé, on voit que les profils de positionnement aux bords des barrières nucléosomales ne permettent pas d'identifier des positions préférentielles,

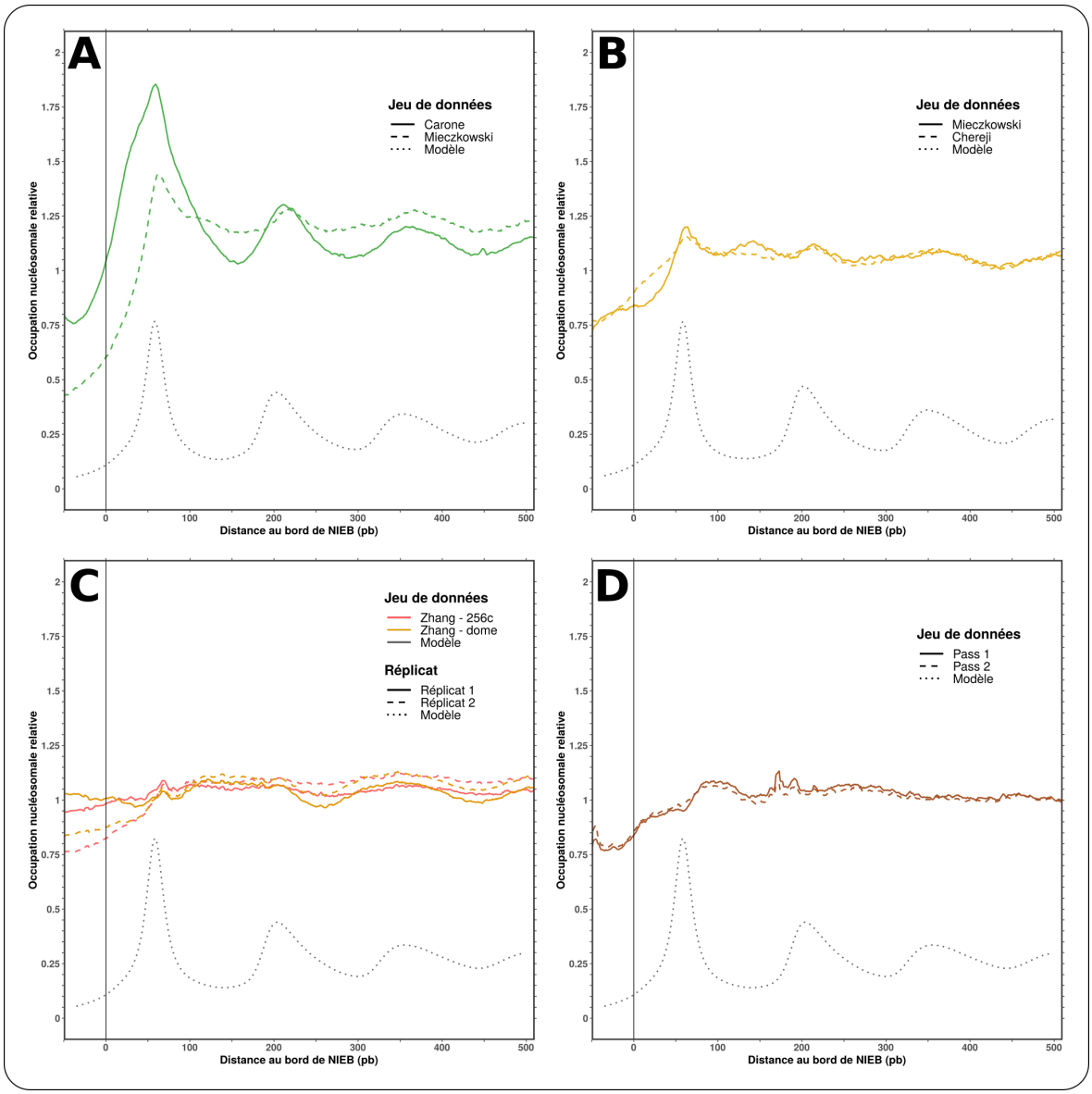


FIGURE 3.5 – **Distributions des nucléosomes aux bords des barrières nucléosomales chez la souris, la drosophile, le poisson-zèbre et l'arabette.** Les données utilisées pour chaque espèce sont résumées dans le **Tableau 3.2**. Les figure A, B C et D représentent respectivement les résultats obtenus chez la souris, la drosophile, le poisson-zèbre et l'arabette. Les distributions ont été obtenues en utilisant le pipeline décrit en **Partie 3.2.1**.

contrairement à ce qui a été prédit à partir de la séquence à ces loci. Le seul point de cohérence entre le modèle et les données concerne l'intérieur des barrières, où l'occupation diminue dans tous les cas par rapport à l'extérieur, ce qui est en accord avec la distribution des z-scores de l'étude précédente (**Figure 3.4**). Chez l'arabette, les deux jeux de données utilisés semblent en accord l'un avec l'autre (**Figure 3.5 - D**). On note que ces jeux de données ne sont pas des réplicats de la même expérience mais deux expérience différentes, une (Pass 1) où les cellules ont été cultivées en présence de lumière et l'autre (Pass 2) en l'absence de lumière. Cela ne semble pas avoir d'impact sur le positionnement nucléosomal aux bords des NIEBs, celui-ci étant remarquablement similaire dans les deux conditions. Chez le poisson-zèbre en revanche, il y a des différences entre les expériences, et même aussi entre les réplicats des expériences (**Figure 3.5 - C**). À l'extérieur des barrières (entre les positions 0 et 500), on observe de légères différences d'occupation, certaines

courbes présentant des diminutions plus ou moins prononcées. Cependant, les effets restent assez faibles, l'occupation n'oscillant qu'entre 0.95 et 1.10. Au bord interne des barrières, les différences observées sont plus marquées, avec une diminution de l'occupation nucléosomale disparate entre les jeux de données. On observe notamment des différences entre les deux réplicats des deux expériences, ce qui diminue globalement la confiance que l'on peut avoir dans les résultats issus de ces jeux de données. En effet, pour l'expérience de MNase effectuée au stade 256 cellules, l'occupation nucléosomale au bord interne des NIEBs diminue à 0.95 pour le réplicat 1 contre 0.75 pour le réplicat 2. La même différence est présente dans l'expérience faite au stade dôme, où l'occupation à l'intérieur des NIEBs est de 1.05 pour le réplicat 1 contre 0.85 pour le réplicat 2. Dans ce dernier cas, cela veut donc dire que le premier réplicat indique que l'occupation nucléosomale à l'intérieur des NIEBs est légèrement supérieure à la moyenne génomique alors que le second réplicat indique au contraire une occupation nucléosomale inférieure à la moyenne génomique. L'interprétation de ces résultats est donc rendue difficile par ces incohérences. De manière plus générale, elles illustrent les problèmes de reproductibilité des expériences de MNase-seq.

3.2.3 L'expérience de MNase-seq comporte des biais que l'on peut en partie corriger avec un nouveau protocole.

Les différences observées précédemment entre des jeux de données pourtant obtenus avec des techniques très similaires, et surtout les incohérences entre les différents réplicats d'une même expérience posent la question de la confiance globale que l'on peut avoir dans les expériences de MNase-seq. Cette enzyme a notamment été identifiée comme ayant des préférences de séquence l'amenant à couper plus facilement l'ADN au niveau des séquences riches en AT (Dingwall et al., 1981). Surtout, les différences d'activité de l'enzyme peuvent rendre difficiles les comparaisons entre expériences (Mieczkowski et al., 2016; Weiner et al., 2010). Dans le protocole de MNase-seq standard, une titration de l'enzyme est effectuée, afin d'obtenir ~ 80 % d'ADN mononucléosomal (Zhang & Pugh, 2011). Cependant, selon les conditions expérimentales, cela peut mener à des niveaux de digestion de la chromatine différents, ce qui complique la comparaison des résultats (Mieczkowski et al., 2016; Weiner et al., 2010). En d'autres termes, des différences dans l'occupation nucléosomale observées entre plusieurs conditions peuvent être interprétées biologiquement alors qu'elles sont en réalité dues à des différences de digestion de la chromatine résultant simplement de légères différences entre protocoles ou de l'efficacité variable de l'enzyme (Mieczkowski et al., 2016).

Les protocoles de digestion standards de la chromatine correspondent à des conditions de digestion intense, ce qui a pour effet de sélectionner préférentiellement les nucléosomes les plus stables, car les nucléosomes plus "fragiles" sont alors déstabilisés puis digérés par l'enzyme avant le séquençage (Chereji et al., 2017; Knight et al., 2014; Mieczkowski et al., 2016; Xi et al., 2011). Le protocole de digestion optimisant la quantité de MNase à utiliser pour obtenir ~ 80 % de mono-nucléosomes a donc tendance à masquer les nucléosomes les plus sensibles à la MNase. Ces nucléosomes sont cependant identifiables, par des protocoles de digestion ménagée, soit en utilisant moins de MNase (Knight et al., 2014; Mieczkowski et al., 2016), soit en stoppant la digestion plus tôt (Chereji et al., 2019; Xi et al., 2011). Ainsi, en comparant les résultats obtenus entre une digestion ménagée et une digestion intensive, on peut mettre en évidence les nucléosomes les

plus sensibles à la MNase. Cela a notamment permis de voir que chez la levure, un certain nombre de régions décrites comme déplétées en nucléosomes comportaient en réalité des nucléosomes sensibles à la MNase qui n'étaient pas visibles dans les résultats dans les conditions standards de digestion (Xi et al., 2011).

L'utilisation de données de MNase-seq standards pour analyser l'occupation des nucléosomes présente donc des biais non négligeables, qui doivent être pris en compte pour interpréter les résultats montrés précédemment dans ce chapitre. En effet, jusqu'ici, nous avons analysé des jeux de données correspondant à des digestions intensives pouvant cacher la présence de certains nucléosomes. Avec ces jeux de données, nous avons vu que les NIEBs correspondent, dans plusieurs génomes, à des régions fortement déplétées en nucléosomes. Or, les mêmes conclusions ont été faites à propos de données expérimentales chez la levure, avant d'être corrigées par l'analyse de données tenant compte du niveau de digestion par la MNase dans les expériences pour mettre en évidence les nucléosomes sensibles à l'enzyme (Xi et al., 2011). Est-il possible que nos barrières nucléosomales soient, à l'instar de certaines NFRs chez la levure, en réalité occupées par des nucléosomes sensibles aux conditions de digestion intensives ? Pour le savoir, il est nécessaire d'analyser des jeux de données comportant plusieurs niveaux de digestion, afin de comparer les résultats obtenus selon que la digestion est ménagée ou intensive. Dans cette optique, j'ai exploré un nouveau type de données, appelé "MNase accessibility" (MACC), dans lequel la digestion de la chromatine n'est plus effectuée avec une seule quantité de MNase mais répétée quatre fois avec quatre quantités différentes correspondant donc à quatre niveaux de digestions (Mieczkowski et al., 2016). L'analyse non plus de l'occupation des nucléosomes mais de leur accessibilité en fonction du niveau de digestion permet également des comparaisons entre espèces plus fiables que les données standards, car l'utilisation de plusieurs niveaux de digestions permet de calibrer les différences d'efficacité de l'enzyme. Des explications détaillées sur les données MACC-seq et l'interprétation des résultats associés sont présentes en **Partie 3.3.1**. Seule ombre au tableau, à l'instar des données MNase classiques, dont la disponibilité limitée a fortement réduit les possibilités de validation du modèle, le type de données nécessaires ici pour l'analyse n'a été produit que dans très peu d'espèces. Dans cette étude, j'ai utilisé les données produites par Mieczkowski et al. (Mieczkowski et al., 2016), chez la souris, l'humain et la drosophile. Les Parties suivantes sont consacrées aux résultats obtenus dans cette analyse, et aux apports de ces résultats dans la compréhension de la chromatine au niveau des barrières nucléosomales.

3.3 Les barrières nucléosomales contiennent des nucléosomes instables : un point d'entrée pour les modifications épigénétiques ?

3.3.1 Comment analyser des données de type MACC-seq ?

Dans la **Partie 3.2.3**, nous avons vu que les expériences de MNase-seq comportent des biais pouvant gêner l'interprétation des résultats ainsi que les comparaisons entre jeux de données, voire entre réplicats d'un même jeu de données. Pour s'affranchir de ces biais, Mieczkowski et al. proposent d'adapter le protocole expérimental pour effectuer ce qu'ils appellent une expérience de MACC-seq (MACC pour MNase accessibility), pour analyser non plus seulement le positionnement

direct des nucléosomes mais l'accessibilité à ces nucléosomes (Mieczkowski et al., 2016). Dans l'expérience de type MACC-seq, la chromatine est toujours digérée avec de la MNase. La principale différence avec les expériences de MNase-seq standards est à trouver dans le niveau de digestion. En effet, là où, dans les expériences de MNase-seq standards, on optimise la quantité de MNase pour obtenir un maximum de mononucléosome (Zhang & Pugh, 2011), dans l'expérience de MACC-seq, on effectue quatre digestions indépendantes de la chromatine avec quatre quantités différentes de MNase prises sur une échelle exponentielle (Mieczkowski et al., 2016). Cela permet d'obtenir quatre niveaux de digestion, que l'on peut identifier comme "très légère", "légère", "moyenne" et "complète". La comparaison des résultats d'occupation obtenus avec les quatre niveaux de digestion permet de déterminer les changements d'occupation en fonction du niveau de digestion, ce qui est une mesure de l'accessibilité aux nucléosomes (Mieczkowski et al., 2016). Pour quantifier ce changement, Mieczkowski et al. proposent de calculer le paramètre MACC, correspondant au coefficient de la courbe de la régression linéaire ajustée aux données. En d'autres termes, il s'agit de comparer, à un locus, le signal obtenu avec les différents niveaux de digestion pour déterminer si :

— Le signal augmente avec la quantité de MNase, ce qui signifie que plus la digestion est intense, plus l'occupation nucléosomale est importante. Dans ce cas, le paramètre MACC est positif.

— Le signal diminue avec la quantité de MNase, ce qui signifie que plus la digestion est intense, plus faible est l'occupation. Dans ce cas, le paramètre MACC est négatif.

Le signe du paramètre MACC peut alors être interprété comme une mesure de l'accessibilité aux nucléosomes (Mieczkowski et al., 2016). Si le paramètre MACC est positif, cela signifie que l'intensité du signal est corrélée à la quantité de MNase, donc au niveau de digestion. Or, les niveaux de digestion importants ont tendance à sélectionner les nucléosomes les plus stables et inaccessibles, les nucléosomes plus instables et accessibles étant, eux, retournés par les niveaux de digestion plus légère (Chereji et al., 2017; Knight et al., 2014; Mieczkowski et al., 2016; Xi et al., 2011). Ainsi, obtenir un signal plus important à un locus avec un fort niveau de digestion qu'avec un faible niveau de digestion indique qu'à ce locus, le nucléosome est plutôt inaccessible, ce qui le rend résistant à la MNase. À l'inverse, un paramètre MACC négatif indique un nucléosome accessible, sensible à la MNase. On peut donc, avec le calcul de la MACC, étudier l'accessibilité des nucléosomes à la MNase. Ce type d'étude permet également les comparaisons entre espèces, car elle dépend de la réponse aux changements de digestion qui ne dépend pas d'une exacte calibration du niveau de digestion. Les différences entre protocoles et conditions expérimentales sont ainsi nivelées. Cette expérience diminue donc les biais liés à l'activité de la MNase, ce qui augmente la confiance que l'on peut avoir dans la comparaison des résultats entre jeux de données (Mieczkowski et al., 2016).

Pour mieux caractériser les nucléosomes aux bords des barrières nucléosomales en ajoutant à nos analyses précédentes une mesure de leur stabilité, j'ai exploré les données MACC-seq produites par Mieczkowski et al. chez l'humain, la souris et la drosophile (Mieczkowski et al., 2016). Le détail des jeux de données, comportant notamment la quantité de MNase utilisées pour chaque niveau de digestion ainsi que le nombre de lectures alignées, est présenté dans le **Tableau 3.3**. Pour obtenir le signal de positionnement aux bords des barrières nucléosomales, j'ai utilisé le pipeline décrit dans la Partie précédente (**Partie 3.2.1**) sur chaque niveau de digestion séparément. J'ai ensuite calculé le paramètre MACC aux bords des NIEBs à trois loci différents :

— 30 pb à l'intérieur des barrière nucléosomales, pour représenter l'accessibilité à d'éventuels

Espèce	Type de cellules	Identifiant	Quantité de MNase (U)	Type de digestion	Nombre de lectures	$\overline{C}$
Humain	K562	Humain-5U	5	Très légère	22 701 752	1.55
Humain	K562	Humain-21U	21	Légère	77 018 510	5.24
Humain	K562	Humain-79U	79	Moyenne	48 277 969	3.29
Humain	K562	Humain-304U	304	Complète	49 113 709	3.34
Souris	J1 ESC	Souris-1U-1	1	Très légère	30 620 507	2.31
Souris	J1 ESC	Souris-4U-1	4	Légère	32 696 576	2.47
Souris	J1 ESC	Souris-16U-1	16	Moyenne	47 868 559	3.62
Souris	J1 ESC	Souris-64U-1	64	Complète	39 963 129	3.02
Souris	E14 ESC	Souris-1U-2	1	Très légère	34 357 257	2.13
Souris	E14 ESC	Souris-4U-2	4	Légère	40 105 312	2.48
Souris	E14 ESC	Souris-16U-2	16	Moyenne	43 029 024	2.86
Souris	E14 ESC	Souris-64U-2	64	Complète	43 479 990	3.15
Souris	NPC	Souris-1U-3	1	Très légère	28 157 467	2.13
Souris	NPC	Souris-4U-3	4	Légère	32 792 008	2.48
Souris	NPC	Souris-16U-3	16	Moyenne	37 805 209	2.86
Souris	NPC	Souris-64U-3	64	Complète	41 748 644	3.15
Souris	eNPC	Souris-1U-4	1	Très légère	38 633 321	2.92
Souris	eNPC	Souris-4U-4	4	Légère	33 227 439	2.51
Souris	eNPC	Souris-16U-4	16	Moyenne	18 592 598	1.40
Souris	eNPC	Souris-64U-4	64	Complète	30 308 452	2.29
Drosophile	S2	Drosophile-1U	1.5	Très légère	18 217 078	25.3
Drosophile	S2	Drosophile-6U	6	Légère	18 277 113	25.3
Drosophile	S2	Drosophile-25U	25	Moyenne	21 008 251	29.1
Drosophile	S2	Drosophile-100U	100	Complète	23 220 773	32.2

TABLEAU 3.3 – **Données expérimentales de positionnement de nucléosome avec la technique de MACC-seq.** Tous les jeux de données sont issus de la même publication de Mieczkowski et al. (Mieczkowski et al., 2016). Le type de cellules NPC correspond à des "neural progenitor" dérivés *in vitro* à partir de cellules J1 ESC. Le type de cellules eNPC correspond à des "Embryonic NPCs". Pour les autres lignées cellulaires, il n'y a pas plus d'information dans la publication qu'indiqué dans le tableau. Les identifiants correspondent à ceux utilisés dans les **Figures 3.6**, **3.7** et **3.10**. Le nombre de lectures et la valeur $\overline{C}$ représentent la même chose que dans le **Tableau 3.1**. Le calcul de la valeur $\overline{C}$ a été effectué avec les mêmes NRL que pour le **Tableau 3.2**.

nucléosomes à l'intérieur des barrières

— aux bords des barrières, en position 65 (correspondant au pic observé dans le positionnement prédit et validé expérimentalement précédemment), pour représenter l'accessibilité aux nucléosomes aux bords des barrières

— À 450 pb des barrières, pour représenter l'accessibilité aux nucléosomes loin des barrières nucléosomales

Pour le calcul du paramètre MACC, j'ai utilisé la fonction lm du logiciel R, qui calcule la régression linéaire entre deux variables (en l'occurence, l'occupation moyenne et la quantité de MNase). Concernant la quantité de MNase, j'ai transformé les quantités du **Tableau 3.3** en une échelle logarithmique, pour conserver une distance similaire entre les différentes quantités. Les résultats obtenus chez l'humain et la souris sont présentés dans la **Partie 3.3.2**. Ceux obtenus chez la

drosophile sont détaillés dans la **Partie 3.4**.

3.3.2 Chez l'humain et la souris, on observe des nucléosomes instables dans les barrières et stables à l'extérieur

Pour explorer les données MACC-seq, j'ai utilisé le pipeline décrit précédemment afin de produire le signal moyen de positionnement nucléosomal aux bords des NIEBs, pour chacun des niveaux de digestion de chaque jeu de données. Les courbes obtenues sont présentées en **Figure 3.6** pour le jeu de données chez l'humain, en **Figures 3.7** et **A.8** pour les jeux de données chez la souris. La valeur des paramètres MACC calculés aux trois positions détaillées précédemment pour chaque jeu de données sont indiqués dans la légende de chaque figure. Les résultats obtenus dans ces deux espèces bousculent la vision que l'on avait jusqu'alors des barrières nucléosomales. En effet, jusqu'ici, que ce soit dans les précédentes publications (Brunet et al., 2018; Drillon et al., 2015, 2016) ou dans la première Partie de ce chapitre, les NIEBs étaient identifiés comme des zones déplétées en nucléosomes. Or, on voit, sur les **Figures 3.6** et **3.7**, que cette affirmation n'est exacte que pour les niveaux de digestion élevés. En effet, sur ces deux figures, on observe une forte diminution du signal d'occupation nucléosomale à l'intérieur des barrières pour les courbes rouges et jaunes, correspondant aux digestions complète et moyenne. En revanche, pour les conditions de digestion plus légères (en bleu et en vert), la diminution est nettement moins marquée (**Figure 3.6**), absente (**Figure 3.7 - courbe verte**), voire même remplacée par une augmentation (**Figure 3.7 - courbe bleue**). Ainsi, lorsque la chromatine est fortement digérée, on retrouve beaucoup moins de nucléosomes dans les barrières nucléosomales qu'en moyenne dans le reste du génome. En revanche, lorsque la digestion est plus ménagée, on en retrouve autant, voire plus, qu'en moyenne dans le reste du génome. Il semble dont qu'à l'instar de ce qui avait été observé chez la levure (Xi et al., 2011), on ait identifié ici des NDRs dans lesquelles sont en réalité formés des nucléosomes. Ces derniers sont plus sensibles à la MNase que la moyenne, ce qui indique une accessibilité importante, et rend invisibles ces nucléosomes dans les données de MNase-seq standards où la digestion et trop importante pour permettre de les observer. La présence de nucléosomes sensibles à la MNase à l'intérieur des NIEBs est confirmée par le calcul du paramètre MACC à ces loci. En effet, ce paramètre a des valeurs négatives à la fois chez l'humain et la souris (respectivement -0.205 et -0.15), indiquant des nucléosomes accessibles à la MNase (**Partie 3.3.1**). On note que cette accessibilité est également présente dans les autres lignées cellulaires de souris (**Figure A.8**).

Aux bords des barrières nucléosomales, l'analyse des données de MNase-seq standards chez l'humain et la souris a mis en évidence des nucléosomes bien positionnés, en accord avec le positionnement prédit à partir de notre modèle physique de positionnement nucléosomal (**Partie 3.2**). L'analyse des données MACC-seq confirme ce postulat, et le calcul du paramètre MACC indique également que ces nucléosomes bien positionnés sont moins sensibles à la MNase que la moyenne. En effet, le paramètre MACC est positif au niveau du premier nucléosome chez l'humain et la souris (respectivement 0.052 et 0.031), indiquant une certaine résistance à la digestion par la MNase, signature d'inaccessibilité au nucléosome (**Partie 3.3.1**). Aux bords des NIEBs, les nucléosomes semblent donc plus stables, moins accessibles qu'à l'intérieur des NIEBs. Ici encore, cette observation est également présente dans les autres lignées cellulaires de souris (**Figure A.8**). Lorsque l'on s'éloigne des barrières nucléosomale, le calcul du paramètre MACC retourne un résultat proche de

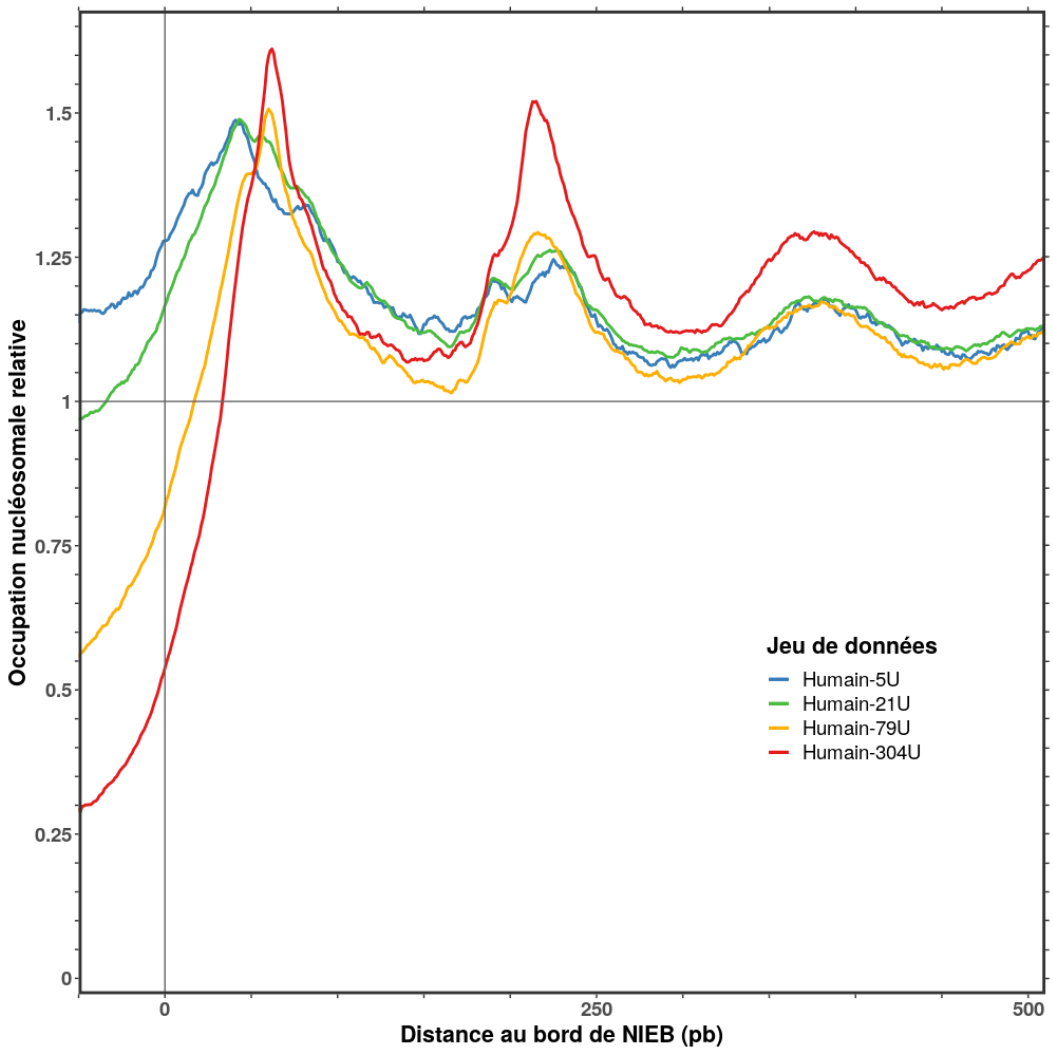


FIGURE 3.6 – **Profils d'occupation en nucléosomes aux bords des barrières nucléosomales chez l'humain selon le niveau de digestion de la chromatine.** Les données utilisées sont résumées dans le **Tableau 3.3**. Les profils obtenus en utilisant le pipeline décrit en **Partie 3.2.1** sont normalisés par la moyenne génomique. Les courbes rouge, jaune, verte et bleue correspondent respectivement aux digestions complète, moyenne, légère et très légère. Les paramètres MACC, coefficients de la régression linéaire calculés selon la méthode décrite en **3.3.1** sont −0.205, 0.052, 0.017 pour les positions -35 pb, 65 pb et 450 pb, respectivement.

0. En effet, à 450 pb des barrières nucléosomales, le paramètre MACC est légèrement positif chez l'humain (0.017) et dans deux des lignées cellulaires de souris (0.017 et 0.016 respectivement pour les lignées eNPC et E14 ESC), et très légèrement négatif dans les deux autres lignées cellulaires de souris (−0.008 et −0.0007 respectivement pour les lignées NPC et J1 ESC). On utilise des signaux qui ont été normalisés par la moyenne génomique. Ainsi, observer un paramètre MACC proche de 0 pour un locus indique qu'à ce locus, quelle que soit la digestion (et donc l'accessibilité des nucléosome), on retrouve le même rapport entre quantité de nucléosomes au locus et moyenne génomique. En d'autres termes, cela veut dire qu'il n'y a pas d'excès d'un certain type de nucléosomes par rapport à un autre. Par contre, la quantité de nucléosomes globale peut être supérieure à la moyenne. C'est ce qu'on observe en s'éloignant des NIEBs chez l'humain et la souris. Les quatre profils d'occupation en nucléosomes y ont des valeurs supérieures à 1 (entre 1.15 et 1.20 pour la souris, 1.10 et 1.25 pour l'humain, **Figures 3.6** et **3.7**). Cela indique qu'à ces loci, il y a légèrement plus de nucléosomes qu'en moyenne sur le génome. Lorsqu'on s'éloigne des NIEBs, il semble donc

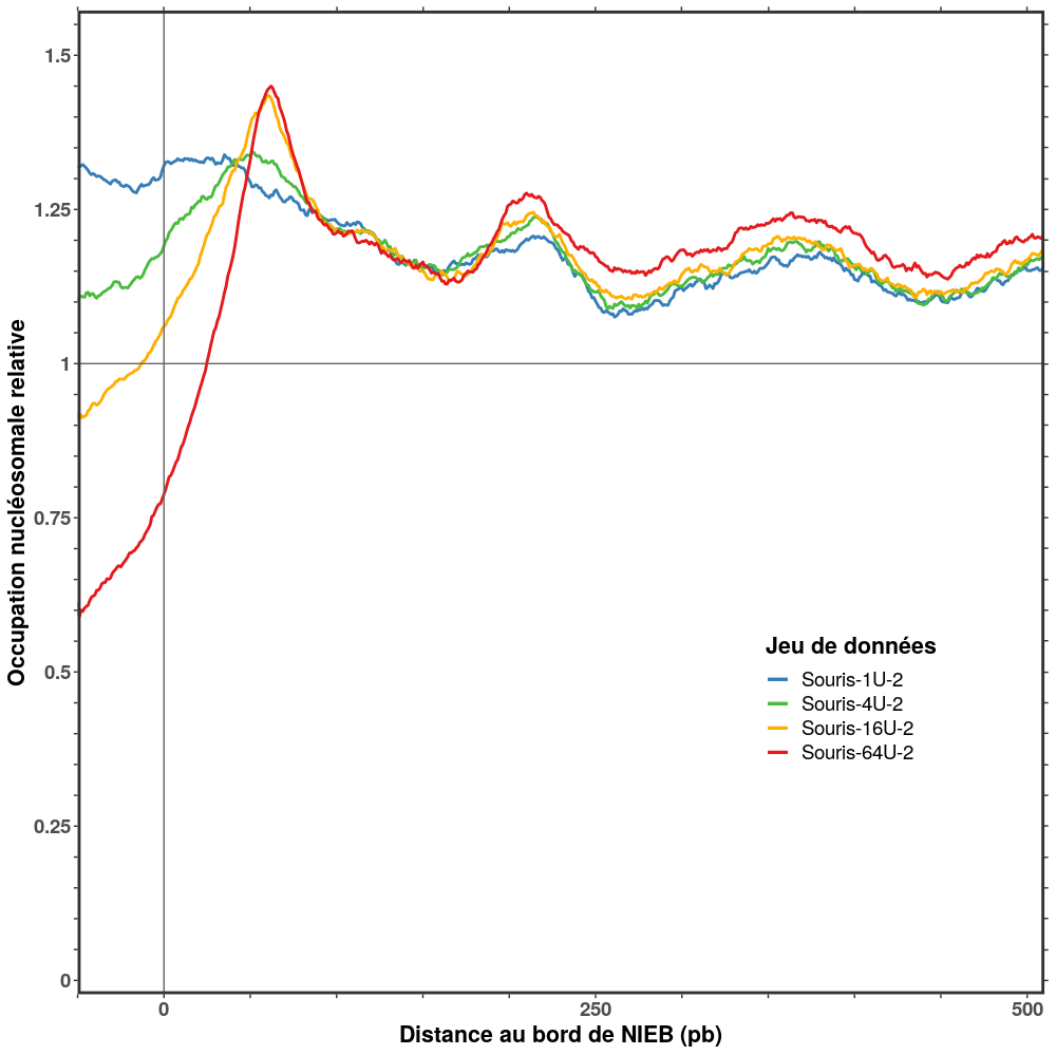


FIGURE 3.7 – **Profils d'occupation en nucléosomes aux bords des barrières nucléosomales chez la souris selon le niveau de digestion de la chromatine.** Les données utilisées sont résumées dans le **Tableau 3.3**. Les profils obtenus en utilisant le pipeline décrit en **Partie 3.2.1** sont normalisés par la moyenne génomique. Les courbes rouge, jaune, verte et bleue correspondent respectivement aux digestions complète, moyenne, légère et très légère. Les paramètres MACC, coefficients de la régression linéaire calculés selon la méthode décrite en **3.3.1** sont −0.15, 0.031, et 0.016 pour les positions -35 pb, 65 pb et 450 pb, respectivement.

que chez l'humain et la souris, on ne retrouve plus un excès de nucléosomes accessibles (comme dans les NIEBs) ou inaccessibles (comme aux bords des NIEBs), mais plutôt une composition en nucléosome proche de la moyenne génomique. En d'autres termes, loin des NIEBs, il ne semble pas que les nucléosomes aient de caractéristiques particulières chez l'humain et la souris.

Les résultats obtenus avec les données de MACC-seq apportent une nouvelle compréhension de la chromatine au niveau des barrières nucléosomales humaines et de souris. Les analyses de MNase-seq standards dans ces deux espèces ont mis en évidence que les barrières nucléosomales correspondaient *in vivo* à des zones déplétées en nucléosomes, et qu'aux bords de ces barrières, on retrouvait deux à trois nucléosomes bien positionnés (**Partie 3.2**). D'après les résultats obtenus avec les données MACC-seq, il semble que des nucléosomes soient en fait formés dans les NIEBs, mais que ceux-ci soient accessibles et sensibles à la MNase, ce qui explique qu'ils aient été absents des données de MNase-seq standards. Aux bords des barrières nucléosomales en revanche, il semble que les nucléosomes soient plutôt résistants à la MNase, indiquant une inaccessibilité qui, couplée à

l'occupation nucléosomale importante observée à ces loci, peut être liée à une certaine stabilité des nucléosomes (Mieczkowski et al., 2016). La dichotomie "zone avec ou sans nucléosome" déduite des expériences de MNase-seq standards semble donc à préciser, il y a plutôt une distinction entre des zones où les nucléosomes sont peu accessibles (aux bords des NIEBs) et des zones où les nucléosomes sont accessibles (dans les NIEBs). Cette nouvelle description des barrières nucléosomales comme des "zones d'accessibilité au nucléosome" pose de nouvelles questions quant à l'association des NIEBs avec d'autres facteurs, particulièrement les facteurs épigénétiques. En effet, l'accessibilité aux nucléosomes pourrait faciliter des mécanismes tels que l'échange d'histones pour remplacer des protéines histones canoniques par des variants d'histones ayant des propriétés épigénétiques particulières. Par exemple, le variant d'histone H3.3, qui a été associé à la fois à des régions de chromatine actives et à des régions réprimées (Szenker et al., 2011), a également été associé à de l'instabilité nucléosomale (Henikoff et al., 2009). Les barrières nucléosomales, de par l'accessibilité aux nucléosomes pouvant s'y former et qui faciliterait l'échange d'histones, pourraient alors être des points d'entrée pour les modifications épigénétiques, ce qui expliquerait leur présence ubiquitaire chez les génome eucaryotes (**Chapitre 2**). Pour étayer cette hypothèse, j'ai voulu savoir si les NIEBs pouvaient effectivement être associés à des variants d'histones, plus particulièrement le variant H3.3. J'ai analysé un jeu de données (pas encore publié) produit par l'équipe de K. Padmanabhan à l'IGFL, issu d'une expérience de digestion de chromatine par de la MNase suivie d'une immunoprécipitation pour sélectionner spécifiquement les nucléosomes contenant le variant d'histone H3.3 chez la souris. On peut donc assimiler cette expérience à du MNase-seq standard, à la différence qu'on ne verra ici que les nucléosomes contenant H3.3 (contre l'ensemble des nucléosomes dans une expérience de MNase-seq sans étape de ChIP). Malheureusement, une seule quantité de MNase a été utilisée pour établir ce jeu de données, aussi il n'est pas possible ici d'évaluer directement l'accessibilité aux nucléosomes qui seront observés. On pourra néanmoins voir si les nucléosomes contenant le variant H3.3 sont enrichis au niveau des barrières nucléosomales, et éventuellement voir si leur occupation suit celle de l'ensemble des nucléosomes.

3.3.3 Chez la souris, on observe un enrichissement du variant d'histone H3.3 au niveau des barrières nucléosomales

Pour étudier le lien entre le variant d'histone H3.3 et les barrières nucléosomales chez la souris, j'ai utilisé le pipeline décrit en **Partie 3.2.1** avec les données de MNase-ChIP-seq décrites ci-dessus. Cela m'a donc permis d'établir la courbe d'occupation des nucléosomes contenant l'histone H3.3 aux bords des barrières (**Figure 3.8**), ainsi que la distribution des z-scores associés à chaque barrière nucléosomale lors du test de l'enrichissement en nucléosomes vis-à-vis des régions adjacentes (**Figure 3.9**).

Le profil d'occupation des nucléosomes contenant H3.3 (**Figure 3.8**) est différent de ceux précédemment observés chez la souris avec les données de MNase-seq standards (**Figure 3.5 - A**). Tout d'abord, on n'observe pas de claire diminution de l'occupation à l'intérieur de la barrière comme ça a pu être le cas précédemment, même si on note que l'occupation diminue légèrement en dessous de la moyenne génomique. Il semble donc que les nucléosomes contenant l'histone H3.3 soient moins déplétés à l'intérieur des barrières nucléosomales que ne le sont les nucléosomes

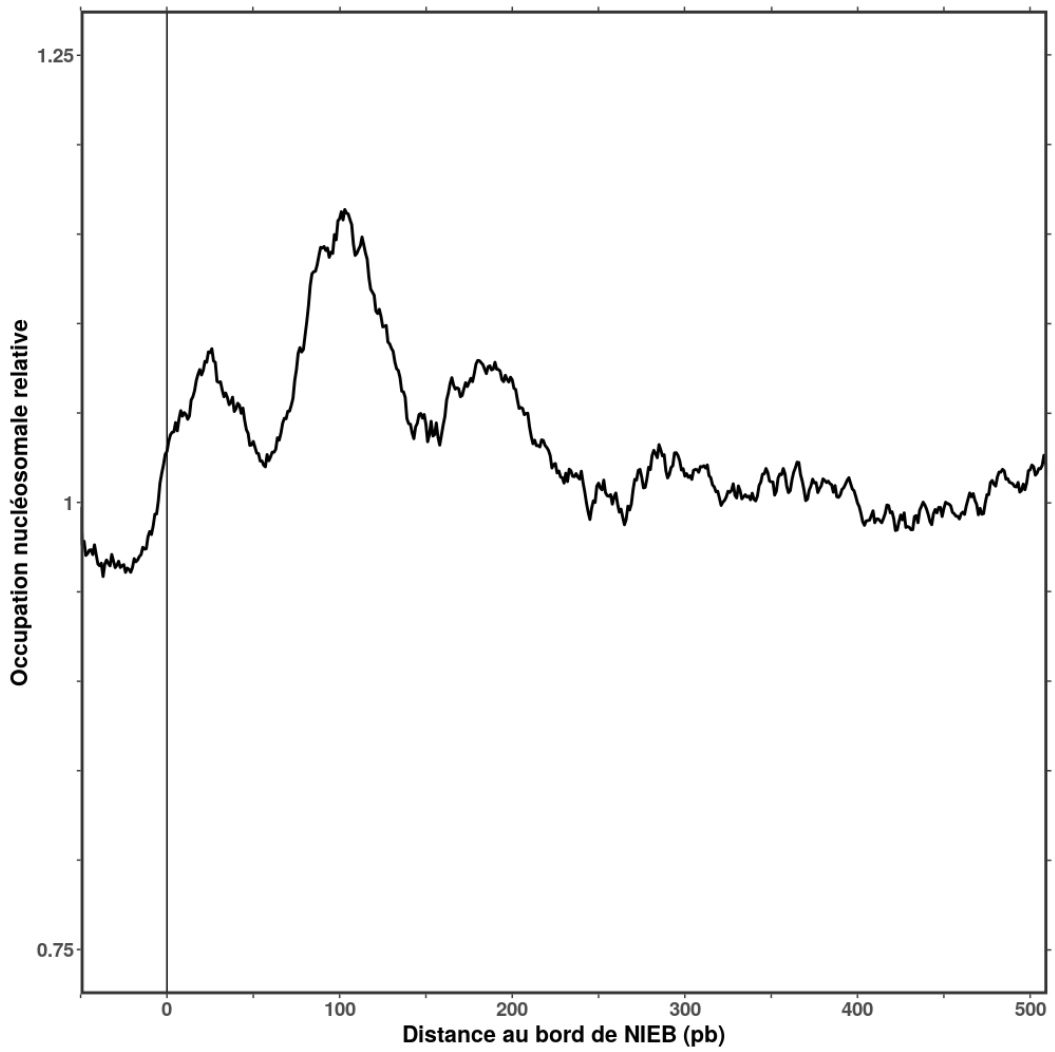


FIGURE 3.8 – **Profil d'occupation des nucléosomes contenant le variant d'histone H3.3 aux bords des barrières nucléosomales chez la souris.** Les données utilisées sont celles produites par l'équipe de Kiran Padmanabhan à l'IGFL (**Partie 3.3.3**). Le profil obtenu en utilisant le pipeline décrit en **Partie 3.2.1** est normalisé pour la moyenne génomique.

retrouvés en MNase-seq standard. Cette observation est confirmée par la distribution des z-scores (**Figure 3.9**). En effet, plus de la moitié des barrières nucléosomales (51.1 %) ne sont ni enrichies, ni déplétées en nucléosome. Le reste des barrières sont très majoritairement (48.7 %) déplétées en nucléosome, seule une infime partie étant enrichies (0.2 %). Cette distribution est clairement décalée vers la gauche du graphe par rapport à celle observée avec les données de MNase-seq classiques (**Figure 3.4 - A** et **A.7 - A**), où la très grande majorité des barrières sont identifiées comme déplétées en nucléosome quel que soit le jeu de données utilisé (77.8 % et 73.7 %). Même si on ne note pas d'enrichissement à proprement parler des NIEBs en nucléosomes H3.3 par rapport aux zones adjacentes (car la quantité de barrières rouges reste infime), on voit bien que le nombre de barrières ni enrichies, ni déplétées est en nette augmentation pour les nucléosomes H3.3. On a donc un enrichissement des nucléosomes contenant H3.3 au niveau des NIEBs par rapport aux nucléosomes observés en MNase-seq standard. On note cependant que la valeur $\overline{C}$ du jeu de données H3.3 est seulement de 1.54 lectures par nucléosomes. Comme on l'a vu précédemment, le nombre de NIEBs pour lesquels le test de déplétion est non-significatif a tendance à augmenter lorsque la valeur $\overline{C}$ diminue. De plus, le nombre de NIEBs considérés dans la distribution des

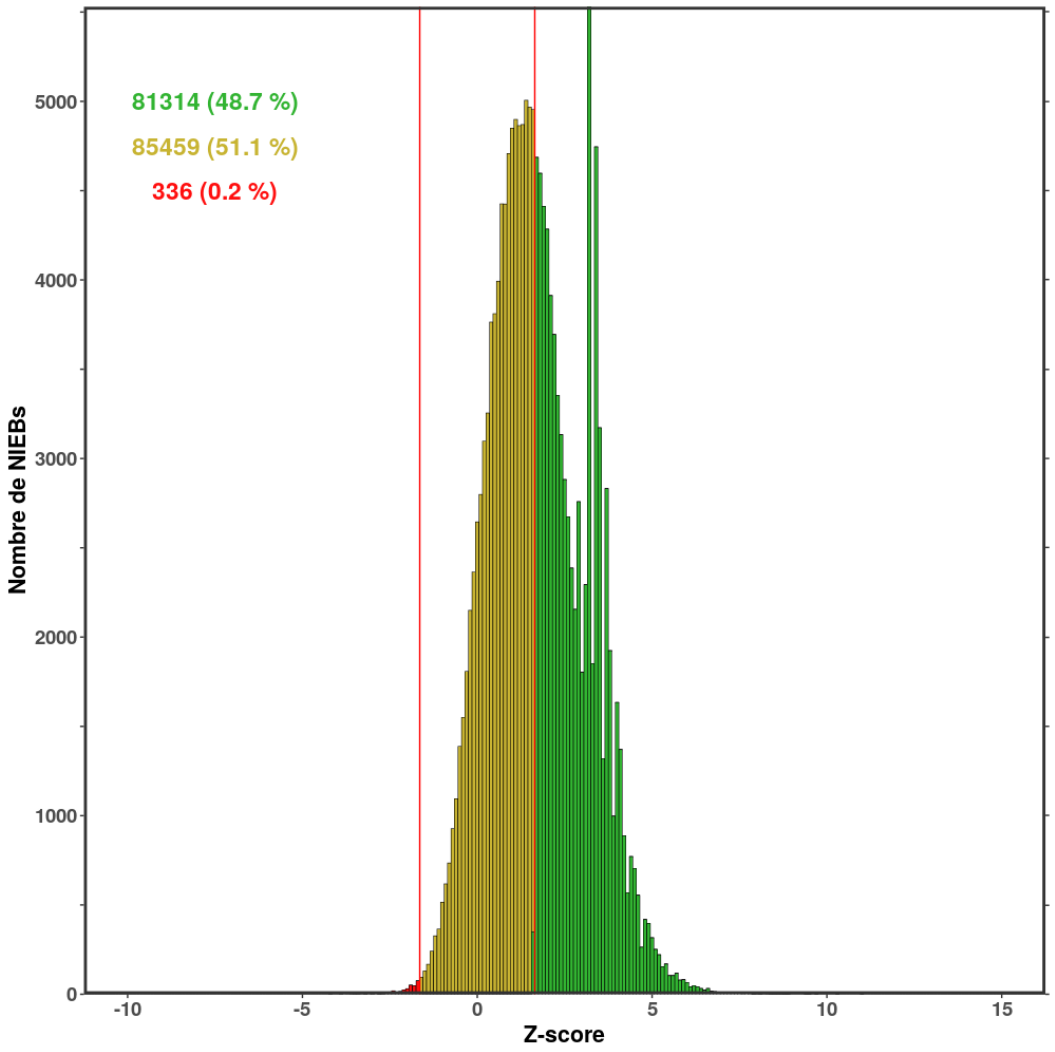


FIGURE 3.9 – **Distribution des z-scores des barrières nucléosomales obtenus avec les données H3.3 chez la souris.** Les données utilisées sont celles produites par l'équipe de Kiran Padmanabhan à l'IGFL (**Partie 3.3.3**). La distribution a été obtenue en utilisant le pipeline décrit en **Partie 3.2.1**.

z-scores est quatre à six fois inférieur à ceux considérés pour les mêmes distribution avec des jeux de données de MNase-seq standards (**Figures 3.4** et **A.7**). En d'autres termes, on a clairement dans le jeu de données H3.3 une faible couverture des nucléosomes qui peut être à l'origine de l'importante proportion de NIEBs identifiés comme ni déplétés ni enrichis en nucléosomes. Il convient donc d'être prudent dans l'interprétation de cette analyse à l'échelle des NIEBs individuels. En revanche, la couverture est tout de même acceptable pour travailler en moyenne sur une population importante de NIEBs, ce qui a été fait dans l'analyse de l'occupation nucléosomale (**Figure 3.8**). Sur cette figure, on observe un enrichissement des nucléosomes contenant l'histone H3.3 au niveau des barrières nucléosomales, à la fois par l'absence de déplétion à l'intérieur des barrières mais aussi par la présence de positions préférentielles aux bords des barrières. Le profil d'occupation nucléosomale (**Figure 3.8**), met en évidence des préférences de positionnement pour les nucléosomes contenant H3.3 aux bords des NIEBs. En effet, 3 pics apparaissent dans la courbe, aux positions 25 pb, 100 pb et 175 pb. On note qu'aucune de ces positions n'est compatible avec la position du premier nucléosome observée précédemment aux bords des barrières (60 pb), ni avec celle du second (210 pb). On note également que la position correspondant au pic le

plus haut (100 pb) n'est pas compatible avec les deux autres. En effet, avec environ 147 pb d'ADN enroulées autour des histones, deux nucléosomes positionnés à 25 pb et 100 pb d'une barrière seraient alors superposés, ce qui est impossible. En revanche, un nucléosome en position 25 pb est compatible avec un autre en position 175 pb si la séquence internucléosomale est très courte, ou que la longueur d'ADN enroulé autour des histones est inférieure à 147 pb. On remarque qu'un nucléosome positionné à 25 pb d'une barrière serait alors en partie formé à l'intérieur de cette barrière, pour environ 50 pb. On pourrait donc avoir deux configurations pour les nucléosomes contenant H3.3 aux bords des barrières nucléosomales. Un seul nucléosome pourrait être formé, à environ 100 pb de la barrière nucléosomale, ou bien deux nucléosomes cohabiteraient au bord de la barrière, au niveau des positions 25 pb et 175 pb, le premier étant également en partie formé à l'intérieur du NIEB.

L'analyse d'une expérience permettant d'étudier spécifiquement les nucléosomes contenant le variant d'histone H3.3 a mis en évidence que les barrières nucléosomales semblent associées à ce variant d'histone. L'occupation des nucléosomes contenant H3.3 aux bords des barrières leur est spécifique, ne reproduisant aucune des positions préférentielles précédemment observées chez la souris. Elle suggère également deux configurations nucléosomales distinctes aux bords des NIEBs. Il reste cependant à déterminer si les deux configurations déduites ici de la courbe d'occupation nucléosomale sont tour à tour formées aux bords des mêmes barrières nucléosomales, ou si les barrières ont une préférence pour l'une ou l'autre des configurations. Le calcul des z-scores illustrant l'enrichissement ou la déplétion en nucléosome H3.3 de chaque barrière nucléosomal a lui permis de mettre en évidence que ce type de nucléosomes semble plus présent à l'intérieur des NIEBs que les nucléosomes classiques, confirmant l'hypothèse d'une association entre les NIEBs et ce variant d'histone. Cette confirmation suggère une potentielle fonction des barrières nucléosomales dans les changements épigénétiques, il serait donc très intéressant de continuer ce travail en tentant d'associer les NIEBs à d'autres variants d'histones, ou marques épigénétiques, pour peut être élucider la fonction de ces barrières nucléosomales dans les différents génomes.

3.4 Contexte chromatinien et positionnement des nucléosomes aux bords des barrières nucléosomales

L'analyse des données expérimentales de type MACC-seq chez l'humain et la souris a mis en évidence, dans ces espèces, que les NIEBs ne correspondent pas à des régions déplétées en nucléosomes mais plutôt à des régions où les nucléosomes sont plus accessibles que dans le reste du génome. Aux bords des NIEBs en revanche, les nucléosomes ont été identifiés comme moins accessibles, et plus résistant à la MNase. Couplée à l'occupation nucléosomale élevée observée à ces loci, l'apparente inaccessibilité indique une stabilité des nucléosomes aux bords des NIEBs (Mieczkowski et al., 2016). Des données de MACC-seq ont également été produites chez la drosophile, dans des cellules de type S2 (Mieczkowski et al., 2016). J'ai donc pu reproduire mon analyse sur une troisième espèce, pour voir si les résultats concordaient avec les observations de la **Partie 3.3**. Le protocole utilisé a été en tous points semblable à celui ayant permis d'obtenir les résultats de la **Partie 3.3.2**. Les résultats obtenus sont présentés en **Figure 3.10**.

Chez la drosophile, la chromatine aux bords des barrières nucléosomales présente de nettes

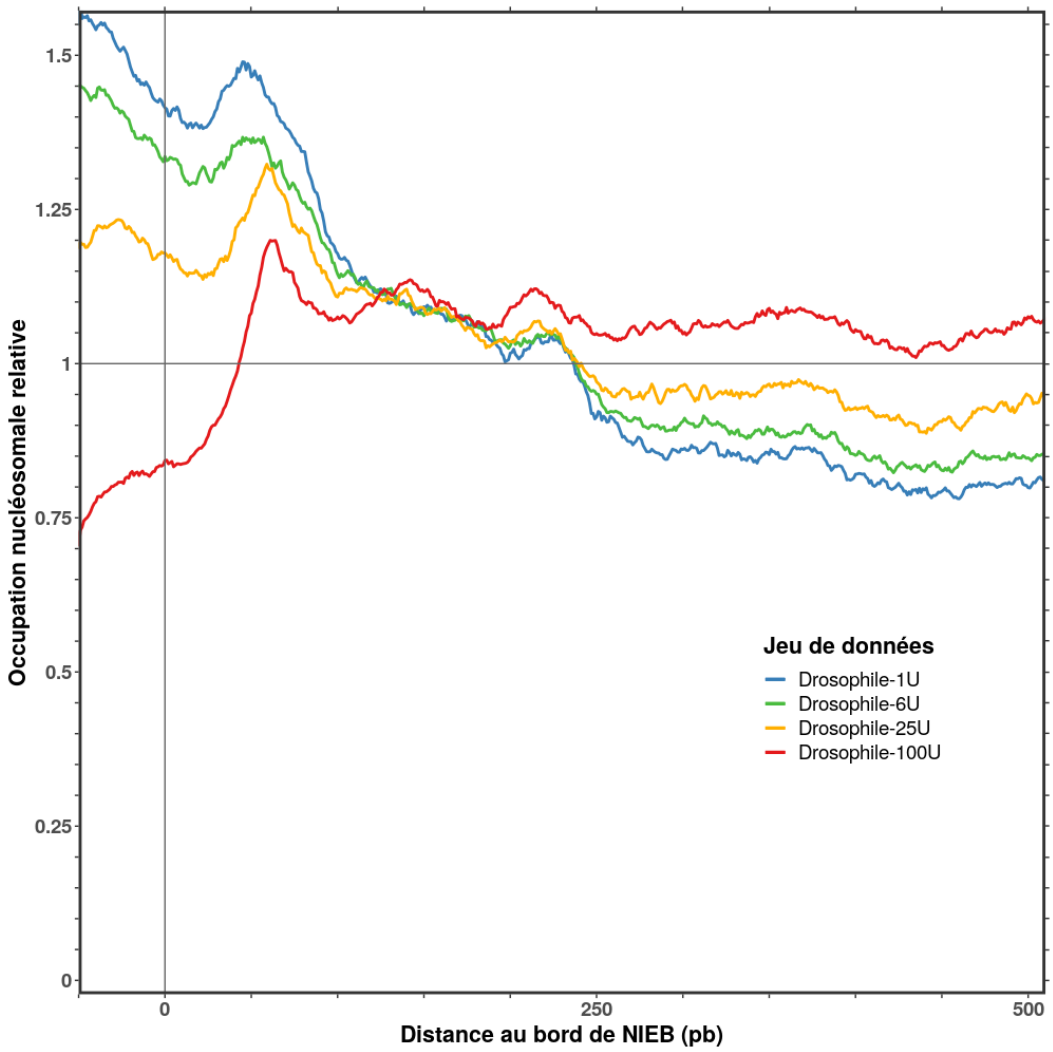


FIGURE 3.10 – **Profils d'occupation en nucléosomes aux bords des barrières nucléosomales chez la drosophile selon le niveau de digestion de la chromatine.** Les données utilisées sont résumées dans le **Tableau 3.3**. Les profils obtenus en utilisant le pipeline décrit en **Partie 3.2.1** sont normalisés pour la moyenne génomique. Les courbes rouge, jaune, verte et bleue correspondent respectivement aux digestions complète, moyenne, légère et très légère. Les paramètres MACC, coefficients de la régression linéaire calculés selon la méthode décrite en **3.3.1** sont −0.170, −0.048, 0.058 pour les positions −35 pb, 65 pb et 450 pb, respectivement

différences avec celle observée chez l'humain et la souris. A l'intérieur des barrières, on observe une sensible augmentation de l'occupation nucléosomale pour tous les niveaux de digestion excepté le plus fort par rapport au niveau observé à 500 pb des NIEBs. Notamment, pour les niveaux de digestion les plus faibles (**Figure 3.10 - courbes bleue et verte**), on observe ~ 1.5 fois plus de nucléosomes à la position −50 pb qu'en moyenne sur tout le génome. Cela correspond également pour ces deux courbes à l'occupation la plus forte observée sur la zone [−50 pb, 500 pb]. On voit aussi une claire anti-corrélation à l'intérieur des barrières nucléosomales entre l'occupation nucléosomale et la quantité de MNase utilisée pour la digestion, illustrée par un paramètre MACC de −0.17, ce qui indique un nucléosome très accessible à ce locus, comme observé précédemment chez l'humain et la souris. L'effet semble plus fort chez la drosophile, avec des courbes d'occupation nettement plus hautes à l'intérieur des barrières qu'à l'extérieur pour les niveaux de digestion faibles, ce qui n'avait été observé qu'une seule fois et dans une moindre mesure, dans une des lignées cellulaires de souris (**Figure 3.7 - courbe bleue**). Cette forme de courbe suggère même

du positionnement nucléosomal à l'intérieur des barrières, ce qui mériterait une étude à part entière qui ne sera pas menée ici. A l'extérieur des barrières, on observe de claires différences entre les résultats chez l'humain et la souris et ceux présentés ici chez la drosophile. Au niveau du premier nucléosome, on a précédemment détecté des signes de stabilité illustrés par un paramètre MACC positif au niveau de la position 65 pb associé à une forte occupation nucléosomale. Ici, on observe un paramètre MACC négatif (−0.048). En revanche, on observe toujours une occupation nucléosomale assez forte, entre 1.20 et 1.55 fois plus élevée que la moyenne génomique selon le niveau de digestion (**Figure 3.10**). D'après Mieczkowski et al., un paramètre MACC négatif couplé à une forte occupation nucléosomale est associée à de l'instabilité pour le nucléosome. Malgré cette apparente instabilité des nucléosomes à ce locus, on observe tout de même du positionnement, car les profils d'occupation en nucléosomes forment tous un pic au niveau de la position 65 pb. Ces résultats indiquent que chez la drosophile, le premier nucléosome aux bords des barrière est bien positionné, comme chez la souris et l'humain, mais qu'il est instable, contrairement à ce qui a été observé chez les deux autres espèces. A distance des barrières (450 pb), les nucléosomes chez l'humain et la souris ne présentaient pas de caractéristiques particulières en terme d'accessibilité, avec des paramètres MACC très proches de 0 (**Figures 3.6** et **3.7**). Chez la drosophile, on observe un comportement différent, avec un paramètre MACC qui s'élève à 0.058, soit un chiffre positif traduisant une résistance à la MNase des nucléosomes loin des barrières nucléosomales, signes d'inaccessibilité. On note d'ailleurs que cette inaccessibilité semble s'étendre entre les positions 250 pb et 500 pb, comme le montre le positionnement des courbes les unes par rapport aux autres sur cette zone dans la **Figure 3.10**. Ce résultat est différent de celui obtenu chez l'humain et la souris, où sur cette zone, les signaux ne montraient rien de particulier en terme de sensibilité des nucléosomes à la MNase (**Figures 3.6**, **3.7** et **A.8**).

Les résultats obtenus avec l'expérience de MACC-seq chez la drosophile mettent en évidence une structure nucléosomale différente dans cette espèce de celle qui a été observée chez l'humain et la souris. Chez la drosophile, on observe des nucléosomes plus accessibles que dans le reste du génome à l'intérieur des NIEBs et au niveau du premier nucléosome à leurs bords. Des nucléosomes moins accessibles sont en revanche identifiés lorsqu'on s'éloigne des barrières nucléosomales d'au moins 250 pb. C'est différent de la structure composée d'une zone accessible suivie d'une zone inaccessible puis d'une zone quelconque qu'on a observé dans les deux autres espèces étudiées. Pour comprendre ces résultats, il est intéressant de prendre en compte l'organisation des différents génomes étudiés. La densité en NIEBs est plus élevée chez la drosophile que chez l'humain et la souris (0.7393 contre 0.6334 et 0.6117, **Figure 2.1**). De même, la densité en gènes est très supérieure dans le génome de la drosophile. En effet, le nombre de gènes des génomes humain et de souris est estimé respectivement à environ 20 000-25 000 et ~ 30 000 (Pray, 2008), soit une densité moyenne d'environ un gène tous les 100 kb étant donnée la taille de ces génomes (**Tableau 2.1**). Chez la drosophile, cette densité est dix fois plus importante, avec environ 14 000 gènes (Pray, 2008) pour un génome ~ 20 fois plus petit, soit une densité d'environ un gène tous les 10 kb. Les barrières nucléosomales chez l'humain et la souris sont donc, mécaniquement, majoritairement positionnées dans des régions intergéniques, alors que celles de la drosophile sont majoritairement positionnées dans des régions géniques. Or, ces deux types de régions peuvent être très différentes d'un point de vue chromatinien, le premier type étant généralement associé à une chromatine plus fermée et moins accessible que le second. Cette importante différence de densité en gènes pourrait donc être

à l'origine des grandes différences d'organisation nucléosomales observées aux bords des NIEBs entre ces trois espèces. Il serait intéressant d'explorer plus avant cette possibilité en analysant des données de MACC-seq dans d'autres espèces, mais comme expliqué précédemment, ces données ne sont pour l'instant pas disponibles. En revanche, on pourrait essayer de différencier les NIEBs des différentes espèces en fonction de caractéristiques de la région dans laquelle ces barrières sont placées, comme la densité en gène, l'état transcriptionnel ou l'état chromatinien, afin de voir si on observe, au sein d'une même espèce, des changements d'organisation nucléosomale aux bords des NIEBs qui seraient liées à la densité en gènes ou d'autres contextes épigénétiques. Ainsi, dans l'analyse des nucléosomes aux bords des NIEBs, il semble nécessaire de tenir compte du contexte chromatinien dans lequel sont placées les NIEBs. En effet, celui-ci pourrait influencer à la fois le positionnement des nucléosomes et leur accessibilité aux bords des NIEBs, ce qui pourrait expliquer les différences observées entre les trois espèces étudiées ici.

3.5 Conclusion

Dans ce chapitre, j'ai analysé des profils expérimentaux de positionnement de nucléosomes afin de confronter les prédictions faites à partir de notre modèle physique de positionnement aux données *in vivo*. Pour faciliter l'obtention des résultats, j'ai mis au point et validé un pipeline automatisant les différentes étapes de l'analyse (**Partie 3.2.1**). J'ai ensuite utilisé ce pipeline sur des données de MNase-seq dans quatre espèces différentes, et pu confirmer la correspondance entre les barrières nucléosomales prédites par notre modèle et des régions déplétées en nucléosomes *in vivo* chez la souris, la drosophile, le poisson-zèbre et l'arabette. Afin de dépasser les limitations inhérentes à l'expérience de MNase-seq classique (**Partie 3.2.3**), j'ai utilisé un jeu de données issu d'une autre expérience, le MACC-seq (**Partie 3.3**). Ce type de données a permis d'étudier l'accessibilité des nucléosomes aux bords des NIEBs, et mis en évidence que la correspondance entre NIEBs et NDRs *in vivo* était principalement due à la sensibilité à la MNase des nucléosomes formés à l'intérieur des NIEBs. Chez l'humain et la souris, des nucléosomes peuvent en fait se former dans les NIEBs, mais ceux-ci sont très accessibles, ce qui les rend sensibles à la MNase et explique qu'on ne les ait jamais observé jusqu'ici. Aux bords des NIEBs, on observe en revanche des nucléosomes résistant à la MNase, stables et inaccessibles. Cette découverte a mené à une hypothèse quant à la fonction de ces barrières nucléosomales, qui pourraient, de par l'accessibilité de leurs nucléosomes, offrir des points d'entrée aux modifications épigénétiques. Les NIEBs pourraient par exemple faciliter les échanges d'histones. Cette hypothèse est d'ailleurs corroborée par l'observation d'un enrichissement des variants d'histones H3.3 au niveau des barrières nucléosomales par rapport aux nucléosomes standards (**Partie 3.3.3**). Pour comprendre complètement l'implication éventuelle des barrières nucléosomales dans les modifications épigénétiques, il est donc nécessaire de continuer cette analyse, par exemple en explorant systématiquement la relation entre les NIEBs et les marques épigénétiques. L'analyse des données de MACC-seq chez la drosophile a produit des résultats sensiblement différents de ceux obtenus chez la souris et l'humain. Cela souligne l'importance de la prise en compte du contexte chromatinien dans lequel sont placées les barrières nucléosomales. De futures analyses séparant les NIEBs selon les caractéristiques chromatiniennes des régions dans lesquelles ils sont placées, par exemple selon la densité en gènes, pourraient donc permettre de mieux comprendre l'organisation nucléosomale aux bords des NIEBs. Cela permettra

également d'appréhender l'importance fonctionnelle des NIEBs dans les génomes eucaryotes, voire de mettre en évidence des fonctions propres à certaines espèces, et in fine contribuer à expliquer l'ubiquité des barrières nucléosomales chez les eucaryotes observée dans le **Chapitre 2**.

4 Les éléments transposables Alu sont à l'origine de nouvelles barrières nucléosomales

Sommaire

4.1 Les éléments Alu induisent des inter-barrières de tailles spécifiques à l'humain

Dans le **Chapitre 2**, nous avons vu que le mécanisme de positionnement des nucléosomes par effet de parcage contre des barrières inhibitrices de la formation du nucléosome est conservé chez les eucaryotes. Cependant, on a observé que certaines tailles d'inter-NIEBs sont spécifiques de certaines espèces (**Figures 2.3** à **2.6**). Particulièrement, au niveau des inter-NIEBs ayant une taille compatible avec la formation de deux nucléosomes entre deux NIEBs, on observe que certaines tailles sont spécifiques des génomes de l'humain et du chimpanzé. Ces inter-NIEBs de tailles spécifiques à ces deux espèces sont légèrement plus grands que les inter-NIEBs de l'ensemble des autres espèces analysées, créant un décalage dans le pic correspondant sur la **Figure 2.6**. On retrouve d'ailleurs le même décalage au niveau des inter-NIEBs de tailles compatibles avec la formation de trois nucléosomes entre deux NIEBs. Comme évoqué dans le **Chapitre 2**, ces pics spécifiques à l'humain et au chimpanzé ont été associés aux éléments Alu. Sur la **Figure 4.1**, on retrouve les distributions de tailles des inter-NIEBs chez l'humain et le chimpanzé, établies exactement de la même manière que dans les **Figures 2.3** à **2.6**, à la différence que cette fois, les inter-NIEBs sont séparés selon la présence ou l'absence d'un élément Alu dans l'inter-NIEB.

Tout d'abord, on note que les distributions sont identiques entre les deux espèces, que ce soit pour les inter-NIEBs contenant des éléments Alu en bleu et turquoise ou pour ceux n'en contenant pas en rouge et rose. On observe également que les distributions correspondant aux inter-NIEBs ne contenant pas d'éléments Alu sont similaires à la distribution observée en **Figure 2.6** pour l'ensemble des espèces exceptée le porc. Enfin, le pic spécifique à l'humain et au chimpanzé que l'on a identifié sur la **Figure 2.6** correspond aux inter-NIEBs contenant un élément Alu sur la **Figure 4.1**. Les tailles d'inter-NIEBs spécifiques à l'humain et au chimpanzé sont expliquées par la présence d'éléments Alu dans ces inter-NIEBs. Cela pose la question de l'association entre les NIEBs et les éléments Alu. Chez l'humain et le chimpanzé, environ 41 % des inter-NIEBs contiennent un éléments transposable Alu. De plus, chez l'humain, plus de la moitié (52 %) de ces ETs ont été retrouvés flanquants un NIEB (Brunet et al., 2018). Ainsi, j'ai établi la distribution des éléments Alu aux bords des NIEBs, afin de préciser comment sont insérés les éléments Alu à ces loci.

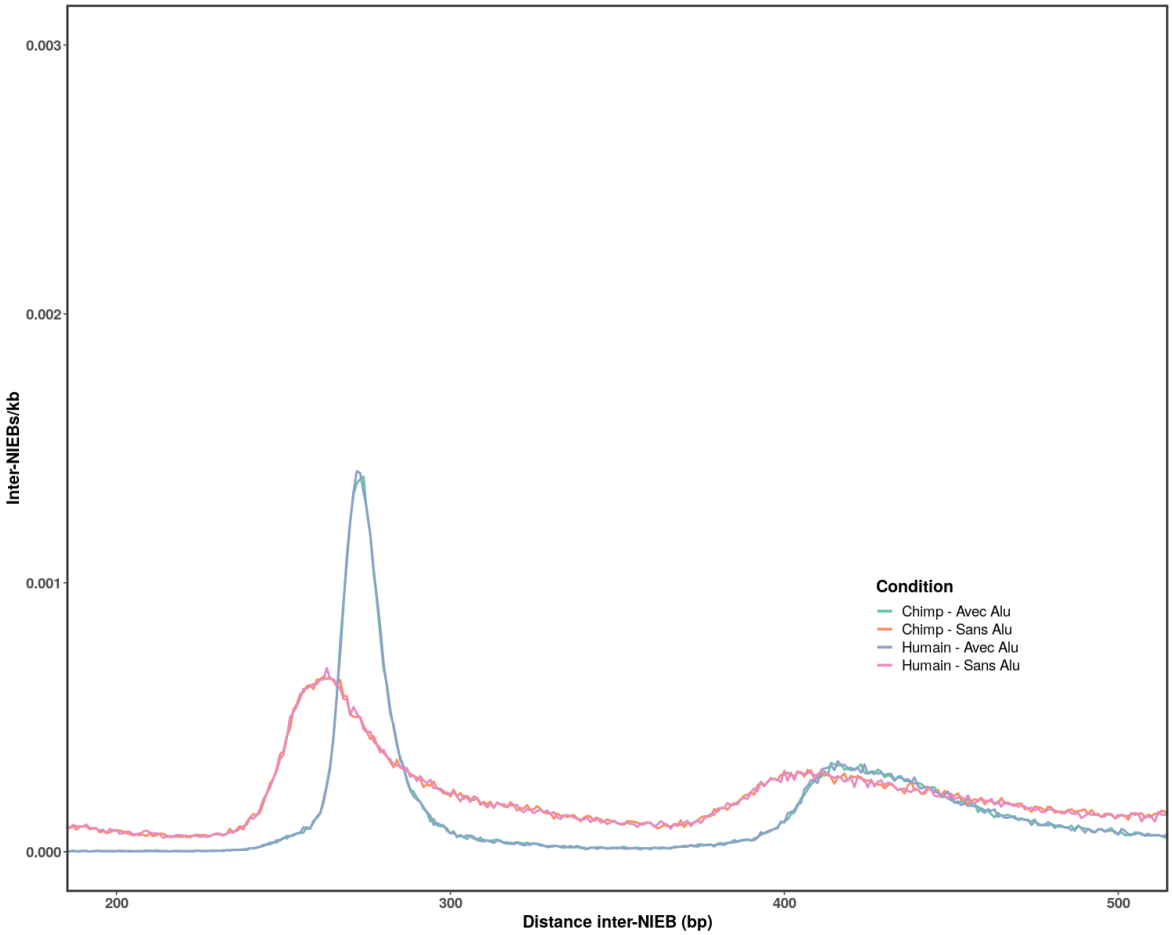


FIGURE 4.1 – **Distribution de tailles d'inter-NIEBs avec et sans éléments Alu.** Les courbes rouge et rose représentent respectivement la distribution de la taille des 1 024 137 inter-NIEBs chimpanzé et des 1 031 481 inter-NIEBs humains ne contenant pas d'éléments Alu. Les courbes de couleur turquoise et bleue représentent respectivement la distribution de la taille des 702 972 inter-NIEBs chimpanzé et des 714 061 inter-NIEBs humains contenant au moins un élément Alu. Ce graphe est un zoom sur les tailles 200 à 500 pb, pour reproduire les mêmes axes que dans la **Figure 2.6**

4.2 La distribution des Alu aux bords des NIEBs humain est non-triviale, et pourrait être expliquée par plusieurs hypothèses

Pour positionner les éléments Alu vis-à-vis des barrières nucléosomales, j'ai d'abord déterminé, pour chaque élément, une position de référence. Pour ceci, j'ai utilisé le polyA terminal de chaque élément. Ce polyA est de taille variable, de quelques à une trentaine de paires de bases. Certains éléments ont même perdu leur polyA. Afin d'éviter des biais liés à cette variabilité, j'ai choisi de prendre comme position de référence pour chaque Alu le milieu de son polyA. Pour les éléments n'en possédant pas, la position de référence se situe à l'extrémité 3' de l'élément. Enfin, pour les éléments insérés dans le brin anti-sens, la position de référence correspond au milieu du polyT présent au début de l'élément (ou à l'extrémité 5' de l'élément pour ceux ne possédant pas de polyT). J'ai ensuite positionné les éléments Alu par rapport aux barrières nucléosomales, en calculant la distance entre leur position de référence et le bord de barrière le plus proche. La **Figure 4.2** représente la distribution des positions de référence des éléments Alu aux bords des barrières nucléosomales. J'ai également effectué cette expérience chez le chimpanzé, où des

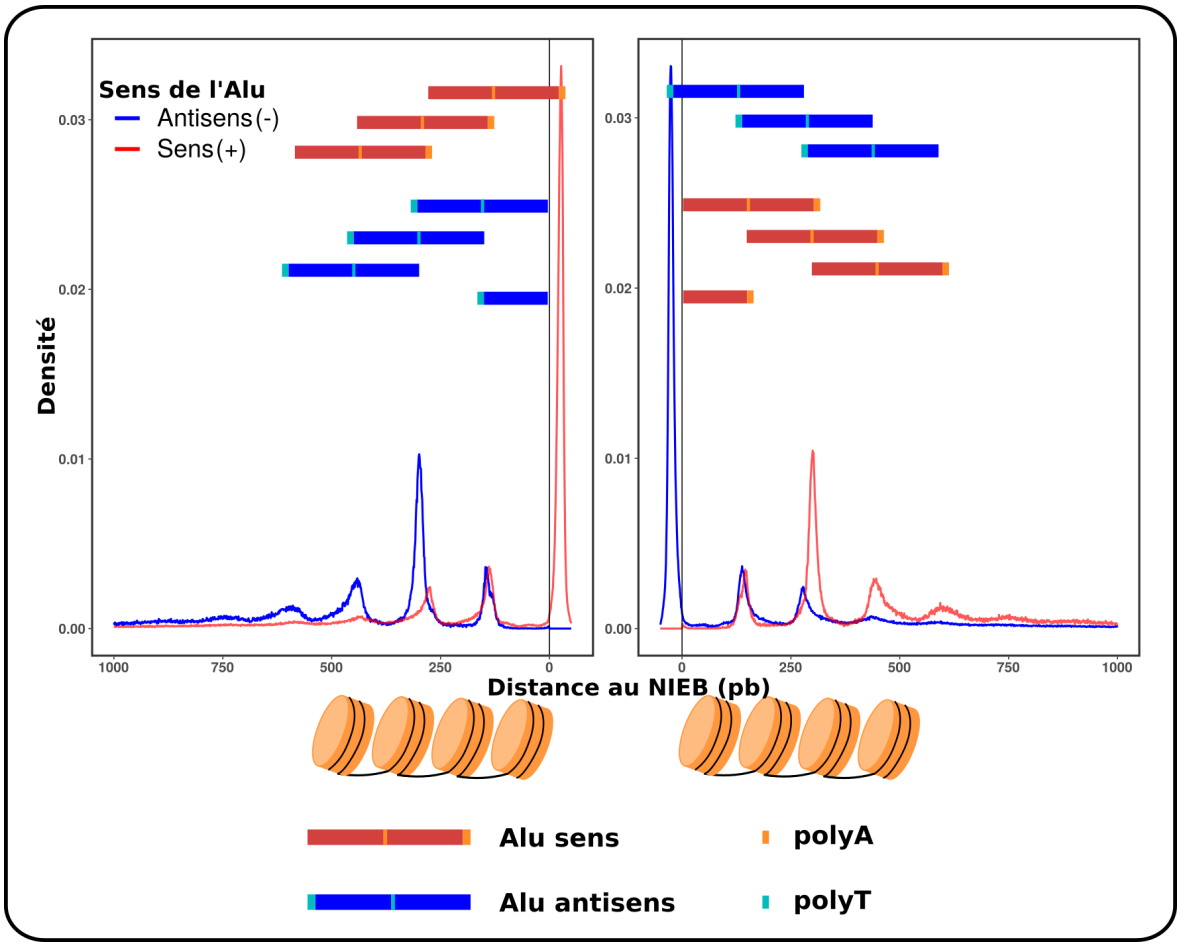


FIGURE 4.2 – **Distribution des** 1 078 322 **éléments Alu des autosomes humains aux bords droits et gauches des barrières nucléosomales.** Le 0 de la partie gauche (resp. droite) de la figure représente le bord gauche (resp. droit) des barrières nucléosomales. La courbe rouge de la partie gauche (reps. droite) représente les 395 076 (reps. 141 548) Alu identifiés sur le brin sens à gauche (resp. droite) d'un NIEB. La courbe bleue de la partie gauche (reps. droite) représente les 142 504 (reps. 398 946) Alu identifiés sur le brin antisens à gauche (reps. droite) d'un NIEB. Chaque courbe est normalisée par le nombre d'Alu de la catégorie correspondantes. Les nucléosomes positionnés par effet de parcage décrits en **Chapitres 1 à 3** sont schématisés sous le graphique.

résultats identiques ont été retrouvés et sont présentés en **Annexe A.9**.

On observe sur la **Figure 4.2** une symétrie entre les deux bords des barrières nucléosomales. Si l'on se place en amont d'une barrière (partie gauche de la figure), on observe une forte densité d'Alu sens flanquants le bord de barrière, avec le polyA terminal de l'Alu au niveau du bord interne de la barrière, et le corps de l'Alu au niveau du bord externe. Inséré de cette manière, on remarque 1. que les deux bras de l'élément Alu correspondent aux positions des deux premiers nucléosomes parqués contre les barrières nucléosomales ; 2. que le polyA interne de l'élément est au niveau du linker entre ces deux nucléosomes. On observe également deux autres positions préférentielles pour les éléments Alu sens en amont des barrières nucléosomales, aux distances 138 pb et 278 pb. Ici encore, les éléments sont placés de manière à ce que les deux polyA (interne et terminal) correspondent aux séquences inter-nucléosomales, et que les régions riches en GC correspondent aux séquences nucléosomales. On observe d'ailleurs le même phénomène avec les polyT des Alu anti-sens. En effet, une des positions préférentielles de ce type d'Alu en amont des barrières se situe à une distance de 300 pb du bord de la barrière. Insérés de cette manière, le polyT de l'Alu se retrouve entre le

second et le troisième nucléosome au bord de la barrière, et les deux bras de l'Alu correspondent aux séquences enroulées autour des deux premiers nucléosomes au bord de la barrière. On retrouve également des Alu anti-sens aux distances 446 pb et 593 pb, avec toujours les polyT au niveau des séquences inter-nucléosomales et les bras des Alu au niveau des séquences nucléosomales. Enfin, on observe qu'une partie des Alu anti-sens sont insérés avec leur polyT 145 pb en amont d'un bord de barrière. Un élément Alu canonique étant long de 280 pb (sans compter son polyA de taille variable), cela pourrait indiquer que la moitié de l'Alu serait alors à l'intérieur de la barrière nucléosomale. Cependant, si l'on concentre l'analyse sur les éléments Alu complets en ne prenant en compte que les éléments de tailles supérieures à 250 pb, ce pic disparaît (**Annexe A.10**). Après vérification, il apparaît que ce pic est constitué d'éléments incomplets, soit car l'un des deux bras de l'Alu a été perdu, soit car il s'agit d'un monomère ancestral de type FRAM ou FLAM. Ce pic correspond donc à des moitiés d'éléments Alu.

Sur la partie droite de la **Figure 4.2**, qui représente la distribution en aval des NIEBs, on observe les mêmes phénomènes qu'en amont concernant le positionnement des polyA, (sur le brin +, et des polyT sur le brin opposé) et des bras riches en GC des Alu par rapport aux nucléosomes. Ici encore, les Alu sont positionnés de manière à ce que leurs polyA (ou polyT) correspondent aux séquences inter-nucléosomales (bords des NIEBs et linkers), et à ce que leurs bras riches en GC correspondent aux séquences nucléosomales. Le positionnement des Alu anti-sens (resp. sens) en aval des NIEBs est parfaitement symétrique au positionnement des Alu sens (resp. antisens) en amont des NIEBs.

Les courbes de la **Figure 4.2** représentent des densités en éléments Alu, c'est-à-dire que le comptage à chaque distance a été normalisé par le nombre total d'Alu dans la catégorie correspondante. Ainsi, un fort pic dans la courbe rouge, comme observé au bord interne gauche des barrières nucléosomales, indique qu'une grande proportion des Alu sens en amont des barrières nucléosomales se situent à cette position. Ce positionnement place le polyA ou le polyT de l'élément Alu au bord interne de la barrière nucléosomale, et le corps de l'Alu au bord externe. Il est très préférentiel, avec 38 % de tous les éléments Alu des autosomes humains qui sont positionnés de cette manière. D'après les pics de la **Figure 4.2**, une autre position préférentielle est celle plaçant le polyA (ou polyT) à 300 pb d'un NIEB avec le reste de l'Alu positionné entre ce polyA et le NIEB. Cependant, ce cas de figure ne représente en réalité que 5.5 % des éléments Alu humains. En effet, on observe environ trois fois plus d'Alu sens en amont d'un NIEB (395 076) que d'Alu antisens (141 548). De la même manière, on observe trois fois plus d'Alu anti-sens en aval d'un NIEB (398 946) que d'Alu sens (142 504). Il semble donc que les éléments Alu soient globalement positionnés de manière à avoir leur polyA terminal faisant face à la barrière nucléosomale, avec environ 40 % des Alu ayant leur polyA terminal directement au bord interne des barrières, 6.7 % avec leur polyA à 138 pb d'une barrière et 5.2 % avec leur polyA à environ 278 pb d'une barrière. Ce sont donc près de 52 % des éléments Alu humains que l'on retrouve aux positions représentées par les trois Alu sens à gauche et les trois Alu antisens à droite de la **Figure 4.2**. Les autres positions représentées sur cette figure correspondent à 10.6 % des éléments Alu humains. En tout, ce sont donc près de deux tiers (62.6 %) des éléments Alu humains qui sont retrouvés proches d'une barrière, avec un positionnement plaçant les polyA et polyT entre les nucléosomes et les bras des Alu aux séquences nucléosomales. Ce positionnement est très éloigné de celui qu'on attendrait si les éléments Alu étaient distribués aléatoirement vis-à-vis des barrières nucléosomales. En effet, dans le cas d'une distribution aléatoire, on s'attendrait à une courbe plate pour la densité, sans positionnement

préférentiel. Il apparait donc des contraintes fortes sur le positionnement des éléments Alu vis-à-vis des barrières nucléosomales. Ces contraintes peuvent agir :

— Lors de l'insertion de l'élément Alu, par exemple en favorisant certaines séquences comme sites d'insertions. Cette hypothèse sera envisagée dans la **Partie 4.2.2**

— Suite à l'insertion, avec une possible sélection de certaines positions dans le génome par rapport à d'autres. Cette hypothèse sera envisagée dans la **Partie 4.2.1**

— Ni l'un ni l'autre, si c'est l'insertion d'Alu en elle-même qui apporte la barrière nucléosomale. Cette hypothèse sera envisagée dans la **Partie 4.2.3**

4.2.1 La perturbation du chapelet nucléosomal par l'insertion d'Alu hors des positions préférentielles pourrait être contre-sélectionnée

Les éléments Alu ont déjà été identifiés comme ayant un effet sur le positionnement nucléosomal (Tanaka et al., 2010). En effet, deux nucléosomes peuvent se former sur un élément Alu, un sur chacun de ses deux bras, séparés par un polyA correspondant à la séquence inter-nucléosomale (Tanaka et al., 2010). De plus, le positionnement des nucléosomes adjacents à l'élément est en phase avec le positionnement sur l'élément. Ainsi, l'insertion d'un élément Alu peut, en plus d'apporter une nouvelle séquence génomique, potentiellement modifier le chapelet nucléosomal au niveau du site d'insertion. Aux barrières nucléosomales, le chapelet nucléosomal a une forme bien définie, suivant l'effet de parcage des nucléosomes contre les séquences inhibitrices. Le positionnement contraint des éléments Alu aux bords des barrières (**Figure 4.2**) place les Alu de manière à ce que les nucléosomes se formant sur ces éléments soient parfaitement en phase avec le positionnement des nucléosomes par effet de parcage aux bords des NIEBs. Les insertions qui ne seraient pas compatibles avec ce positionnement sont absentes du génome humain. Elles pourraient donc être un évènement évolutif délétère, qui serait soumis à la sélection purificatrice. Dans ce scénario, les insertions seraient placées aléatoirement vis-à-vis des barrières nucléosomales (**Figure 4.3 - C**, panneau gauche). Certaines seraient en phase avec le positionnement nucléosomal pré-existant (**Figure 4.3 - A**). D'autres modifieraient le chapelet nucléosomal aux bords des barrières (**Figure 4.3 - B**). Post-insertion, celles modifiant le chapelet nucléosomal seraient contre sélectionnées, ce qui ne laisserait plus que les insertions en phase avec le positionnement nucléosomal aux bords des NIEBs. Cette hypothèse de sélection purifiante s'appuie sur le fait que l'encodage du positionnement nucléosomal via des séquences inhibitrices semble conservé entre les espèces, comme nous l'avons détaillé dans le **Chapitre 2**. Les NIEBs ainsi que le positionnement nucléosomal à leurs bords pourraient donc avoir une importance fonctionnelle pour les génomes, notamment concernant les modifications épigénétiques telles que l'échange d'histones (**Chapitre 3**). Cela pourrait expliquer que des insertions apportant des modifications importantes à l'organisation chromatinienne aux NIEBs soient délétères et donc contre-sélectionnées. Cette hypothèse de purification des insertions perturbant le chapelet nucléosomal aux bords des NIEBs est résumée dans la **Figure 4.3**. Il est possible de tester cette hypothèse en séparant les éléments Alu selon le moment de leur insertion dans les génomes (**Partie 4.3**).

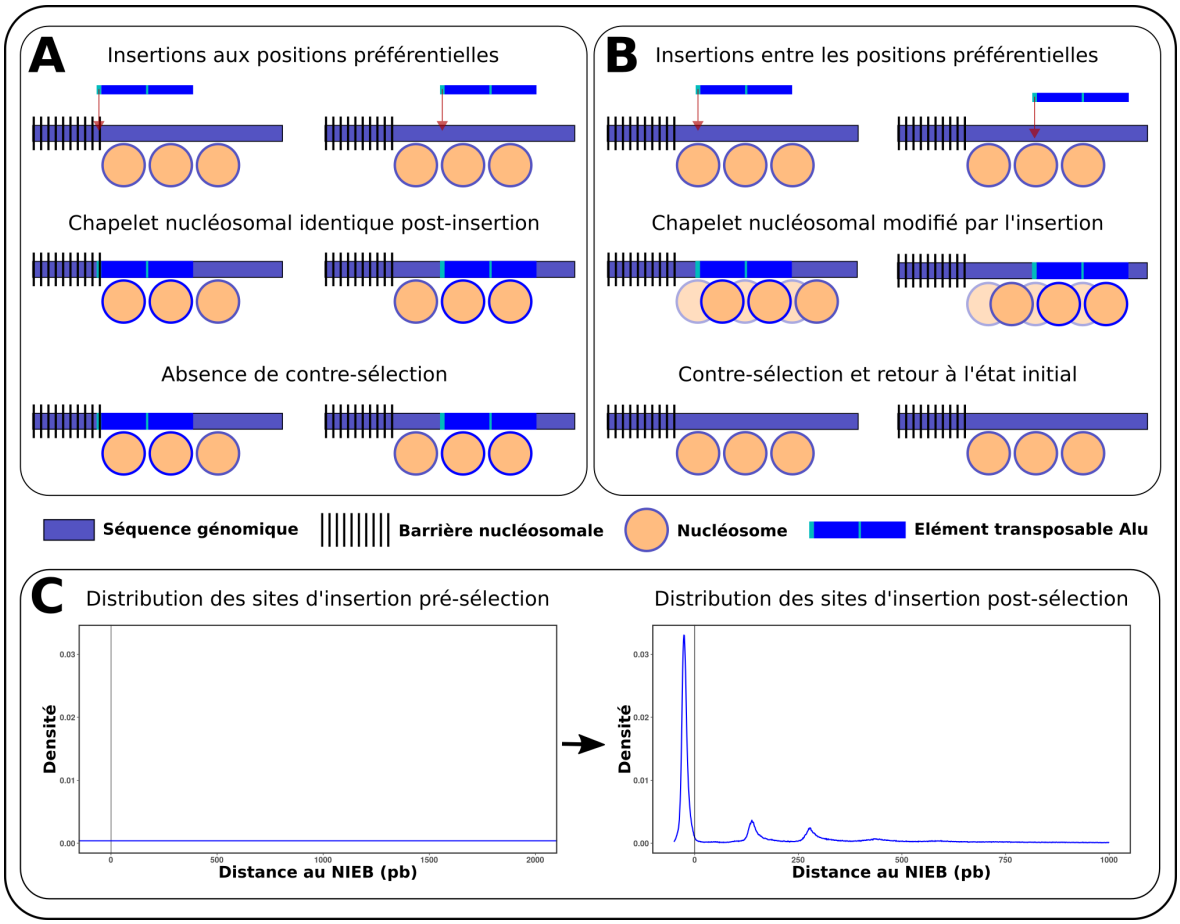


FIGURE 4.3 – **Schéma représentatif de l'hypothèse de sélection purifiante des insertions modifiant le positionnement nucléosomal.** Le panneau A représente deux exemples d'insertions où les nucléosomes formés sur l'élément Alu sont en phase avec ceux formés au bord de la barrière pré-insertion. Le panneau B représente deux exemples d'insertions hors-phase, où le positionnement nucléosomal est modifié par l'insertion. Le panneau C représente la distribution des sites d'insertion pré-sélection à gauche, et la distribution obtenue post-sélection à droite (identique à celle de la **Figure 4.2**).

4.2.2 Les polyA aux bords des NIEBs pourraient offrir des plateformes d'insertion pour les éléments Alu directement à leurs positions préférentielles

La distribution observée des Alu aux bords des NIEBs (**Figure 4.2**) pourrait résulter d'une distribution inhomogène des insertions de ces ETs. En effet, le mécanisme d'insertion des Alu implique la présence d'une séquence de type polyT dans le génome, qui est clivée par l'activité endonucléase d'ORF2p et à laquelle le polyA de l'élément Alu se fixe pour que l'élément soit rétro-transcrit (Deininger, 2011). Or, les séquences de type polyT et polyA sont sur-représentées aux bords internes des NIEBs dépourvus d'Alu ainsi qu'aux positions correspondants au premier linker (Brunet et al. (2018) et **Figure 4.4 - A**). Ces positions sont précisément celles où l'on retrouve majoritairement les éléments Alu dans le génome humain (**Figure 4.2**). La composition des séquences aux bords des NIEBs pourrait donc être directement à l'origine de la distribution des éléments Alu à ces loci, les polyA et polyT étant des plateformes d'insertion pour les éléments Alu. De plus, le positionnement nucléosomal aux bords des NIEBs place ces séquences entre les nucléosomes, les rendant potentiellement plus accessibles que les séquences nucléosomales, ce qui pourrait également faciliter l'insertion.

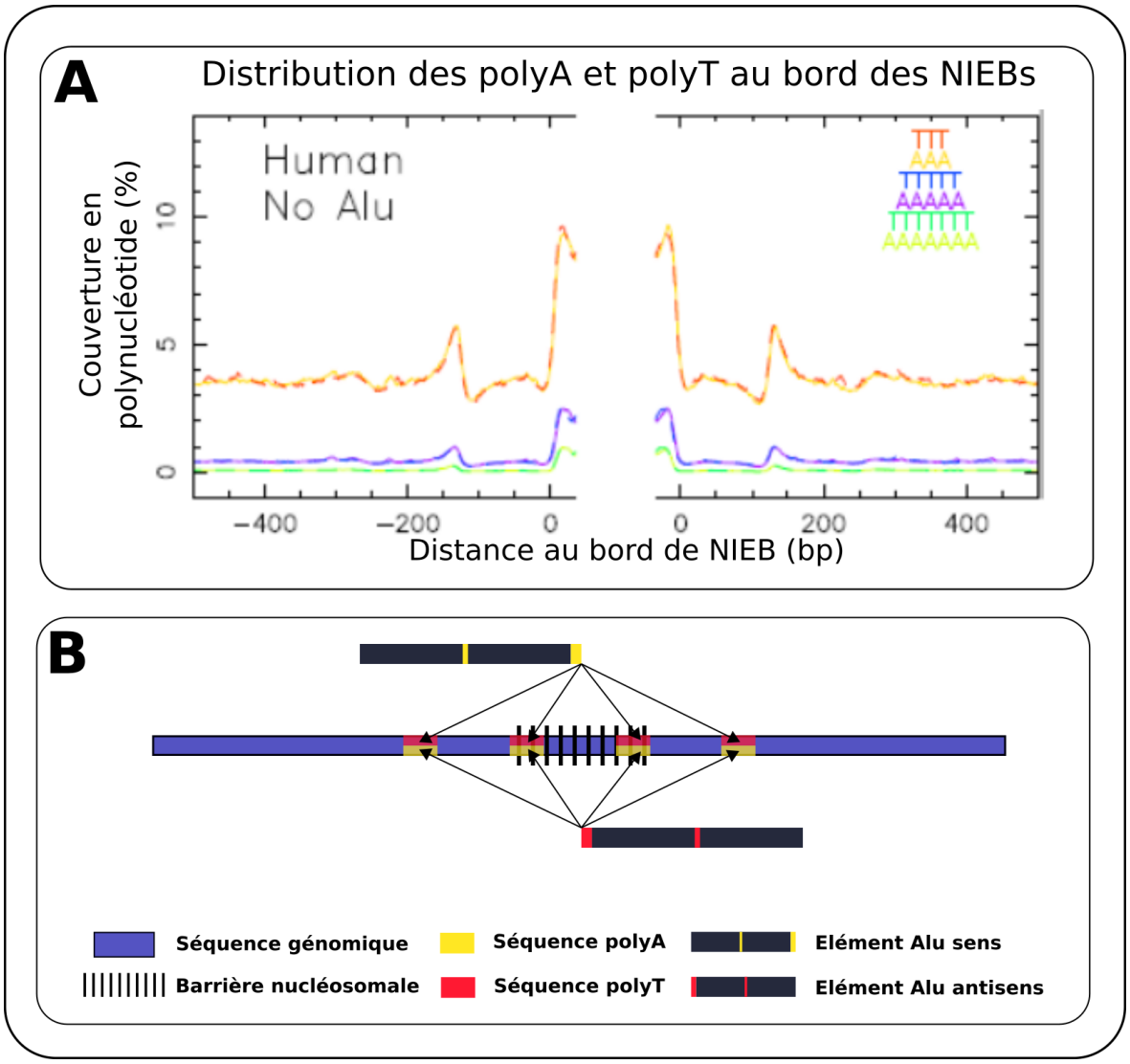


FIGURE 4.4 – **Schéma représentatif de l'hypothèse de mécanisme d'insertion.** Le panneau A représente la distribution des séquences polyA et polyT de taille 3, 5 et 7 pb aux bords des barrières nucléosomales dépourvues d'éléments Alu, adaptée de Brunet et al. (2018). Le panneau B est un schéma des possibilités d'insertion des éléments Alu selon leur sens et la présence de séquences polyA et polyT qui expliquerait la distribution observée de ces éléments aux bords des NIEBs.

Dans cette hypothèse, le mécanisme d'insertion est en cause pour expliquer la distribution des Alu aux bords des NIEBs. Il est à noter cependant que cette hypothèse n'est pas incompatible avec l'hypothèse de sélection développée en **Partie 4.2.1**. En effet, il est tout à fait possible que les insertions aient lieu majoritairement au niveau des séquences inter-nucléosomales et qu'il y ait également une purification des quelques insertions hors-phase modifiant le positionnement nucléosomal. En fait, la validation de ces deux hypothèses pourrait même expliquer le succès évolutif des éléments Alu durant l'évolution des primates. Pour tester l'hypothèse du mécanisme d'insertion comme origine du positionnement des éléments Alu aux bords des NIEBs, il est nécessaire d'étudier les insertions les plus récentes possibles afin de s'affranchir au maximum des contraintes sélectives. Cela a été réalisé en utilisant les données de polymorphisme d'insertions issues du Projet 1000 Genomes humains (Durbin et al. (2010), **Partie 4.4**)

4.2.3 Les Alu comme source de nouvelles barrières nucléosomales

La troisième possibilité pour expliquer la distribution des éléments Alu aux bords des barrières nucléosomales suppose la formation de nouvelles barrières nucléosomales par l'insertion d'un élément Alu (**Figure 4.5**). En effet, les éléments Alu, lors de leur insertion, provoquent l'apparition d'une séquence de type polyA à l'extrémité 3' de l'élément. Ces séquences sont connues pour être globalement absentes des séquences nucléosomales (Satchwell et al., 1986; Segal & Widom, 2009b; Struhl & Segal, 2013). On suppose que ces séquences sont particulièrement résistantes à la courbure nécessaire pour enrouler l'ADN autour des protéines histones pour former le nucléosome, et donc que leur présence inhibe la formation du nucléosome (Segal & Widom, 2009b). Cependant, cette explication n'a pas pu être confirmée expérimentalement, aussi la raison de l'inhibition de la formation des nucléosomes par les polyA reste sujet à controverse (Segal & Widom, 2009a).

Quel que soit le réel mécanisme impliqué, l'inhibition de la formation des nucléosomes par les séquences polyA a été observé et confirmé par de multiples expériences (Field et al., 2008; Satchwell et al., 1986; Segal & Widom, 2009a; Struhl, 1985). L'apport, par l'insertion d'éléments Alu, de polyA de plusieurs dizaines de paires de bases pourrait donc avoir d'importantes conséquences sur la formation des nucléosomes. De fait, en plus d'un polyA terminal inhibiteur de la formation du nucléosome, l'insertion d'un élément Alu apporte une séquence particulièrement compatible avec la formation des nucléosomes (Tanaka et al., 2010). En effet, un élément Alu possède non seulement un polyA terminal assez long et très inhibiteur du nucléosome, mais aussi un polyA interne qui peut lui aussi inhiber le positionnement nucléosomal. Ces deux polyA sont séparés par une séquence riche en GC (l'un des deux bras de l'Alu), favorable à la formation du nucléosome, et de taille ~ 150 pb, donc compatible avec les ~ 147 pb d'ADN enroulées autour des protéines histones pour former le nucléosome. Il a été observé que ce type de structure avec une séquence favorable flanquée de deux séquences inhibitrices a un effet particulièrement positionnant pour le nucléosome, en créant ce qui a été appelé un effet d'ancrage (Valouev et al., 2011). Un autre nucléosome peut se former sur les éléments Alu. En effet, le second bras de l'Alu, en amont du polyA interne, est également riche en GC, et d'une taille d'environ 150 pb, le rendant donc compatible avec la formation d'un nucléosome. L'insertion d'un Alu apporte donc une séquence avec les propriétés nucléosomales suivantes :

— Une inhibition forte de la formation du nucléosome à l'extrémité 3' de l'Alu, au niveau du polyA terminal.

— Un positionnement fort d'un nucléosome sur le bras adjacent à ce polyA terminal, par effet d'ancrage entre les deux polyA de l'élément

— Une inhibition de la formation du nucléosome par le polyA interne de l'élément.

— Le positionnement d'un second nucléosome sur le second bras de l'Alu, en 5', par effet de parcage contre le nucléosome ancré.

— Un potentiel positionnement du nucléosome en aval du polyA terminal de l'Alu, par effet de parcage au bord d'une séquence inhibitrice.

Après insertion d'un Alu, on a donc une séquence inhibitrice (le polyA terminal de l'Alu) flanquée de nucléosomes positionnés par effet de parcage, avec même un renforcement du positionnement par effet d'ancrage entre les deux polyA de l'élément Alu. Cette structure est précisément

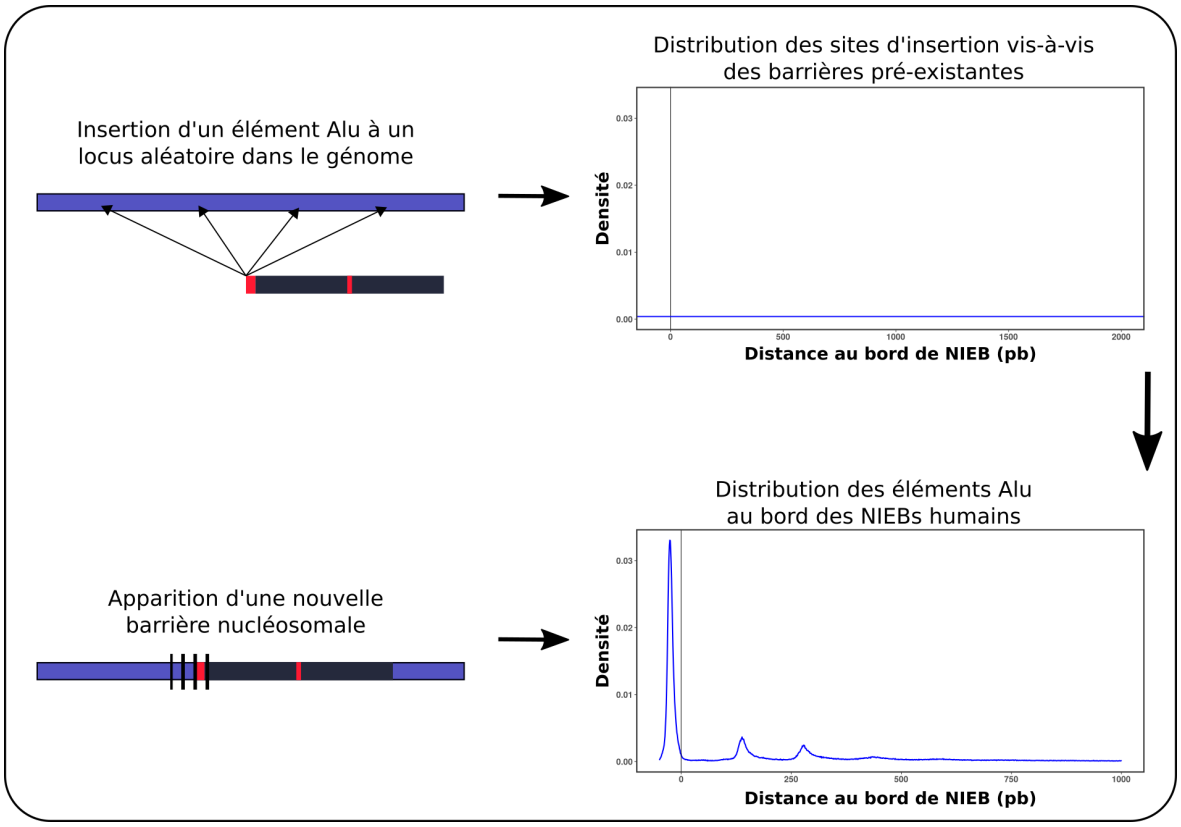


FIGURE 4.5 – **Schéma représentatif de l'hypothèse de création de NIEBs par l'insertion d'éléments Alu.** Dans cette hypothèse, un élément Alu s'insère aléatoirement dans une séquence génomique, ce qui se traduit par un positionnement aléatoire des sites d'insertion vis-à-vis des barrières pré-existantes (partie haute). L'insertion est à l'origine d'un NIEB par l'apport du polyA terminal (partie basse), ce qui explique le pic le plus fort de la distribution observée en **Figure 4.2**. Dans cette hypothèse, c'est le NIEB qui est formé au bord de l'Alu et non pas l'Alu qui est inséré au bord du NIEB.

celle décrite au niveau des barrières nucléosomales, avec une séquences inhibitrice flanquée de nucléosomes positionnés par effet de parcage contre cette barrière (**Chapitres 2** et **3**). L'insertion d'un élément Alu pourrait donc, principalement par l'apport de son polyA terminal, être à l'origine de nouvelles barrières nucléosomales. Si cette hypothèse est exacte, les barrières nucléosomales ainsi formées se retrouveraient de fait au niveau du polyA terminal des éléments Alu, ce qui placerait ces derniers en grande majorité de manière à avoir leur polyA terminal au bord interne d'une barrière. Cela pourrait donc expliquer en grande partie la distribution observée en **Figure 4.2**. On remarque cependant que les pics secondaires nécessitent que cette hypothèse soit affinée, par exemple pour inclure une dérive des polyA post-insertion pouvant amener à des Alu dont le polyA terminal ne forme plus de NIEB. On peut également imaginer des insertions d'éléments Alu en tandem pouvant décaler certains éléments de deux nucléosomes. Pour explorer l'hypothèse de formation de nouveaux NIEBs par l'insertion d'Alu, j'ai utilisé les insertions identifiées comme spécifiques aux génomes de l'humain et du chimpanzé par Tang et al. (Tang & Liang, 2019, 2020). Ainsi, j'ai pu comparer des sites d'insertion dans leurs états pré-insertion et post-insertion d'Alu pour déterminer si l'insertion est associée à la formation d'une nouvelle barrière nucléosomale(**Partie 4.5**).

4.3 Les insertions récentes d'Alu ont un positionnement plus contraint

Une des hypothèses pour expliquer la distribution des éléments Alu aux bords des barrières nucléosomales suppose une sélection purifiante des insertions destabilisant le positionnement nucléosomal (**Partie 4.2.1**). On peut tester cette hypothèse en tenant compte de l'âge des insertions d'éléments Alu. Dans le cadre d'une sélection purifiante des insertions hors phase par rapport aux NIEBs, on s'attend à ce que cette sélection soit plus efficace sur les insertions ayant eu lieu précocement dans l'évolution des primates, donc sur les Alu identifiés comme les plus anciens, simplement car le temps pendant lequel la sélection a joué est plus long. Ainsi, en suivant cette hypothèse, les éléments Alu les plus anciens devraient aussi être ceux pour lesquels les contraintes de positionnement aux bords des barrières sont les plus fortes.

Une manière de tenir compte de l'ancienneté des insertions Alu est de les séparer selon les sous-familles d'éléments. En effet, les éléments Alu sont classés en trois grandes sous-familles, à savoir, de la plus ancienne à la plus récente, les AluJ, les AluS et les AluY (Deininger (2011) et **Chapitre 1**). En plus de cette classification, j'ai ajouté, dans cette analyse, une catégorie correspondant aux éléments identifiés comme spécifiques au génome humain identifiés par Tang et Liang (Tang & Liang, 2019, 2020). Ces insertions ont donc eu lieu après la différenciation entre humain et chimpanzé, et peuvent être considérées comme les plus récentes insertions fixées chez l'humain. Il est à noter que les éléments Alu appartenant à cette dernière catégorie appartiennent principalement à la famille des AluY. En effet, les Alu actifs chez l'humain font principalement partie de cette famille, en particulier des sous-familles AluYa5 et AluYb8 (Deininger, 2011). Cependant, seulement 8259 insertions ont été identifiées comme spécifiques au génome humain, à comparer à plusieurs dizaines voire centaines de milliers d'éléments dans les différentes familles. La contribution apportée par les insertions spécifiques à l'humain aux résultats obtenus avec les familles d'Alu est donc négligeable.

J'ai catégorisé les éléments Alu en AluJ, AluS, AluY et Alu spécifiques à l'humain puis établi la distribution de chaque catégorie aux bords des barrières nucléosomales en suivant le même protocole que pour les distributions présentées en **Figure 4.2**. Afin de simplifier les interprétations, j'ai concentré mon analyse ici seulement sur les Alu insérés dans le brin anti-sens, et en aval d'une barrière nucléosomale. En effet, comme détaillé dans la **Partie 4.2**, les insertions positionnant le polyA terminal des éléments face à la barrière nucléosomale représentent la grande majorité des éléments insérés proches d'un NIEB. De plus, les distributions observées sont parfaitement symétriques entre les éléments insérés dans le brin sens en amont d'un NIEB et ceux insérés dans le brin antisens en aval d'un NIEB. Ainsi, on peut se concentrer sur un seul bord de NIEB et un seul sens d'élément Alu pour cette analyse. Les résultats obtenus pour les différentes catégories sont présentés en **Figure 4.6**.

Les résultats présentés sur la **Figure 4.6** vont totalement à l'encontre de ceux attendus dans le cadre de l'hypothèse de sélection purifiante (**Partie 4.2.1**). On observe que la proportion d'éléments Alu dont le polyA est situé au bord interne d'une barrière nucléosomale est bien plus importante pour les éléments les plus récents en rouge que pour les plus anciens en bleu. En fait, on remarque même un gradient dans la hauteur du premier et du second pic, avec une baisse de la proportion corrélée à l'ancienneté. En effet, près de 70 % des insertions spécifiques au génome humain sont

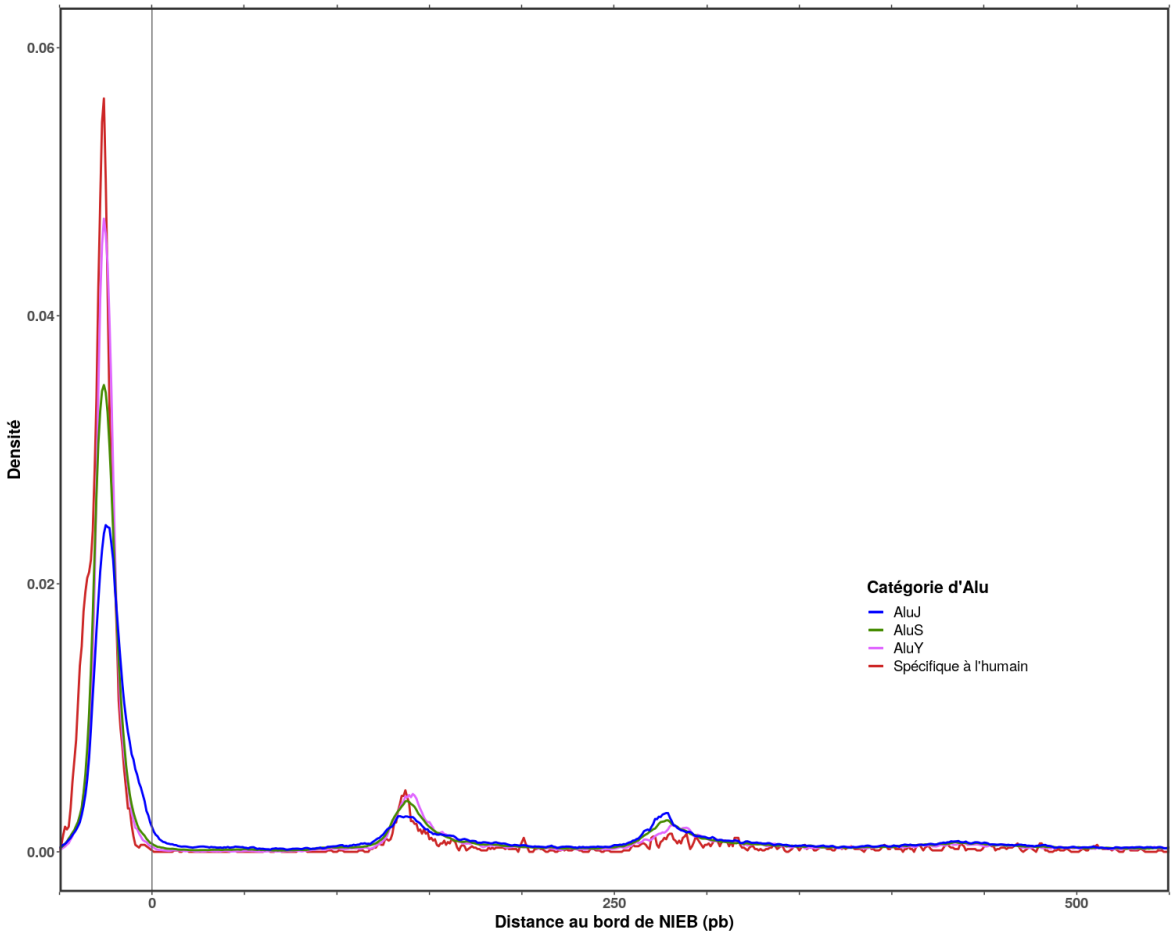


FIGURE 4.6 – **Distribution des différentes familles d'éléments Alu aux bords des barrières nucléosomales.** Sont représentés ici les distributions des 95 859 AluJ (en bleu), 234 121 AluS (en vert), 49 188 AluY (en rose) et des 3160 Alu spécifiques au génome humain, antisens, et situés à moins de 2000 paires de base d'une barrière. Les éléments Alu identifiés comme spécifiques au génome humain sont également inclus dans les autres catégories (la majorité étant également des AluY). 43 251 éléments dont la famille n'a pas été identifiée ou étant identifiés comme des FRAM ou FLAM ont été retirés de l'analyse. Enfin, la figure ne représente que les abscisses −50 à 550, les profils obtenus entre 550 et 2000 pb étant plats, et quasi égal à 0.

positionnées au niveau du premier ou du second pic (68.7 %), pour 54.3 % des AluY, 47.4 % des AluS et seulement 39.1 % des AluJ. Il semble donc que plus un élément est ancien, moins il est positionné au niveau des positions préférentielles identifiées en **Partie 4.2**. Cela va à l'encontre de l'hypothèse de sélection purifiante des insertions non-phasées avec le positionnement nucléosomal qui stipulait que les insertions anciennes devraient voir les contraintes sur leur positionnement être plus fortes. Ici, on observe l'exact opposé, excepté au niveau du troisième pic de la **Figure 4.6**. En effet, la hauteur de ce troisième pic semble positivement corrélée à l'âge des éléments Alu. Cependant, l'effet semble assez léger, les pourcentages d'AluJ, d'AluS, d'AluY et d'Alu spécifiques à l'humain contenus dans ces pics étant respectivement de 5.70 %, 5.39 %, 4.93 % et 3.50 %. En somme, les résultats obtenus en séparant les éléments Alu selon leur ancienneté dans le génome rejettent l'hypothèse de sélection purifiante développée en **Partie 4.2.1**.

4.4 Le positionnement des sites d'insertion de nouveaux Alu exclu l'hypothèse de plateforme d'insertion

L'hypothèse de sélection purifiante ayant été rejetée par les résultats obtenus dans la partie précédente, j'ai exploré la seconde hypothèse formulée en **Partie 4.2**. Dans cette hypothèse, on suppose que c'est le mécanisme d'insertion des éléments Alu et la distribution des polyA et des polyT aux bords des barrières qui sont à l'origine du positionnement des Alu aux bords des barrières observé en **Figure 4.2**. L'insertion de ces éléments se fait par l'appariement de leur polyA terminal a une séquence polyT pré-existante dans le génome (**Partie 4.2.2**). Or, on retrouve des polyT au niveau du bord interne des barrières nucléosomales et du premier linker, donc aux positions préférentielles pour les éléments Alu. L'hypothèse impliquant le mécanisme d'insertion des éléments Alu prédit donc que la distribution des Alu insérés *de novo* suivra celle des polyT aux bords des barrières nucléosomales, et donc a fortiori celle des éléments Alu déjà présents dans le génome observée précédemment. Pour tester cette hypothèse, j'ai positionné le site d'insertion des insertions d'éléments Alu les plus récentes possible dans le génome humain par rapport aux barrières nucléosomales. Pour se faire, j'ai utilisé les données du projet 1000 Genomes (Durbin et al., 2010). Dans ce projet, le génome d'un grand nombre d'individus de différentes populations a été séquencé afin de mettre en évidence les polymorphismes associés aux différentes populations. Parmi les données générées par ce projet, les insertions spécifiques de certains individus ou groupes d'individus ont été mises en évidence, et donc parmi elles les insertions d'ETs, incluant les éléments Alu. Il est donc possible de retrouver dans ces données les sites d'insertion de ces éléments absents du génome de référence et que l'on considère ici comme des insertions *de novo*. Il s'agit alors de voir si leur positionnement vis-à-vis des barrières nucléosomales humaines suit celui des Alu du génome. Les résultats obtenus par cette expérience sont présentés en **Figure 4.7**.

La distribution des Alu polymorphes, en vert sur la **Figure 4.7**, ne suit pas du tout celle des autres catégories d'Alu dans le génome humain, et donc pas celle des polyT aux bords des barrières nucléosomales. On observe un profil plus ou moins plat, traduisant des sites d'insertion qui semblent distribués aléatoirement vis-à-vis des barrières nucléosomales. On observe néanmoins une légère augmentation au niveau du bord interne des barrières nucléosomale, mais la densité reste bien plus faible que celle observée avec les éléments Alu du génome de référence humain (en bleu et rouge sur la **Figure 4.7**). En effet, seulement 10.8 % des sites d'insertions sont positionnés au bord interne d'une barrière nucléosomale. Ces résultats tendent à rejeter l'hypothèse développée en **Partie 4.2.2** selon laquelle le mécanisme d'insertion des éléments Alu expliquerait leur positionnement aux bords des barrières.

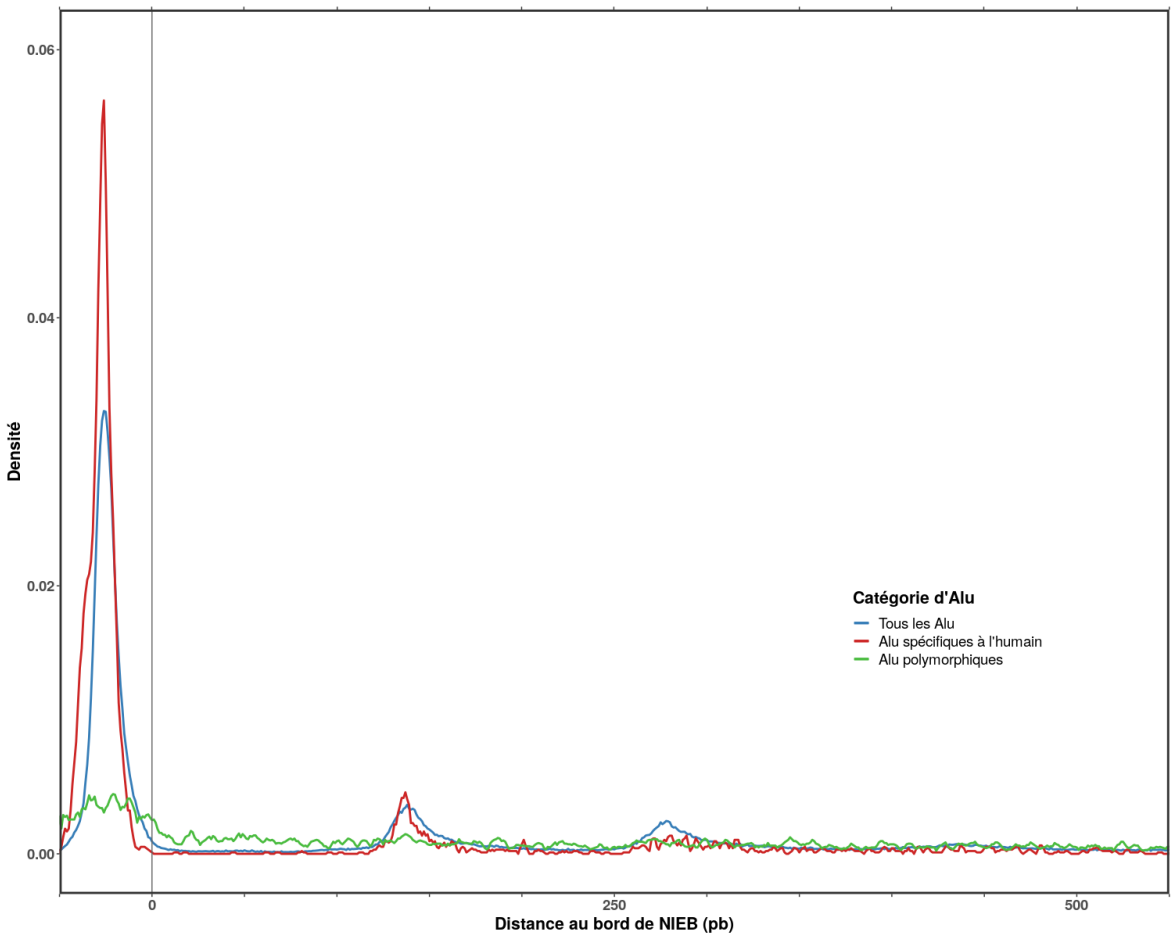


FIGURE 4.7 – **Distribution des sites d'insertions d'éléments Alu polymorphes issus du 1000 Genomes Project.** La courbe bleue représente la distribution des 392 807 éléments Alu du génome humain antisens et situés à moins de 2000 pb d'une barrière nucléosomale. Elle correspond à la courbe bleue de la partie droite de la **Figure 4.2**. La courbe rouge représente la distribution des 3160 éléments Alu identifiés comme spécifiques au génome humain, antisens, et situés à moins de 2000 pb d'une barrière. Elle correspond à la courbe rouge de la **Figure 4.6**. La courbe verte représente la distribution des 3610 sites d'insertion d'éléments Alu antisens identifiés dans les données du 1000 Genomes Project et situés à moins de 2000 bp d'une barrière nucléosomale.

4.5 L'insertion d'Alu est à l'origine de nouvelles barrières nucléosomales

4.5.1 Vers une hypothèse de formation de nouvelles barrières nucléosomales par l'insertion d'éléments Alu

Les résultats présentés en **Parties 4.3** et **4.4** ont non seulement invalidé les deux premières hypothèses formulées pour expliquer la distribution des éléments Alu observée aux bords des barrières nucléosomales, mais pas seulement. En effet, on peut également interpréter ces résultats comme des indices tendant à confirmer la troisième hypothèse. Celle-ci stipule que l'insertion d'un élément Alu, par l'apport d'un polyA terminal assez long (de plusieurs dizaines de paires de base), modifie le positionnement nucléosomal, y apportant des contraintes, notamment l'apparition de nouvelles barrières nucléosomales. Les sites d'insertion *de novo* d'éléments Alu pourraient donc être aléatoirement distribués vis-à-vis des barrières nucléosomales, comme on l'observe sur la **Figure 4.7**, et ce serait la création de nouvelles barrières nucléosomales au site d'insertion qui

résulterait en la contrainte apparente de positionnement des éléments Alu aux bords de celles-ci, comme observé sur la **Figure 4.2**. De plus, il a été démontré que le polyA terminal des éléments Alu est très sujet à dégradation post-insertion, notamment pour réduire le potentiel de transposition de l'élément (Deininger, 2011; Roy-Engel et al., 2002). Comme on suppose que c'est ce polyA qui serait à l'origine des barrières nucléosomales, sa dégradation progressive pourrait amener à la disparition de ces barrières nouvellement créées, ce qui pourrait expliquer le relâchement des contraintes de positionnement sur les Alu anciens par rapport aux Alu récents que l'on observe en **Figure 4.6**.

La distribution des sites d'insertion d'éléments Alu polymorphes présentée en **Figure 4.7** est également un indice fort vers l'hypothèse de formation de nouvelles barrières nucléosomales par l'insertion d'éléments Alu. En effet, sur ce graphique, trois catégories d'éléments Alu sont représentées : en bleu, tous les Alu du génome de référence humain ; en rouge, les Alu spécifiques au génome humain ; et en vert, les Alu polymorphes détectés chez certains individus grâce au projet 1000 Genomes. Les courbes rouge et bleue représentent donc des éléments Alu qui sont présents dans le génome de référence. A l'inverse, la courbe verte représente des éléments Alu qui sont absents du génome de référence. Ainsi, les éléments Alu représentés par les courbes rouge et bleue ont été pris en compte lors de la détection des barrières nucléosomales basée sur le modèle physique de positionnement de nucléosomes dépendant de la séquence. En revanche, les éléments Alu représentés par la courbe verte n'ont pas pu être pris en compte, car ils sont, par construction, absents du génome de référence. En somme, lorsqu'on prend en compte les éléments Alu dans la détection des barrières nucléosomales, on retrouve des barrières au niveau du polyA terminal de ces éléments. Ce phénomène disparaît lorsqu'on ne prend plus en compte les éléments Alu dans la détection des barrières. Ce résultat peut s'expliquer si les éléments Alu sont à l'origine des barrières nucléosomales. Les éléments polymorphes, absents du génome de référence, sont donc distribués plus aléatoirement vis-à-vis des barrières pré-existantes.

Pour résumer, le résultat obtenu avec les sites d'insertion d'éléments Alu polymorphes (**Figure 4.7**) semble indiquer que l'insertion d'Alu pourrait être à l'origine de nouvelles barrières nucléosomales. Cela serait provoqué par l'apport du polyA terminal des éléments Alu, ces séquences étant connues pour être fortement inhibitrices de la formation du nucléosome (Field et al., 2008; Satchwell et al., 1986; Segal & Widom, 2009a; Struhl, 1985). Post-insertion, la dégradation de ce polyA terminal pourrait amener ces NIEBs nouvellement formés à disparaître. Ainsi, les éléments Alu anciens, ayant des polyA plus dégradés que les éléments Alu récents, se retrouveraient alors moins souvent aux bords des barrières nucléosomales, ce qui expliquerait le résultat obtenu en séparant les éléments Alu selon leur famille (**Figure 4.6**). Les résultats des **Parties 4.3** et **4.4** peuvent donc être interprétés comme des preuves indirectes de l'hypothèse de formation de nouvelles barrières nucléosomales par l'insertion d'éléments Alu développée en **Partie 4.2.3**. Cependant, il est possible d'obtenir une confirmation directe de cette hypothèse, particulièrement en étudiant les insertions spécifiques au génome humain.

4.5.2 La comparaison des séquences pré et post-insertion d'élément Alu illustre la formation des barrières nucléosomales par l'insertion de ces éléments transposables

Cas 1 (1/1)	Cas 2 (1/0)	Cas 3 (0/0)	Cas 4 (0/1)	Total
572	2229	1831	96	4728
12.1 %	47.1 %	38.7 %	2.0 %	100 %

TABLEAU 4.1 – **Nombre de sites d'insertion d'Alu spécifiques au génome humain appartenant à un NIEB de l'humain (1/-) et/ou du chimpanzé (-/1).**

4.5.2.1 La moitié des insertions d'Alu spécifiques au génome humain sont à l'origine d'une nouvelle barrière nucléosomale

L'hypothèse développée en **Partie 4.2.3** stipule que l'insertion d'un élément Alu peut amener la formation d'une nouvelle barrière nucléosomale. Si certaines analyses semblent indiquer que cette hypothèse est exacte, il reste à obtenir une confirmation directe. Une manière de l'obtenir est de comparer une même région avant et après insertion d'un élément Alu, afin de voir si la présence des Alu est effectivement associée à la présence de nouvelles barrières nucléosomales. Pour obtenir ce jeu de données, j'ai utilisé les 8259 insertions d'Alu identifiées comme spécifiques au génome humain par Tang et al. (Tang & Liang, 2019, 2020). Pour chaque insertion, j'ai récupéré les 100 pb en amont et en aval de l'élément Alu, et aligné ces 2×100 pb sur le génome du chimpanzé. Cela m'a permis, pour chaque élément Alu spécifique au génome humain, de localiser le site d'insertion correspondant chez le chimpanzé. J'ai restreint mon analyse aux séquences parfaitement alignées entre les deux espèces, sans autre insertion que celle de l'élément Alu et sans délétion, en tout cas dans les 100 pb en amont et en aval du site d'insertion. En effet, il faut que les séquences adjacentes à l'élément Alu inséré soient identiques afin de s'assurer que c'est bien la présence de l'élément qui entraîne celle de la barrière et non un autre évènement mutationnel. Cela m'a permis d'obtenir un jeu de données de 4728 paires de sites pour lesquels un Alu est présent chez l'humain et absent chez le chimpanzé. J'ai ensuite positionné et calculé la distance de ces sites aux barrières nucléosomales prédites chez ces deux espèces. Quatre cas sont alors possibles :

1. Le site humain (avec Alu) et le site chimpanzé (sans Alu) sont tous les deux à l'intérieur d'une barrière nucléosomale. Il y avait donc une barrière nucléosomale à l'origine, dans laquelle un élément Alu s'est inséré.
2. Le site humain (avec Alu) est à l'intérieur d'une barrière nucléosomale, mais pas le site chimpanzé (sans Alu). L'élément Alu s'est donc inséré hors d'une barrière nucléosomale, et est à l'origine d'un nouveau NIEB.
3. Le site humain (avec Alu) et le site chimpanzé (sans Alu) ne sont ni l'un ni l'autre à l'intérieur d'une barrière nucléosomale. L'élément Alu s'est donc inséré hors d'une barrière nucléosomale, et n'en a pas formé de nouvelle.
4. Le site humain (avec Alu) est hors d'une barrière nucléosomale, mais le site chimpanzé (sans Alu) est à l'intérieur d'une barrière nucléosomale. L'insertion de l'élément Alu a donc fait disparaître une barrière nucléosomale.

Le nombre d'éléments Alu dans chacun des quatre cas est présenté dans le **Tableau 4.1**.

On observe que, parmi les 4728 insertions analysées, seule une petite partie ont été faites au niveau d'une barrière nucléosomale pré-existante (les cas 1 et 4, soit 14.1 %). C'est cohérent avec ce que l'on a mis en évidence dans la **Partie 4.4**, où 10.8 % des sites d'insertion d'éléments Alu polymorphes sont retrouvés au bord interne d'une barrière nucléosomale chez l'humain. Il

semble que ces insertions conservent les barrières nucléosomales, car dans 86 % de ces cas, on retrouve une barrière nucléosomale à la fois chez le chimpanzé et chez l'humain. La majorité des insertions d'Alu spécifiques à l'humain semblent cependant avoir été effectuées en dehors des barrières nucléosomales (cas -/0 : 85.8 %). Et dans ce cas, pour plus de la moitié des éléments (54.9 %), on retrouve chez l'humain une barrière nucléosomale, absente chez le chimpanzé. Il semble donc que lorsque les éléments sont insérés au niveau d'une séquence qui n'est pas une barrière nucléosomale, dans plus de la moitié des cas, une barrière nucléosomale apparaît au niveau du polyA terminal des éléments. Cependant, 45.1 % de ces éléments ont été insérés sans qu'une nouvelle barrière nucléosomale apparaisse. Il semble donc que la formation d'une nouvelle barrière nucléosomale par l'insertion d'un élément Alu soit dépendante de certaines conditions. Afin de préciser ces conditions, j'ai comparé les profils d'énergie de formation du nucléosome et d'occupation nucléosomale prédits par le modèle, à la fois aux sites d'insertions chez le chimpanzé et aux éléments Alu chez l'humain, pour chacun des cas énumérés ci-dessus. L'étude de ces profils permettra aussi de voir si l'absence de barrière nucléosomale pour une partie des insertions peut être due à des faux négatifs dans la détection de NIEBs, comme ce que l'on a pu observer dans le **Chapitre 2** lors de l'étude de la conservation des NIEBs entre l'humain et le chimpanzé. Les profils d'énergie produits dans chaque cas sont présentés en **Figure 4.8**. Les profils d'occupation nucléosomale sont en **Figure 4.9**. Pour interpréter ces profils, on peut considérer que le profil obtenu chez le chimpanzé correspond à l'état ancestral, pré-insertion d'Alu, et que celui obtenu chez l'humain correspond à l'état post-insertion.

4.5.2.2 Les profils d'énergie et d'occupation nucléosomale prédits par le modèle indiquent que les éléments Alu sont quasi-systématiquement à l'origine de nouvelles barrières nucléosomales

Pour le cas 1 (présence d'un NIEB à la fois à la position de référence de l'Alu chez l'humain et au site d'insertion correspondant chez le chimpanzé, **Figure 4.8 - A**), on observe que la présence d'un élément Alu change considérablement le profil énergétique. En effet, à l'état ancestral (en rouge), on observe une augmentation de l'énergie autour du 0, et une baisse progressive de celle-ci en s'éloignant de ce 0, pour arriver à un profil plat. L'augmentation s'explique par la présence d'une barrière nucléosomale au niveau du 0. Chaque site 0 peut tout aussi bien être au bord interne gauche, au bord interne droit ou au milieu d'une barrière nucléosomale, ce qui explique la forme en cloche de la courbe rouge autour du 0. En présence d'un élément Alu (courbe bleue), le profil énergétique change radicalement. On observe, comme chez le chimpanzé, une forte hausse de l'énergie autour de la position 0, correspondant ici au polyA terminal de l'élément. On note cependant que la hausse est plus importante chez l'humain que chez le chimpanzé, ce qui pourrait traduire d'un renforcement de la barrière nucléosomale par le polyA terminal de l'élément. Au niveau du corps de l'élément Alu, le profil énergétique baisse très fortement, avec une valeur minimale au niveau du premier bras de l'élément, suivie d'une remontée au niveau du polyA interne, et d'une nouvelle baisse, moins importante, au niveau du second bras de l'élément. Enfin, on observe un pic d'énergie de positionnement de nucléosome au niveau de la position -375 pb, donc en amont de l'élément. Ce pic pourrait correspondre à une partie de la barrière nucléosomale initiale qui aurait été décalée par l'insertion de l'Alu. On aurait alors une barrière initiale qui aurait

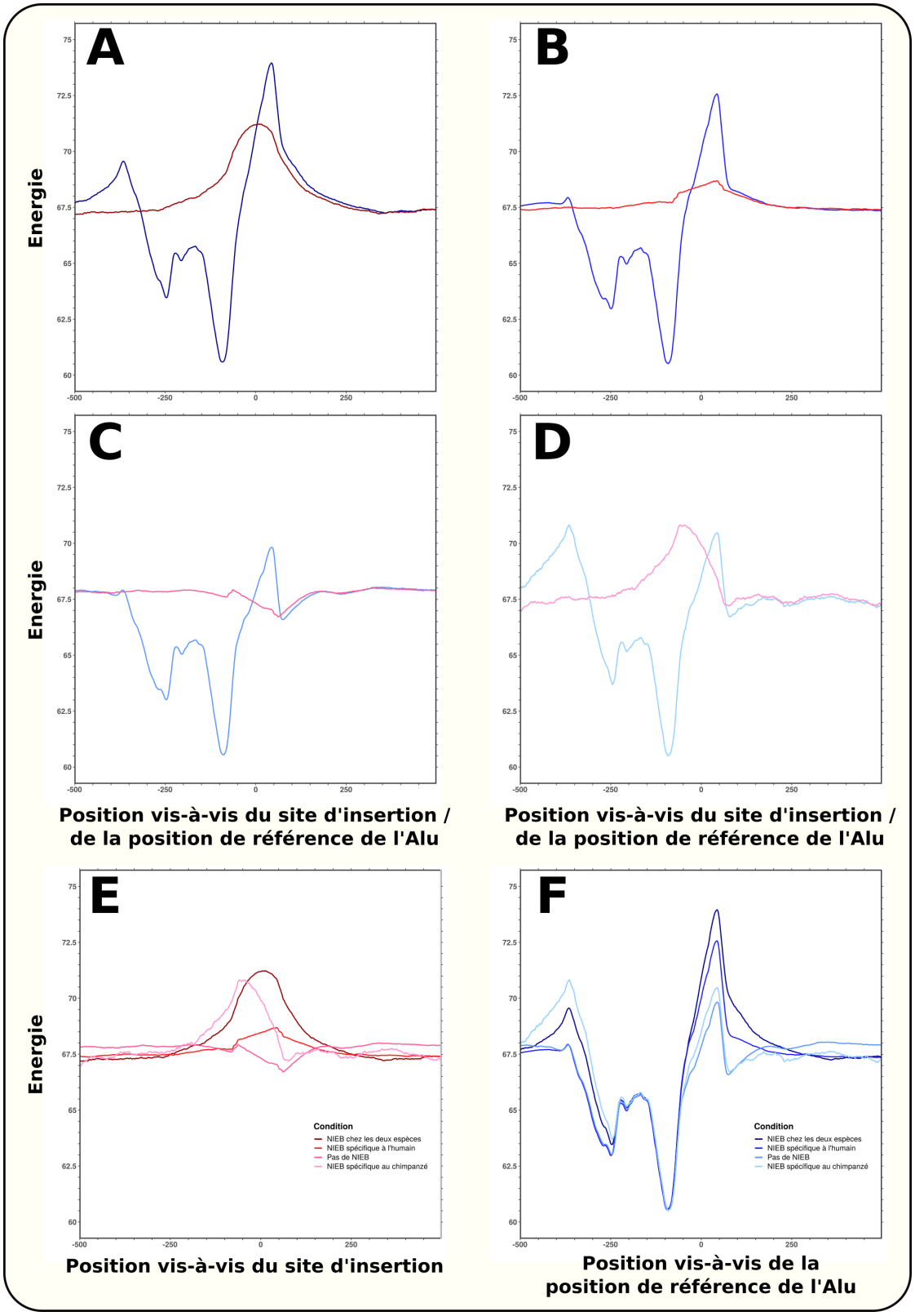


FIGURE 4.8 – **Profil énergétique des sites avec et sans Alu chez l'humain et le chimpanzé.** Les courbes rouges représentent les profils obtenus chez le chimpanzé, les courbes bleues les profils obtenus chez l'humain. Les panneaux A, B, C et D représentent respectivement les cas 1, 3, 2 et 4 que l'on vient d'énumérer. La partie E est une superposition des 4 courbes obtenues chez le chimpanzé, la partie F une superposition des 4 courbes obtenues chez l'humain. La position 0 correspond chez l'humain à la position de référence de l'élément Alu, celui-ci est donc positionné entre les positions -350 et 0 selon la taille de l'élément. Chez le chimpanzé, la position 0 correspond au site d'insertion retrouvé par l'alignement des 2×100 pb flanquants l'élément chez l'humain sur le génome du chimpanzé.

été séparée en deux par l'insertion, une partie se retrouvant au niveau du polyA terminal de l'Alu (en 0 pb), et l'autre juste en amont de l'élément (en -375 pb).

Pour le cas 2 (présence d'un NIEB uniquement chez l'humain, **Figure 4.8 - B**), on observe que le profil chez l'humain est très similaire à celui décrit pour le premier cas. On note cependant quelques différences. D'abord, les valeurs d'énergie semblent moins importantes, ce que l'on peut confirmer avec la **Figure 4.8 - F** où les profils humains sont superposés. Ensuite, on voit que le pic d'énergie observé en amont de l'Alu pour le cas 1 n'est plus présent dans le cas 2. Or, on a interprété ce pic comme un décalage de la barrière présente à l'état ancestral, coupée en deux par l'insertion. Dans le cas 2, il n'y a plus de barrière nucléosomale à l'état ancestral, ce qui peut expliquer l'absence de pic en amont de l'Alu et tend donc à confirmer notre hypothèse. Enfin, le profil de l'état ancestral est également différent. Il est beaucoup plus plat que dans le cas 1, ce qui explique l'absence de barrière nucléosomale détectée. Cependant, on note un léger pic d'énergie autour du 0, ce qui pourrait traduire d'une insertion dans des séquences déjà légèrement défavorables à la formation du nucléosome.

Concernant le cas 3, qui est l'absence de NIEB dans les deux espèces (**Figure 4.8 - C**) on observe des différences intéressantes avec les deux cas précédents, particulièrement avec le cas 2. La forme du profil énergétique chez l'humain est sensiblement identique à celle des cas 1 et 2, cependant les valeurs d'énergie sont considérablement plus basses (**figure 4.8 - F**). Cela explique l'absence de barrière nucléosomale détectée chez l'humain. Le polyA terminal reste inhibiteur de la formation du nucléosome, comme en témoigne le pic d'énergie au niveau de la position 0. Cependant, cette inhibition semble ici insuffisante pour provoquer la détection d'une barrière nucléosomale, contrairement au cas 2. La différence entre ces deux cas se trouve dans le profil énergétique de l'état ancestral. En effet, on a observé que dans le cas 2, au niveau du site d'insertion, la séquence semble légèrement défavorable à la formation du nucléosome. A l'inverse, dans le cas 3, on observe au légère baisse de l'énergie, traduisant des séquences qui semblent plutôt favorables à la formation du nucléosome.

Enfin, le cas 4, soit la présence d'un NIEB à l'état ancestral en l'absence d'Alu mais pas chez l'humain en présence de l'Alu (**Figure 4.8 - D**), bien que largement minoritaire (2 % des cas, **Tableau 4.1**), est un cas intéressant à analyser. Ici, on a une insertion qui a eu lieu dans un NIEB, comme dans le cas 1. Cependant, on remarque que le profil énergétique chez le chimpanzé diffère légèrement entre ces deux cas (**Figure 4.8 - E**). Dans le cas 1, on observe une forme de cloche, alors que dans le cas 4, bien que la partie gauche de la cloche soit retrouvée, la partie droite est quant à elle remplacée par une diminution beaucoup plus rapide de l'énergie. En fait, c'est parce qu'ici on est dans le cas d'une insertion dans une barrière nucléosomale, mais très proche du bord de cette barrière et dans le brin plaçant le corps de l'élément Alu à l'intérieur de la barrière. Post-insertion, on a donc une toute petite partie de la barrière ancestrale au niveau du polyA terminal de l'élément Alu, puis le corps de l'élément, puis le reste de la barrière ancestrale. C'est visible sur la **Figure 4.8 - D**, par le pic en position 0 pb, puis la baisse déjà observée dans les 3 cas précédents le long du corps de l'élément Alu, puis de nouveau un pic très important en amont de l'Alu (-375 pb), plus haut que celui observé dans le cas 1, et aussi intense que celui au niveau de la position 0 pb chez le chimpanzé. Ce que l'on observe ici, c'est un cas de "décalage" d'une barrière nucléosomale par l'insertion d'un élément Alu. C'est d'ailleurs ce type d'évènement qui explique les pics à 300 pb sur

la courbe bleue (resp. rouge) de la partie gauche (resp. droite) de la **Figure 4.2**.

Parmi les résultats obtenus pour les différents cas identifiés précédemment, la comparaison des deux cas majoritaires (les cas 2 et 3, **Figure 4.8 - B et C**, représentant 85.8 % des cas totaux) semble être la clé pour comprendre la formation de nouvelles barrières nucléosomales par les éléments Alu. Quelles que soient les conséquences sur la détection des barrières nucléosomales, l'insertion d'un élément Alu apporte une séquence fortement défavorable au nucléosome, au niveau du polyA terminal de l'Alu, illustrée par le pic observé au niveau de la position 0 pb sur tous les profils humains (**Figure 4.8 - F**). Cependant, selon si l'élément est inséré dans une séquence déjà légèrement défavorable (**Figure 4.8 - B**), ou légèrement favorable (**Figure 4.8 - C**) à la formation d'un nucléosome, alors le pouvoir inhibiteur à la formation du nucléosome associé au polyA terminal de l'élément Alu est plus ou moins fort. Il semble ici qu'on se trouve encore une fois face à l'effet de seuil, que l'on a décrit en détail dans le **Chapitre 2**. L'absence de barrière nucléosomale observée dans le cas 3 pourrait donc être principalement due à des faux négatifs dans la détection des NIEBs. Pour répondre à cette question, il est nécessaire d'analyser si les différences observées en terme de profils énergétiques se traduisent par des différences dans l'occupation nucléosomale.

Sur la **Figure 4.9**, on retrouve les profils d'occupation nucléosomale prédite, aux mêmes loci que pour les profils énergétiques de la **Figure 4.8**. On observe que l'hypothèse d'un effet de seuil est confirmée par ces profils. En effet, on voit très clairement que, quel que soit le cas analysé, le profil d'occupation nucléosomale prédite en présence de l'élément Alu chez l'humain (en bleu sur chaque panneau) est exactement le même au niveau de l'élément, et très proche en amont et en aval de celui-ci. Ainsi, que le modèle détecte une barrière nucléosomale ou non au niveau du polyA de l'élément Alu, il prédit de toute manière qu'à ce locus le nucléosome a bien moins de chance de se former qu'aux loci adjacents (autour de −100 pb et +100 pb). De plus, en amont de cette région inhibitrice située autour du 0, on observe un positionnement très fort des nucléosomes, avec deux positions préférentielles très claires en −100 pb et −250 pb. Ces positions correspondent aux deux bras de l'élément Alu. On observe également un positionnement des nucléosomes en aval de l'élément Alu sur les panneaux B, C et D. Ce positionnement est d'ailleurs plus marqué sur les panneaux C et D que sur le B. Or, paradoxalement, les panneaux C et D correspondent aux conditions pour lesquelles on ne détecte pas de barrière nucléosomale chez l'humain au niveau de la position 0 pb. D'après ces résultats, il semble donc que le positionnement soit plus fort en l'absence de barrière nucléosomale. Cependant, il faut tenir compte du fait qu'on a pris ici comme position de référence celle de l'élément Alu (à savoir le milieu du polyA terminal). La présence de l'Alu en amont de ce polyA "cale" le bord gauche de la barrière, mais on n'a imposé ici aucune contrainte de taille sur cette barrière. Ainsi, la position de référence utilisée ici peut tout aussi bien être à +50 ou à +150 pb du bord 3' de la barrière. Pour le panneau A, on obtient donc logiquement un profil plutôt plat en aval de l'Alu car les tailles de barrières sont très variables (**Figure 4.9**), on moyenne donc sur des situations variables dans ces régions. Pour le panneau B, on remarque qu'on voit apparaître du positionnement. Ce panneau représente les cas pour lesquels aucune barrière n'est détectée chez le chimpanzé alors qu'il y en a une détectée chez l'humain. Il semble donc que ce soit l'élément Alu qui soit à l'origine de la barrière. Il est donc logique de voir apparaître du positionnement en aval de la barrière, car, contrairement au panneau A, la taille de la barrière est dépendante de celle du polyA terminal de l'élément Alu, dont la taille varie peu (de 20 pb à 30 pb). Cependant, la taille des barrières semble également légèrement variable dans ce cas là car le

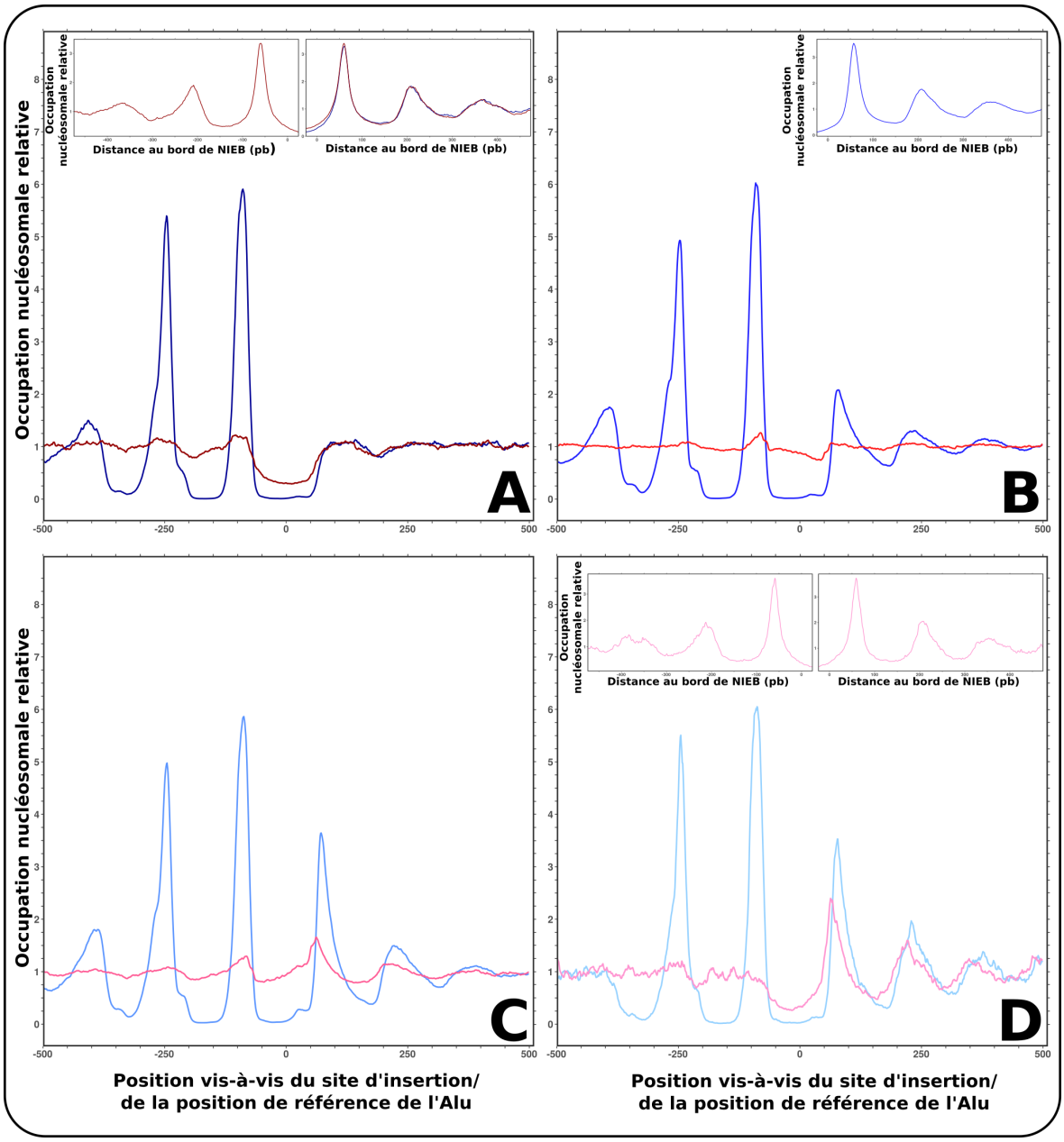


FIGURE 4.9 – **Profil d'occupation nucléosomale prédite des sites avec et sans Alu chez l'humain et le chimpanzé.** Les courbes rouges représentent les profils obtenus chez le chimpanzé, les courbes bleues les profils obtenus chez l'humain. Les panneaux A, B, C et D représentent respectivement les cas 1, 2, 3 et 4. La position 0 correspond chez l'humain à la position de référence de l'élément Alu, celui-ci est donc positionné entre les positions -350 et 0 selon la taille de l'élément. Chez le chimpanzé, la position 0 correspond au site d'insertion retrouvé par l'alignement des 2×100 pb flanquant l'élément chez l'humain sur le génome du chimpanzé. Les encadrés dans les parties A, B et D représentent les profils obtenus aux mêmes loci, mais en utilisant comme position de référence le bord des NIEBs plutôt que les positions de référence des éléments Alu ou les sites d'insertion. L'ensemble de ces profils a été normalisé par la moyenne génomique du signal correspondant.

positionnement en aval reste moins marqué que pour les panneaux C ou D. Cela est probablement dû à des différences dans le pouvoir inhibiteur des séquences au niveau du site initial, amenant à des barrières plus ou moins grandes. Quoi qu'il en soit, pour ces deux cas de figure, si l'on change de position de référence, en prenant comme 0 les bords des barrières, on obtient les profils présentés en encadrés dans les panneaux A et B. On y retrouve bien le positionnement précédemment décrit aux bords des barrières, avec 2 à 3 nucléosomes très positionnés. Cet exemple illustre bien toute

l'importance du référentiel dans l'interprétation de ces résultats. Pour les panneaux C et D, on est dans des cas où aucune barrière n'a été détectée chez l'humain. Or, on observe un positionnement très fort, à la fois au niveau de l'élément Alu mais également en aval du polyA terminal. Ce dernier joue ici le rôle de barrière nucléosomale. Comme sa taille est peu variable, le positionnement à son bord est "calé" sur le bord du polyA. Le positionnement des nucléosomes en aval de l'élément Alu sur le panneau C, ainsi que la prédiction d'une forte inhibition au niveau de la séquence polyA illustrée par les valeurs très basses autour de la position 0 pb, confirment l'hypothèse d'effet de seuil évoquée lors de l'analyse des profils énergétiques. Même si aucune barrière nucléosomale n'a été détectée ici, les profils d'occupation prédite indiquent que le polyA terminal est inhibiteur de la formation du nucléosome, et que des nucléosomes se forment au bord de cette séquence par effet de parcage. Le polyA terminal a donc, dans ce cas toutes les caractéristiques d'une barrière nucléosomale.

L'insertion d'un élément Alu, si elle n'a pas lieu directement dans une barrière nucléosomale (**Figure 4.9 - B et C**), provoque donc l'apparition d'une nouvelle barrière. Cette barrière nucléosomale est formée au niveau du polyA terminal de l'élément. En amont de ce polyA, la composition en GC de l'élément, avec deux bras d'environ 150 pb riches en GC et séparés par un court polyA, amène un positionnement très fort de deux nucléosomes. Un troisième nucléosome semble également être positionné en amont de l'élément (au niveau de la position -400 pb), par effet de parcage sur les deux premiers. En aval du polyA, on observe également un fort positionnement de deux à trois nucléosomes, dans la configuration de parcage des nucléosomes aux bords des NIEBs décrite dans les **Chapitres 1** à **3**. Ces barrières nucléosomales ne sont pas toujours détectées par le modèle (un nouveau NIEB n'est détecté que pour 55 % des insertions, **Partie 4.5.2.1**), principalement à cause de faux négatifs dans la détection de NIEBs.

Si l'insertion d'un élément Alu hors d'une barrière nucléosomale semble systématiquement mener à la formation d'une nouvelle barrière, les insertions à l'intérieur d'une barrière peuvent amener à différents cas de figures. Dans la grande majorité des cas (86 %, **Partie 4.5.2.1**), la barrière nucléosomale est conservée au niveau du polyA terminal de l'Alu, et est renforcée par celui-ci (**Figure 4.8 - A** et **4.9 - A**). Il semble également que ce type d'insertion puisse diviser une barrière existante en deux barrières distinctes, séparées alors par un élément Alu. Ce cas de figure fait alors apparaître les inter-barrières de tailles spécifiques à l'humain et au chimpanzé observés dans la **Figure 4.1**. Enfin, une petite partie des insertions semblent décaler les barrières nucléosomales. Cela arrive lorsqu'un élément Alu est inséré au bord interne d'une barrière nucléosomale, mais dans le brin faisant que le corps de l'élément se retrouve à l'intérieur de la barrière (**Figure 4.8 - D** et **4.9 - D**). Par exemple, si un Alu est inséré dans le brin sens, au bord interne droit d'une barrière. Dans ce cas, le polyA terminal est bien au bord interne droit, mais le corps de l'Alu, en amont du polyA, se retrouve au niveau de la barrière nucléosomale. Or, le corps de l'Alu est particulièrement compatible avec la formation du nucléosome, la structure de cet élément semblant même positionner deux nucléosomes. À la place de la barrière nucléosomale ancestrale, on a donc, post-insertion, des nucléosomes. On pourrait donc penser que l'insertion à "détruit" une barrière nucléosomale. Cependant, la séquence ancestrale inhibitrice de la formation du nucléosome n'a pas disparu du génome avec l'insertion d'Alu. En fait, elle a juste été décalée en amont de l'élément Alu. De plus, l'insertion ayant été effectuée au bord de cette séquence, cette dernière est donc presque intacte. Son pouvoir inhibiteur ne devrait donc être que peu réduit. Les courbes d'occupation

nucléosomale prédite illustrent l'insertion au bord d'une barrière (**Figure 4.9 - D**). En effet, on observe du positionnement préférentiel pour les nucléosomes entre les positions 0 et 500. Or, on a vu sur le panneau A qu'en prenant comme référentiel le site d'insertion, on n'est pas "calé" sur le bord de barrière, et le positionnement qu'on devrait observer disparaît à cause de l'effet de moyenne. Si on l'observe quand même ici, cela veut dire que même avec une moyenne centrée sur le site d'insertion, on reste "calé" sur le bord 3' de la barrière nucléosomale. Le site d'insertion est donc bien positionné au bord interne d'une barrière, à l'extrémité 3' si l'insertion est faite dans le brin sens (et à l'extrémité 5' si elle est faite dans le brin antisens). En revanche, la taille variable des NIEBs fait que l'on n'est pas aligné sur l'autre bord de NIEB, ce qui explique le profil plat sur la gauche du panneau (< -300 pb), à cause du même effet de moyenne que pour la **Figure 4.9 - A**. Dans le cas d'un décalage de barrière nucléosomale, on observe d'ailleurs aussi un effet de seuil, comme pour la **Figure 4.9 - C**. En effet, ici aussi, la formation du nucléosome est très inhibée autour du 0, au niveau du polyA terminal, et on observe un fort positionnement par effet de parcage en aval de ce polyA terminal. Il a donc, ici encore, toutes les caractéristiques d'une barrière nucléosomale, mais, comme illustré par la **Figure 4.8 - D et F**, le profil énergétique présente des valeurs trop basses pour permettre la détection d'une barrière nucléosomale.

De manière générale, l'insertion d'un élément Alu a donc d'importantes conséquences sur le positionnement nucléosomal. En effet, les prédictions de ce positionnement dans la séquence sans l'insertion (**Figure 4.9, en rouge**) sont sensiblement différentes de celles obtenues avec l'insertion (**Figure 4.9, en bleu**). Les insertions dans les régions dépourvues de NIEB (86 %) sont majoritairement à l'origine de nouvelles barrières nucléosomales, même si celles-ci ne sont pas toujours correctement détectées. Les insertions dans des NIEBs ancestraux peuvent amener des changements de positionnement nucléosomal par décalage d'une barrière ou la division d'une barrière en deux barrières distinctes. Dans tous les cas, l'insertion d'éléments Alu affecte fortement le positionnement nucléosomal.

4.5.2.3 Les profils d'occupation nucléosomale expérimentale confirment la formation de nouvelles barrières nucléosomales

Les effets de l'insertion d'éléments Alus sur le positionnement nucléosomal décrits dans les parties précédentes ont été observés à partir des prédictions faite par notre modèle physique de positionnement des nucléosomes. Ce modèle a été confirmé expérimentalement chez l'humain, ce qui nous permet d'avoir confiance dans nos observations précédentes. Cependant, ces confirmations étant en partie faites sur des moyennes, il apparaît important de tester directement nos prédictions quant aux conséquences de l'insertion d'éléments Alu sur le positionnement nucléosomal avec des données expérimentales, pour voir si ces éléments sont bien à l'origine de nouveaux NIEBs. Pour tenter de confirmer les résultats de la **Partie 4.5.2.2**, j'ai utilisé les données expérimentales de positionnement nucléosomal chez l'humain produites par Mieczkowski et al. (Mieczkowski et al. (2016), **Chapitre 3**). Je n'ai ici utilisé qu'un seul jeu de données expérimentales, car celui produit par Valouev et al. (Valouev et al. (2011), **Chapitre 3**) est de type single-end avec des lectures de taille 35 pb. Or, nous avons observé que les éléments Alu ne présentent pas une mappabilité suffisante pour permettre les analyses avec un tel jeu de données. Les éléments Alu spécifiques à l'humain sont, par définition, ceux insérés le plus récemment dans le génome humain. En ce sens, il s'agit

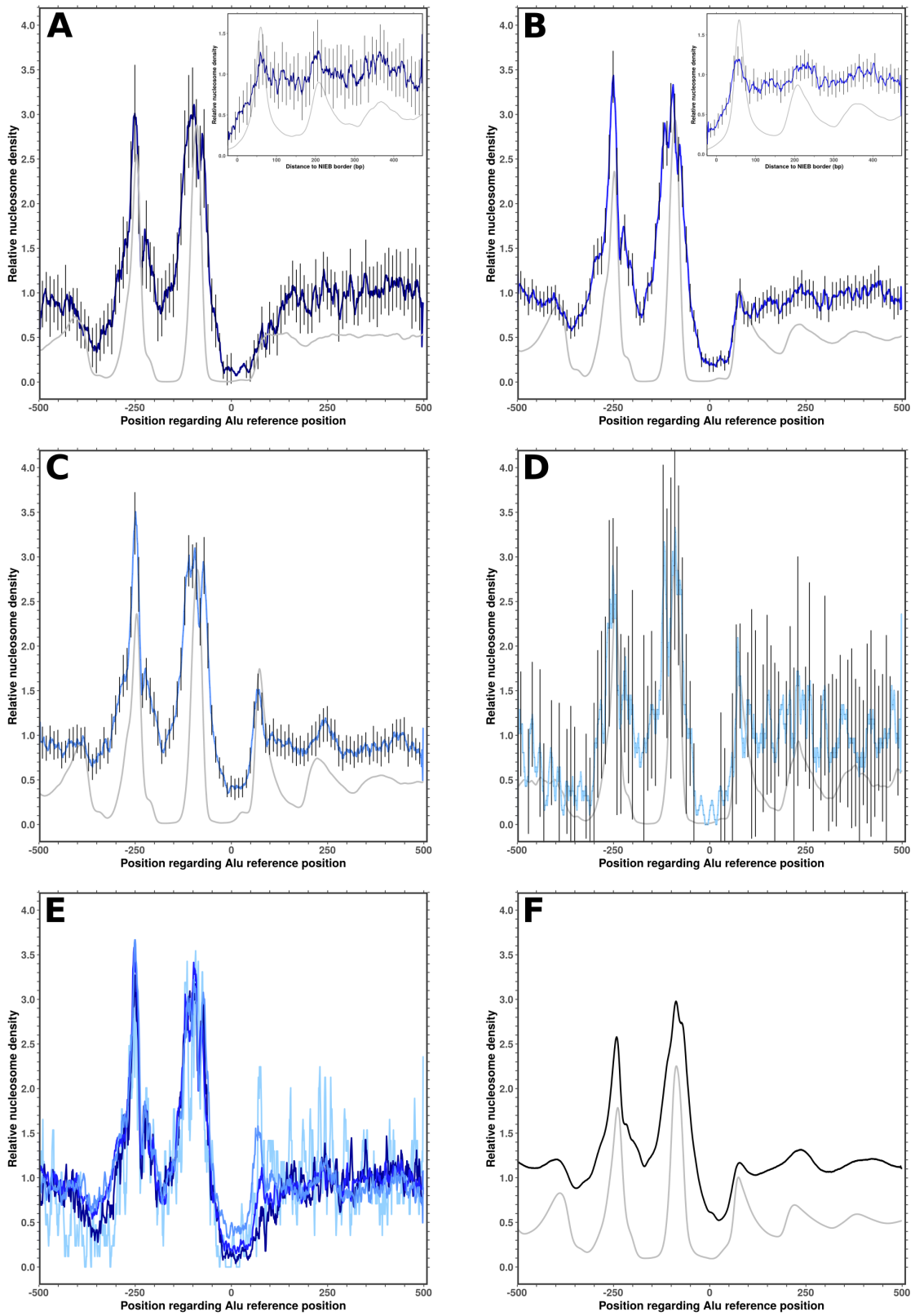


FIGURE 4.10 – **Profil d'occupation nucléosomale expérimentale sur les éléments Alu spécifiques à l'humain obtenus à partir des données de Mieczkowski et al.** La légende est à retrouver sur la page suivante.

FIGURE 4.10 – Les panneaux A à D représentent, respectivement, les 572, 2229, 1831 et 96 éléments Alus identifiés dans chaque cas de la **Partie 4.5**. Les courbes bleues représentent les profils moyens expérimentaux d'occupation nucléosomale. Ces profils moyens ont été obtenus à partir du signal issu des données expérimentales publiées par Mieczkowski et al. (Mieczkowski et al., 2016), en suivant le protocole développé au **Chapitre 3** pour les données paired-end, mais en ne filtrant pas les lectures selon leur qualité d'alignement (voir **Partie 3.2.1**). Les encadrés des panneaux A et B représentent les mêmes profils mais avec comme position de référence le bord de barrière nucléosomale, comme dans la **Figure 4.9**. Les barres d'erreurs ont été obtenues en prenant l'écart-type d'une loi binomiale de paramètres n et p avec n : le nombre de sites considérés (par exemple 572 pour le panneau A, 2229 pour le panneau B, etc...), et p : la valeur de la courbe à la position i. Pour faciliter la lecture du panneau, ces barres d'erreurs n'ont été indiquées que tous les 10 pb. Sur la figure E sont présentes les 3 courbes des panneaux A à C, pour faciliter les comparaisons. La courbe de la figure D n'y est pas présente car les barres d'erreurs qui lui sont associées ainsi que le nombre de cas très réduit la rende peu pertinente. La figure F représente le profil moyen obtenu sur les 1 078 322 éléments Alu des autosomes humains. Des barres d'erreurs ont aussi été calculées pour cette courbe, mais elles sont si petites qu'elles n'apparaissent pas sur le graphe. Pour finir, les courbes grises des panneaux A et D ainsi que du panneau F représentent les profils moyens obtenus en prenant comme signal l'occupation nucléosomale prédite par le modèle. Il s'agit donc des mêmes courbes que celles de la **Figure 4.9**, à l'exception de la figure F où il s'agit du profil obtenu sur tous les Alu des autosomes humains et qui n'avait pas été montré précédemment. L'ensemble de ces profils a été normalisé par la moyenne génomique du signal correspondant. Enfin, pour faciliter la lecture des courbes, les profils d'occupation nucléosomale prédits à l'aide du modèle ont été divisé par deux afin de rapprocher leur valeur de celles des profils expérimentaux, excepté pour les encadrés.

du sous-groupe sur lequel la dérive génétique s'est appliquée sur le laps de temps le plus court. Si des analyses plus poussées sont nécessaires pour affirmer avec certitude que ce groupe d'Alu est le moins divergent du génome humain, la mappabilité plus faible de ces Alu par rapport aux familles J, S et Y prises dans leur ensemble témoigne d'une proximité de séquence très importante. Cette proximité de séquence complique l'alignement car il devient alors plus difficile voire impossible d'aligner les lectures de manière unique. Pour pallier ce problème, il est nécessaire d'augmenter la taille des lectures, d'où l'utilisation du jeu de données produit par Mieczkowski et al., dans lequel le séquençage est de type paired-end, avec 2x50 pb séquencées par fragment. J'ai également modifié le protocole d'alignement des lectures pour cette analyse, en retirant le filtrage effectué sur la qualité d'alignement (voir **Partie 3.2.1**) pour autoriser l'analyse des lectures alignées de manière équivalente à plusieurs loci du génome. Cependant, pour ce cas de figure, j'ai fait le choix de sélectionner un alignement au hasard parmi les alignements possibles. Ainsi, chaque lecture continue de n'être traitée qu'une seule fois. Pour produire le profil expérimental, je me suis concentré sur l'approche MNase-seq standard, à savoir celle avec la plus forte concentration de MNase (304U). Cela m'a permis de construire les profils expérimentaux pour les sites d'insertion d'éléments Alu spécifiques à l'humain, chez l'humain, dans les différents cas décrits précédemment (**Partie 4.5.2.1**). Malheureusement, aucun jeu de données expérimentales de positionnement nucléosomal n'est disponible chez le chimpanzé, je n'ai donc pas pu produire les profils expérimentaux des sites d'insertion sans élément Alu. Pour ces derniers, on devra donc se contenter des prédictions du modèle. Les profils expérimentaux obtenus pour les différents cas sont présentés en **Figure 4.10**.

La **Figure 4.10** est organisée de la même manière que les figures présentant les profils d'énergie de positionnement du nucléosome et d'occupation nucléosomale prédite précédemment analysées (**Figures 4.8** et **4.9**). Avant d'en faire une analyse détaillée, il convient de noter que pour la **Figure 4.10 - D**, on observe que les barres d'erreurs sont trop importantes pour avoir confiance dans le profil expérimental. Cela est du au manque de couverture. En effet, cette figure correspond aux cas où une barrière est détectée chez le chimpanzé et pas chez l'humain, décrit précédemment comme

un décalage de barrière nucléosomale. Ce cas est rare (moins de 100 cas détectés, **Tableau 4.1**). De plus, la profondeur du séquençage dans les données expérimentales utilisées ne permettent d'obtenir que quelques lectures tout au plus pour chaque nucléosome ($\overline{C} = 3.3$, **Tableau 3.1**). Pour le reste des figures, les barres d'erreurs semblent acceptables pour permettre les interprétations.

On observe que dans les trois premiers cas, les prédictions du modèle semblent être globalement confirmées par les données expérimentales (**Figures 4.10 - A à C**). En effet, on retrouve la diminution de l'occupation au niveau de la position 0 pb, illustrant la présence d'une barrière nucléosomale. En aval de cette barrière, sur la partie droite des figures, on observe du positionnement des nucléosomes lorsqu'aucune barrière n'est detectée ni chez l'humain ni chez le chimpanzé (**Figure 4.10 - C**). Lorsqu'une barrière est détectée dans les deux espèces, on perd le positionnement (**Figure 4.10 - A**). Cependant, comme on l'avait noté avec les profils d'occupation nucléosomale prédite, on retrouve ce positionnement en prenant comme référence le bord 3' de la barrière nucléosomale (**Figure 4.10 - A, encadré**). On observe le même phénomène pour le cas où on a une barrière nucléosomale chez l'humain et pas chez le chimpanzé (**Figures 4.10 - B et encadré**). Les profils expérimentaux confirment donc les prédictions de formation d'une nouvelle barrière nucléosomale, ainsi que le positionnement nucléosomal prédit au bord de cette barrière pour le côté sans élément Alu. Pour ce qui est du côté avec l'élément Alu, on confirme également le positionnement prédit. En effet, on observe un fort positionnement à la fois au niveau de la position -250 pb et de la position -100 pb, dans les trois cas analysés précédemment (**Figure 4.10 - A à C**), comme attendu d'après les prédictions. On note également une forte baisse de l'occupation nucléosomale au niveau du polyA interne de l'élément Alu (entre -150 pb et -200 pb). Enfin, au niveau de l'élément Alu, on observe que les quatre courbes produites se superposent parfaitement (**Figure 4.10 - E**). On note également que ces courbes sont compatibles avec celle obtenue en prenant l'ensemble des éléments Alu du génome humain (**Figure 4.10 - F**). Sur cette dernière figure, on observe que, comme pour les éléments spécifiques à l'humain, les profils expérimentaux sur les éléments Alu (en noir) sont en accord avec les prédictions (en gris). En conclusion, le positionnement observé expérimentalement sur les éléments Alu, et particulièrement sur les éléments spécifiques au génome humain reproduit les prédictions, ce qui confirme l'hypothèse de la formation de nouvelles barrières nucléosomales par l'insertion d'éléments Alu. Il serait intéressant d'utiliser les données de gradients de MNase présentées dans le **Chapitre 3** pour déterminer l'accessibilité aux nucléosomes aux différentes positions observées ici, pour voir dans quelle mesure les caractéristiques d'accessibilité de ces nouvelles barrières nucléosomales sont similaires à celles des barrières plus anciennes.

4.6 Conclusion

Dans ce chapitre, on a d'abord observé que les éléments Alu ont un positionnement contraint aux bords des barrières nucléosomales (**Figure 4.2**). Plusieurs hypothèses ont été émises pour expliquer la distribution de ces éléments aux bords des NIEBs. Les travaux présentés dans les **Parties 4.3**, **4.4** et **4.5** ont mis en évidence que les insertions d'éléments Alu sont à l'origine de la formation de nouvelles barrières nucléosomales, par l'apport du polyA terminal de ces éléments, fortement inhibiteur de la formation du nucléosome. Aux bords de ces barrières nucléosomales, on observe, comme détaillé dans les **chapitres 1** et **3**, du positionnement des nucléosomes par effet de parcage contre la barrière, du côté sans élément Alu. Ce positionnement a été confirmé expérimentalement par l'analyse de jeux de données expérimentales de type MNase-seq. Toujours aux bords de ces nouvelles barrières mais cette fois du côté avec Alu, les données expérimentales de positionnement ont mis en évidence un fort positionnement des nucléosomes sur les deux bras des éléments Alu, en accord total avec les prédictions de notre modèle. La formation de nouvelles barrières nucléosomales a donc pu être directement confirmée expérimentalement. Des analyses de type gradients de MNase, permettant d'avoir une information sur l'accessibilité des nucléosomes, pourraient permettre de mieux caractériser ces nouvelles barrières et de voir si des différences existent avec les barrières plus anciennes du génome humain. Plusieurs analyses restent à mener pour comprendre complètement l'impact de l'insertion des éléments Alu sur le positionnement nucléosomal. Néanmoins, on a démontré ici que les insertions d'éléments Alu ne sont pas neutres vis-à-vis du positionnement des nucléosomes. Elles ont même un très fort impact sur ce dernier, étant à l'origine de nouvelles barrières nucléosomales, et donc de positionnement par effet de parcage. Les barrières nucléosomales semblant être un objet conservé dans les génomes eucaryotes (**Chapitre 2**), potentiellement pour faciliter les modifications épigénétiques (**Chapitre 3**), le lien entre les éléments Alu et ces barrières pourrait en partie expliquer le succès invasif de ces éléments dans les génomes de primates.

5 Conclusion et perspectives générales

Sommaire

L'importance de la séquence dans le positionnement nucléosomal n'est plus à démontrer, même si les stratégies peuvent diverger entre séquences positionnantes et séquences inhibitrices de la formation du nucléosome positionnant par effet de parcage (Segal & Widom, 2009b). L'écriture de la position des nucléosomes dans les séquences d'ADN implique que l'évolution des séquences a une influence sur l'évolution du positionnement nucléosomal. Cette influence peut même amener à de la sélection de certaines séquences pour leurs effets sur le positionnement des nucléosomes. Aujourd'hui, de plus en plus d'indices suggèrent que la relation entre évolution des séquences et évolution du positionnement nucléosomal n'est pas unidirectionnelle. En effet, le nucléosome a été, ces dernières années, associés à une variété de biais dans les patrons de mutations. Ainsi, l'évolution des séquences influence l'évolution du positionnement nucléosomal, qui lui même influence l'évolution des séquences (Barbier et al., 2021).

Cette relation bidirectionnelle entre évolution des séquences d'ADN et du positionnement nucléosomal a été au centre de ma thèse. Pour l'explorer, j'ai étudié un mécanisme de positionnement des nucléosomes par effet de parcage contre des séquences inhibitrices de la formation du nucléosome (qualifiées de barrières nucléosomales, ou NIEBs). J'ai abordé trois questions fondamentales vis à vis des NIEBs :

— Les NIEBs sont-ils universels aux eucaryotes ? Quelle est leur conservation chez ces espèces ?

— Quelle est l'importance fonctionnelle des NIEBs ?

— Dans quelle mesure la dynamique évolutive des NIEBs est-elle liée à celle des éléments transposables ?

5.1 Les barrières nucléosomales, une écriture ubiquitaire du nucléosome dans les séquences...

Dans le **Chapitre 2**, on a pu voir que les NIEBs ont été détectés dans chacune des espèces eucaryotes que nous avons étudiées, que ce soit chez les mammifères, les oiseaux, les poissons, les plantes ou encore les levures. De manière générale, les caractéristiques associées à ces barrières sont conservées entre les espèces, même si leur densité dans les génomes peut différer. La taille des NIEBs et les distances inter-NIEBs ont une distribution très proche entre les dix espèces étudiées. Les distances inter-NIEBs sont quantifiées, traduisant une contrainte pour un nombre entier de nucléosomes entre deux NIEBs proches, contrainte remarquablement conservée chez tous les eucaryotes étudiés. L'étude de la composition en GC aux bords des NIEBs ainsi que de l'occupation nucléosomale intrinsèque prédite à l'aide de notre modèle de positionnement des nucléosomes a également montré une grande conservation du positionnement nucléosomal à ces loci. En somme, les caractéristiques des NIEBs sont partagées entre l'ensemble des eucaryotes. Une étude de la conservation NIEB par NIEB entre l'humain et le chimpanzé a également mis en évidence une conservation très forte de ces structures entre ces deux espèces proches. En effet, la majorité des différences de NIEBs observées entre ces deux espèces ont été identifiées comme des faux-négatifs dans la détection de ces barrières dans l'une ou l'autre des deux espèces.

De manière générale, il semble donc que l'encodage du positionnement des nucléosomes dans la séquence au moyen de séquences inhibitrices à leur formation contre lesquelles ils se

positionnent par effet de parcage soit un mécanisme commun aux eucaryotes. L'analyse comparative menée dans le **Chapitre 2** pourrait être étendue à un nombre plus important d'espèces. Cela permettrait de préciser les caractéristiques des NIEBs selon les groupes d'espèces, pour voir si des particularités émergent selon les clades, comme celle qu'on a pu observer pour les distances inter-NIEBs chez le porc (**Partie 2.2.3**). Il serait également intéressant de trouver des espèces pour lesquelles ce mécanisme de positionnement nucléosomal n'est que peu utilisé (voire pas du tout). L'étude de ces dernières (si elles existent) permettrait alors de mettre en évidence des stratégies alternatives de positionnement nucléosomal à l'échelle génomique. Enfin, il serait aussi intéressant d'appliquer notre modèle de positionnement nucléosomal à certaines archées, où la compaction du génome se fait autour de tétramères d'histones. Cela permettrait de voir si la contrainte évolutive associée aux NIEBs chez les eucaryotes est également pertinente pour des nucléosomes différents de la forme canonique.

Les résultats montrés dans la dernière partie du **Chapitre 2** ont mis en évidence plusieurs aspects de la relation entre NIEBs et évolution des séquences. Tout d'abord, les patrons de mutations observés précédemment chez l'humain (Drillon et al., 2016) sont retrouvés chez le chimpanzé. Ce résultat suggère que la sélection renforçant l'oscillation du taux de GC (et donc le positionnement des nucléosomes) aux bords des NIEBs n'est pas une spécificité humaine mais est également présente dans au moins une autre espèce. Des analyses comparatives entre d'autres quadruplets d'espèces proches entre elles permettraient de tester la généralité de cette sélection liée au positionnement nucléosomal chez les eucaryotes. Il est cependant à noter que pour tester des effets de sélection, il est nécessaire d'accéder aux patrons de mutations neutres. Chez l'humain, cela a été fait en utilisant des jeux de données de SNPs issus du projet 1000 Genomes (Durbin et al., 2010). Pour observer des effets de sélection dans d'autres espèces, il sera donc nécessaire de mettre au point ce type de jeux de données. On note cependant que les SNPs ne sont pas totalement exempts de contraintes sélectives. Dans l'idéal, il faudrait donc mettre au point des jeux de données de mutations *de novo* dans les espèces considérées, à l'image de ce qui a pu être fait pour étudier les préférences d'insertions de certaines familles d'ETs (Sultana et al., 2017). Si la sélection semble, chez l'humain (et probablement le chimpanzé), renforcer le positionnement nucléosomal aux bords des NIEBs, ce positionnement est également à l'origine de biais de mutations comme celui observé au niveau des dinucléotides CpG. En effet, les mutations de C vers T en contexte CpG sont distribuées de manière hétérogène aux bords des NIEBs, étant clairement favorisées dans les séquences inter-nucléosomales par rapport aux séquences nucléosomales. Ce résultat est en accord avec une désamination spontanée des cytosines plus facile dans l'ADN accessible que dans celui contraint structurellement par l'interaction avec les histones (Makova & Hardison, 2015). Il reste néanmoins à déterminer quel pourrait être l'impact de la sélection dans les différences observées. En effet, on remarque que le biais observé ici favorise les mutations vers T dans les séquences inter-nucléosomales, exactement de la même manière que les effets de sélection observés précédemment. Il est donc nécessaire ici de déterminer quelle est l'importance relative du mécanisme mutationnel et de la sélection dans l'hétérogénéité de la distribution des mutations C vers T en contexte CpG aux bords des NIEBs. L'analyse de ce type de mutations pour les SNPs humain pourrait apporter un élément de réponse à cette question en permettant d'évaluer les effets de sélection dans cette espèce.

Si les biais de mutations C vers T en contexte CpG sont confirmés comme étant associés au

positionnement nucléosomal plutôt qu'à des effets de sélection, voire en plus de ces effets, ils pourraient alors être utilisés comme une confirmation indirecte du positionnement nucléosomal dans les espèces pour lesquelles les données expérimentales ne sont pas disponibles. Il ne serait alors plus forcément nécessaire d'accéder à des données de type MNase-seq pour confirmer le positionnement des nucléosomes aux bords des NIEBs dans un génome, même si l'étude de ce genre de données resterait une preuve plus directe. On note cependant que l'étude des taux de mutations en contexte CpG implique de travailler en moyenne sur une population de NIEBs, et ne permet donc pas les confirmations à l'échelle du NIEB individuel comme peut le faire l'analyse de MNase-seq avec le test de déplétion en nucléosome présenté en **Partie 3.2.1.3**. De plus, pour étudier les mutations en contexte CpG, il est nécessaire de comparer au minimum trois génomes (deux génomes d'intérêt et un groupe externe), et dans l'idéal au moins quatre (pour avoir deux groupes externes), pour déterminer la séquence ancestrale afin d'étudier les substitutions dans le génome d'intérêt. Si ce genre de données est plus accessible que les données de type MNase-seq car de plus en plus d'espèces voient leurs génomes séquencés et assemblés, cela reste assez contraignant. Pour s'affranchir de cette dernière contrainte, on pourrait utiliser le taux des CpG observés sur les CpG attendus. Pour calculer ce taux, on estime, à partir de la composition en G et C d'un génome, le nombre de dinucléotides CpG attendus dans le génome. On divise alors le nombre de dinucléotides CpG effectivement observés par le nombre de dinucléotides CpG attendus, ce qui nous donne un taux qui permet d'estimer la mutation de ce type de dinucléotides. Si les CpG sont particulièrement sujets aux mutations, alors le nombre de CpG observés sera en deça du nombre de CpG attendus, et le taux sera alors inférieur à 1. Chez l'humain par exemple, où 80% des dinucléotides CpG sont méthylés (Jang et al., 2017), le génome contient cinq fois moins de CpG que le nombre attendu selon sa composition en G et C (Moore et al., 2013). Ainsi chez l'humain, le taux de CpG observés sur attendus est de 0.2 (1/5). Il est possible de calculer ce taux en fonction de la distance aux bords des NIEBs, afin d'y évaluer le taux de mutations des CpG. Chez l'humain et le chimpanzé, on observe un taux de mutations des CpG inégal entre NIEBs/linker et séquences nucléosomales (**Partie 2.5.2**). On observe de même un taux des CpG observés sur les CpG attendus oscillant entre valeurs faibles dans les NIEBs et linkers (où les CpG sont plus souvent mutés) et fortes dans les nucléosomes (où les CpG sont moins souvent mutés) (Résultats préliminaires non-montrés). Le calcul de cette métrique pour d'autres espèces concernées par ce type de mutations comme les mammifères permettrait d'obtenir des indications sur le positionnement nucléosomal aux bords des NIEBs d'une espèce. On pourrait donc confirmer les prédictions faites à partir de notre modèle en moyenne sur des populations de NIEBs, même lorsque la seule donnée disponible est le génome de référence de l'espèce en question.

5.2 Permettant les changements chromatiniens à l'échelle génomique...

Le **Chapitre 3** a été consacré à l'analyse de jeux de données expérimentaux de positionnement de nucléosomes. Tout d'abord, l'étude des jeux de données de MNase-seq standards disponibles a permis d'apporter une confirmation expérimentale que les NIEBs sont des régions déplétées en nucléosomes (NDRs) *in vivo* chez la souris, la drosophile, le poisson-zèbre et l'arabette. Cependant, on a discuté que l'expérience de MNase-seq peut être associée à des problèmes de reproductibilité des résultats qui rendent difficiles les comparaisons entre espèces (Mieczkowski et al., 2016). Pour

pallier ce problème, j'ai utilisé d'autres jeux de données expérimentales issus d'une expérience de MACC-seq (MACC pour MNase accessibility), dans laquelle plusieurs niveaux de digestion de la chromatine permettent d'extraire des informations sur l'accessibilité des nucléosomes à la MNase. L'analyse des jeux de données MACC-seq chez l'humain, la souris et la drosophile a mis en évidence que les NIEBs ne sont pas des régions sans nucléosomes mais plutôt des régions où le nucléosome est plus accessible que dans le reste du génome. Cela le rend plus sensible à la MNase, ce qui explique que jusqu'ici on ait assimilé les NIEBs à des NDRs. Aux bords des NIEBs, les nucléosomes apparaissent comme peu accessibles et assez stables chez l'humain et la souris. A l'inverse, chez la drosophile, aux mêmes loci, les nucléosomes ont été observés comme accessibles et instables. Cette différence suggère une importance du contexte chromatinien dans le positionnement et l'accessibilité aux nucléosomes aux bords des NIEBs.

D'autres analyses sont nécessaires pour déterminer l'origine des différences d'accessibilité observées aux bords des NIEBs entre les deux mammifères et la drosophile. Par exemple, on pourrait étudier l'éventuel impact de la densité en gènes de la région dans laquelle se trouvent les NIEBs sur le positionnement/l'accessibilité des nucléosomes à leurs bords. En effet, chez l'humain, les régions riches en gènes sont associées à une chromatine plutôt ouverte et accessible, tandis que les régions pauvres en gènes sont associées à une chromatine plutôt fermée et inaccessible (Gilbert et al., 2004). Il est donc possible que la densité en gènes de la région, en étant associée à différents contexte chromatiniens, influence le positionnement/l'accessibilité des nucléosomes aux bords des NIEBs. La séparation des NIEBs humains en sous-groupes selon la densité en gènes de la région dans laquelle ils se trouvent devrait permettre de mettre en évidence un tel effet. La répétition de cette analyse chez la souris et la drosophile permettrait de voir si les particularités observées chez la drosophile en prenant l'ensemble des NIEBs s'expliquent par les différences globales de densité en gènes entre les génomes étudiés. De manière générale, il sera nécessaire de tenir compte, dans les futures analyses, du contexte chromatinien dans lequel sont placées les barrières nucléosomales lors de l'étude du positionnement et de l'accessibilité des nucléosomes à leurs bords. Il pourrait être intéressant de coupler l'étude de l'organisation nucléosomale aux bords des NIEBs à des caractéristiques génomiques comme le timing de réplication ou les compartiments GC. En effet, les résultats obtenus dans cette thèse suggèrent une organisation nucléosomale flexible et dynamique aux bords des NIEBs. Un travail de modélisation est actuellement en cours au Laboratoire de Physique à l'ENS de Lyon pour inclure la dynamique de formation/déformation des nucléosomes aux bords des NIEBs à nos analyses. L'étude des données expérimentales de type MACC-seq selon le contexte chromatinien pourrait faciliter cette démarche, en précisant les possibles variations de positionnement et d'accessibilité des nucléosomes aux bords des NIEBs.

La mise en évidence de la présence de nucléosomes accessibles à l'intérieur des NIEBs a suggéré qu'ils pourraient être utilisés comme des "points chauds de modifications épigénétiques", particulièrement celles effectuées par échange d'histones. Le variant d'histone H3.3, associé à la fois à des structures chromatiniennes ouvertes ou fermées (Szenker et al., 2011), a d'ailleurs été retrouvé comme enrichi aux bords des NIEBs chez la souris, ce qui corrobore cette hypothèse. L'analyse, au niveau des NIEBs, de données de type MNase-ChIP-seq comme celles produites pour le variant H3.3 chez la souris (**Partie 3.3.3**), pour d'autres variants d'histones, permettra de préciser l'association entre NIEBs et échange d'histones. L'analyse de données de type ChIP-seq permettrait, lorsqu'elles ont une résolution assez importante (de l'ordre de la dizaine ou centaine

de paires de bases), de mettre en évidence une éventuelle association des NIEBs avec d'autres marques épigénétiques comme les modifications post-traductionnelles des histones (O'Geen et al., 2011). Ces études permettraient de préciser la relation entre NIEBs et épigénétique, et peut être d'expliquer la dissémination de ces structures dans tous les génomes eucaryotes, ainsi que l'apparente conservation des NIEBs durant l'évolution observée dans le **Chapitre2**.

5.3 Et dont la dynamique évolutive dépendrait des éléments transposables

Le **Chapitre 4** de cette thèse a été consacré aux interactions entre les NIEBs et les éléments transposables, à travers l'exploration, chez l'humain, de la relation entre NIEBs et éléments Alu. En effet, il a été remarqué que ces éléments sont sur-représentés aux bords des barrières nucléosomales (Brunet et al., 2018; Drillon et al., 2016). Les travaux effectués durant cette thèse ont mis en évidence que la distribution des éléments Alu aux bords des NIEBs s'explique par la formation de nouveaux NIEBs lors de l'insertion de ces éléments. Ces NIEBs sont formés au niveau du polyA terminal des éléments Alu, les séquences polyA étant d'ailleurs connues pour être inhibitrices de la formation du nucléosome (Segal & Widom, 2009a,b). La formation de nouvelles barrières nucléosomales par l'insertion d'Alu a pu être confirmée expérimentalement avec des données de MNase-seq standards. Il pourrait être intéressant ici d'analyser les données de MACC-seq le long des éléments Alu pour accéder à l'information d'accessibilité de leurs nucléosomes. Cela permettrait de préciser les caractéristiques des nucléosomes se formant sur et autour des éléments Alu spécifiques à l'humain, et de voir si les NIEBs nouvellement formés ont les mêmes caractéristiques que ceux plus anciens. On pourrait également étendre l'analyse des données de type MACC-seq aux autres éléments Alu du génome humain, notamment en reprenant la séparation selon les familles d'Alu de la **Partie 4.3**. Cela pourrait permettre de déterminer si l'âge des éléments Alu peut avoir une influence sur le positionnement nucléosomal. De manière générale, l'étude des jeux de données expérimentaux, particulièrement ceux de type MACC-seq, au niveau des éléments Alu permettra de mieux caractériser l'impact de l'insertion de ces ETs sur la position et l'accessibilité aux nucléosomes.

Les éléments Alu ont été mis en évidence comme de potentiels facteurs de création de NIEBs. Dans le **Chapitre 3**, une hypothèse plaçant les NIEBs comme des portes d'entrée aux modifications épigénétiques a été émise, et en partie confirmée par l'enrichissement du variant d'histones H3.3 à ces loci. Ces résultats amènent à une potentielle importance fonctionnelle des NIEBs dans les génomes. Si celle-ci se confirme, il sera important de répéter les analyses ayant mené à cette confirmation spécifiquement sur les NIEBs associés à des éléments Alu, pour déterminer s'ils possèdent les mêmes caractéristiques et fonctions que les autres NIEBs des génomes de primates. En effet, associer une fonction aux NIEBs formés par l'insertion d'Alu pourrait permettre d'expliquer le formidable succès invasif de ces ETs dans les génomes de primates.

Les résultats produits lors de l'exploration, durant cette thèse, de la relation Alu-NIEBs amènent à la question d'une éventuelle relation similaire entre les NIEBs et d'autres familles d'ETs. L'oscillation du taux de GC sur les éléments Alu suggère qu'ils sont particulièrement propices à la formation des nucléosomes. Cependant, il semble que c'est surtout la présence d'un polyA terminal

amenant à la formation de nouvelles barrières nucléosomales qui est l'élément le plus important dans l'impact de l'insertion d'Alu sur le positionnement nucléosomal. Or, de nombreuses familles d'ETs, particulièrement de SINEs comportent des polyA terminaux. Par exemple, les éléments B1 chez la souris sont issus du même ARN7SL que les éléments Alu (Labuda et al., 1991; Quentin, 1994). La principale différence entre ces deux familles d'éléments est que les Alu sont des dimères alors que les B1 sont des monomères. En fait, la structure des B1, formée d'une séquence d'environ 150 pb riche en GC suivie d'un polyA terminal fait qu'on peut assimiler ces ETs à des "moitiés d'Alu" (Dridi, 2012). Aux bords des NIEBs de souris, les éléments B1 se comportent exactement ainsi, avec des positions préférentielles plaçant leur polyA terminal majoritairement au bord interne du NIEB (**Figure 5.1 - A**). Les autres positions possibles placent toujours ce polyA au niveau des séquences inter-nucléosomales. Les éléments Pre0_SS chez le porc suivent les mêmes contraintes de positionnement aux bords des NIEBs (**Figure 5.1 - B**). Ces éléments, spécifiques au porc (PRE signifiant Porcine Repetitive Element), dérivent d'un ARN de transfert d'acide glutamique (Groenen et al., 2012). Ils ont une taille assez proche de celle des éléments Alu, d'environ 250 pb. Enfin, ils comportent un polyA terminal, placés, à l'image de celui des Alu et des B1, majoritairement aux bords internes des NIEBs (**Figure 5.1 - B**). Le positionnement de ces deux familles d'éléments aux bords des NIEBs en fait de bons sujets d'étude de la relation entre ETs et NIEBs, pour voir si l'on retrouve la même interaction qu'avec les éléments Alu chez l'humain et le chimpanzé. Le fait de retrouver un positionnement similaire pour trois familles d'ETs différentes, dans quatre espèces différentes, suggère une relation générale entre éléments transposables et NIEBs, qu'il convient donc de caractériser. Dans quelle mesure l'insertion d'ETs fait évoluer le positionnement nucléosomal ? La relation ETs-NIEBs influence-t-elle le pouvoir invasif des familles d'ETs ? Les ETs sont-ils les propagateurs de NIEBs dans les génomes ? La continuation de ces travaux de thèse proposée ici permettra d'apporter des éléments de réponse à ces questions, et de mieux comprendre la co-évolution des séquences et du positionnement nucléosomal.

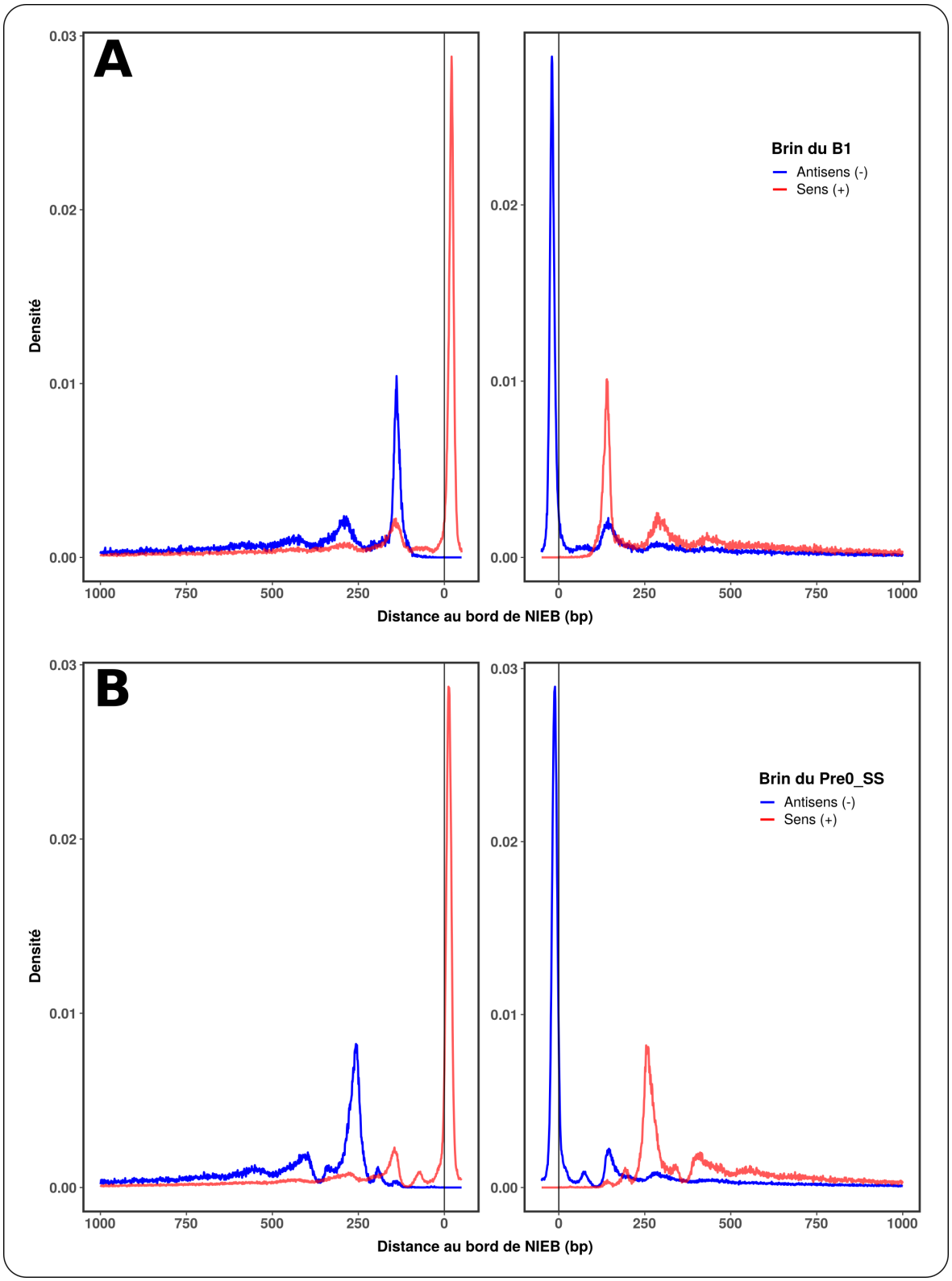


FIGURE 5.1 – **Distribution des** 360 181 **éléments B1 du génome de souris et** 724 559 **éléments Pre0_SS du génome de porc aux bords droits et gauches des barrières nucléosomales.** La partie A représente la distribution des B1, la partie B celle des Pre0_SS. Les courbes rouges représentent les 180 273 (B1) et 362 291 (Pre0_SS) éléments identifiés sur le brin sens. Les courbes bleues représentent les 179 908 (B1) et 362 268 (Pre0_SS) éléments identifiés sur le brin antisens. Le 0 de la partie gauche (resp. droite) de la figure représente le bord gauche (resp. droit) des barrières nucléosomales. Les comptages sont normalisés par le nombre total d'ETs dans chaque catégorie.

A Annexes

Sommaire

A.1 Matériel et logiciels utilisés

A.1.1 Données génomiques

A.1.1.1 Génomes de références

Espèce	Version du génome	Base de données	Date d'accès
Homo Sapiens	hg38	UCSC	12/11/18
Pan Troglodytes	panTro5	UCSC	12/11/20
Sus Scrofa	susScr3	UCSC	18/11/21
Mus Musculus	mm10	UCSC	15/10/18
Gallus Gallus	galGal5	UCSC	18/11/21
Danio Rerio	danRer11	UCSC	14/10/19
Drosophila Melanogaster	dm6	UCSC	14/10/19
Arabidopsis Thaliana	tair10	TAIR	12/03/20
Rosa chinensis	RcHm2	RDP - ENSL	12/03/20
Saccharomyces Cerevisiae	sacCer3	UCSC	24/11/21

TABLEAU A.1 – **Génomes de références utilisés dans le Chapitre 2.** La base de données UCSC se trouve à l'URL suivante : https://hgdownload.soe.ucsc.edu/downloads.html. A partir de celle-ci, on peut retrouver les fichiers fasta utilisés en retrouvant l'organisme concerné dans la liste, puis la version du génome, pour arriver à une URL de type https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/ (ici, pour l'humain, version hg38). Les fichiers utilisés sont ceux avec le nom suivant : version_genome.fa.gz, ainsi que version_genome.chrom.sizes.
La base de données TAIR (The Arabidopsis Information Resource) se trouve à l'URL suivante : https://www.arabidopsis.org/index.jsp.
Le génome du rosier a été obtenu directement auprès des auteurs de la publication décrivant ce génome (Jérémy Just, laboratoire de Reproduction et Développement des Plantes, ENS de Lyon) (Raymond et al., 2018)

A.1.1.2 Chaînes d'alignements de génomes issues du logiciel liftOver

Logiciel liftOver développé à l'UCSC : https://genome.ucsc.edu/cgi-bin/hgLiftOver

Les fichiers de chaînes correspondant aux alignements de génomes complets ont été téléchargés sur la base de données de l'UCSC, à l'adresse suivante : https://hgdownload.cse.ucsc.edu/goldenpath/hg38/liftOver/. Les fichiers utilisés sont les suivants :

— Alignement humain - chimpanzé (hg38-panTro5) : hg38ToPanTro5.over.chain.gz

— Alignement humain - gorille (hg38-gorGor4) : hg38ToGorGor4.over.chain.gz

— Alignement humain - orang-outan (hg38-ponAbe2) : hg38ToPonAbe2.over.chain.gz

A.1.1.3 Positions des éléments transposables

Pour positionner les éléments transposables, j'ai utilisé les annotations d'ETs fournies par l'UCSC avec les génomes de références. Typiquement, pour l'humain, version hg38, elle se trouve à l'URL suivante : https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/rmsk.txt.gz. Il suffit, pour les autres espèces dont le génome a été récupéré dans la base de données de l'UCSC, de

remplacer "hg38" par la version du génome utilisée ("panTro5" ou "susScr3" par exemple). Les dates d'accès sont identiques à celles du **Tableau A.1**.

A.1.2 Liste des logiciels/langages utilisés et de leur version

A.1.2.1 Langages

— Python3 - versions 3.7, 3.8 et 3.9

— Bash - version 4.4.20

— R - versions 3.5.1 à 3.6.3

A.1.2.2 Logiciels et packages principaux

Logiciels

— Bowtie2 - version 2.3.4.1

— Bowtie - version 1.2.2

— Samtools - 1.7 (Utilise htslib v1.7-2)

— Bedtools - version 2.26.0

Principaux packages

— NumPy (python3) - versions 1.15.3 à 1.18.1

— Biopython (python3) - versions 1.73 à 1.79

— ggplot2 (R) - version 3.0.0 à 3.3.5

— Tidyverse (R) - version 1.3.1

A.2 Figures

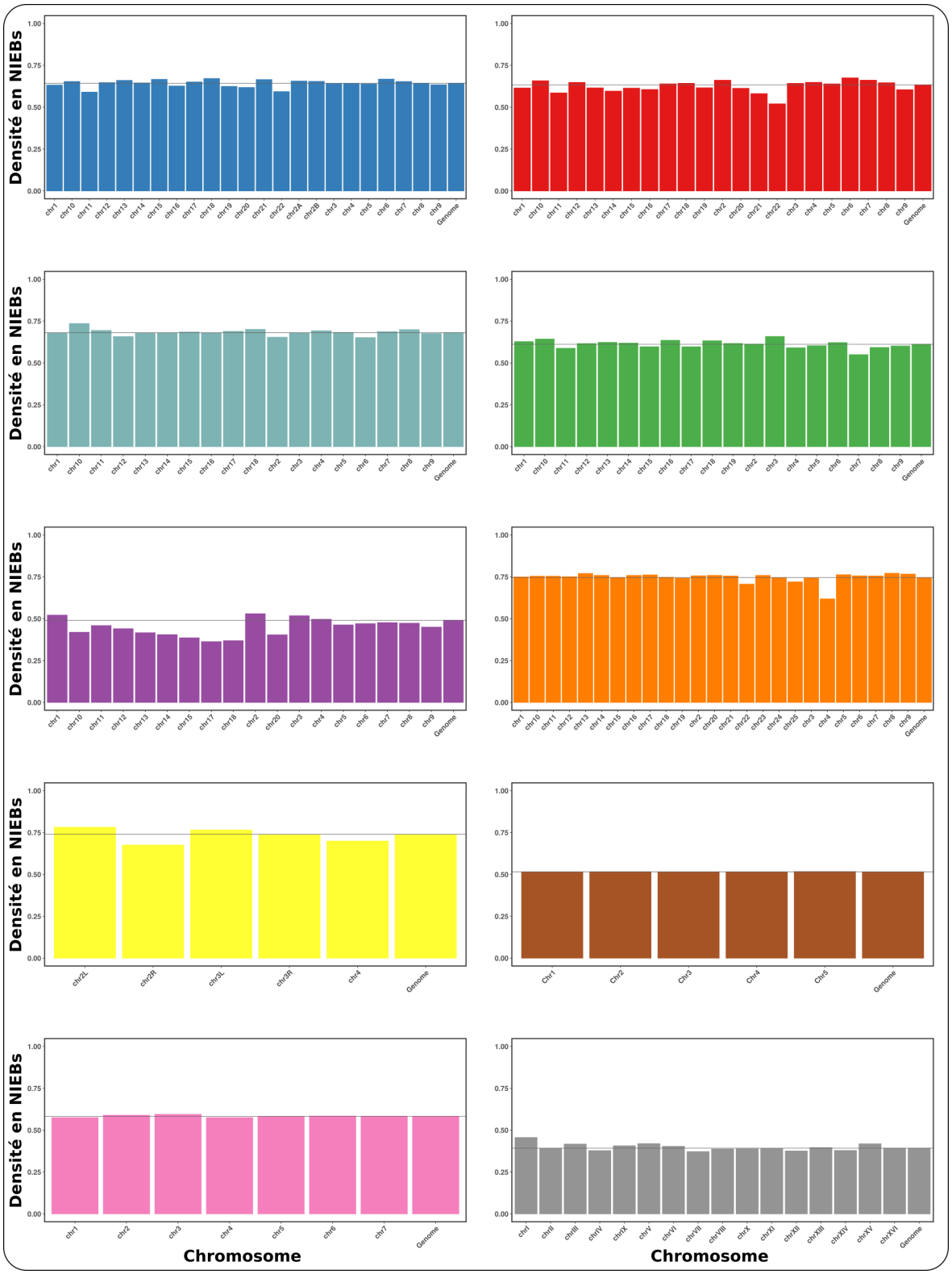


FIGURE A.1 – **Densité en NIEBs dans chaque chromosome des 10 génomes analysés.** Chaque figure correspond à un génome. Les couleurs correspondent à celles utilisées pour la **figure 2.1**. Sur chaque figure, une ligne horizontale grise indique la densité en NIEBs sur le génome complet.

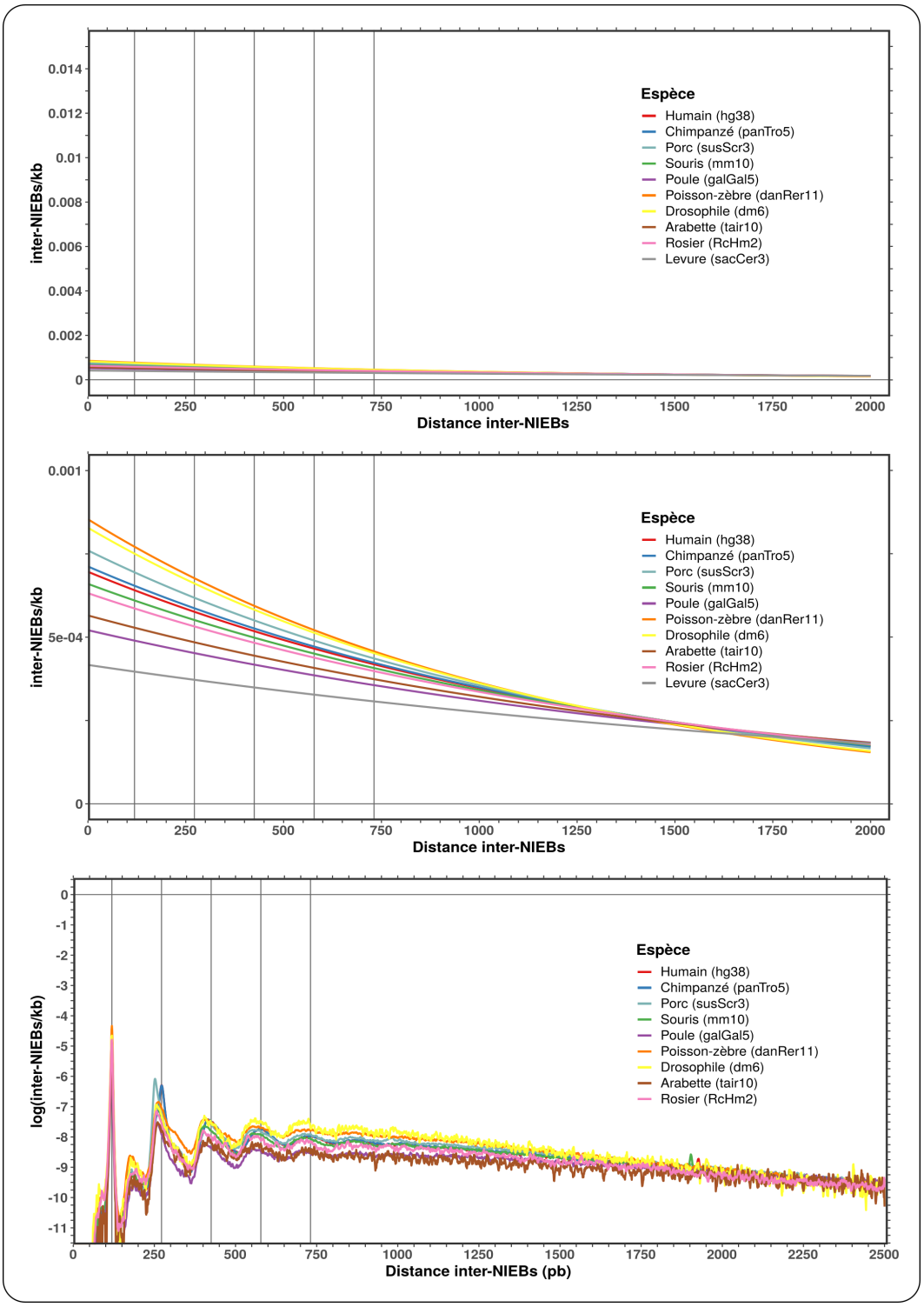


FIGURE A.2 – **Tailles d'inter-NIEBs attendues dans le cas d'une distribution aléatoire des barrières nucléosomales.** La figure du haut correspond aux courbes attendues selon les espèces, avec la même échelle que celle utilisée pour la **Figure 2.3**. La figure du milieu est un zoom de la partie haute, entre 0 et 0.001. La figure du bas correspond au résultat obtenu en passant les données de la **Figure 2.3** en échelle logarithmique. On observe qu'à partir de la distance 1000 pb, les courbes forment une droite à pente négative, ce qui confirme la distribution exponentielle des distances inter-NIEBs à partir des distances 1000 pb. Les couleurs correspondent à celles utilisées pour la **Figure 2.3**. Les moyennes utilisées pour générer les distributions attendues sont celles répertoriées dans le **Tableau 2.2**. Les lignes verticales correspondent aux abscisses 119, 272, 425, 578 et 731 pb, comme dans la **Figure 2.3**. On note que la levure a été retirée de la figure du bas car un nombre trop important de distance inter-NIEBs ne sont pas représentées dans cette espèce, et cela rendait la figure illisible.

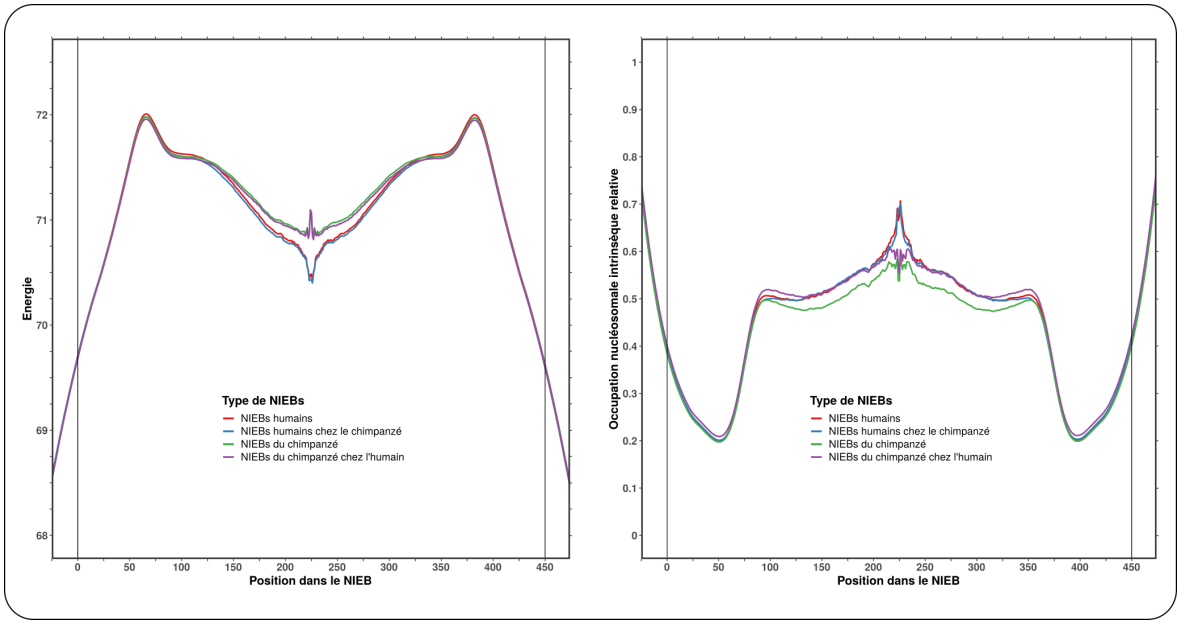


FIGURE A.3 – **Profils d'énergie de formation du nucléosome et d'occupation nucléosomale prédits pour tous les couples de NIEBs homologues.** Les panneaux A et B représentent respectivement les profils d'énergie et d'occupation nucléosomale prédits par le modèle. Les courbes rouge et bleue sont à comparer aux courbes de mêmes couleurs de la **Figure 2.13 - A et C**. Les courbes verte et violette sont à comparer aux courbes de mêmes couleurs de la **Figure 2.13 - B et D**. Chacune de ces courbes a été obtenue à en faisant la moyenne des 810 002 couples de NIEBs identifiés entre l'humain et le chimpanzé, quelle que soit leur couverture mutuelle. L'axe des abscisses représente la même chose que dans la **Figure 2.13**.

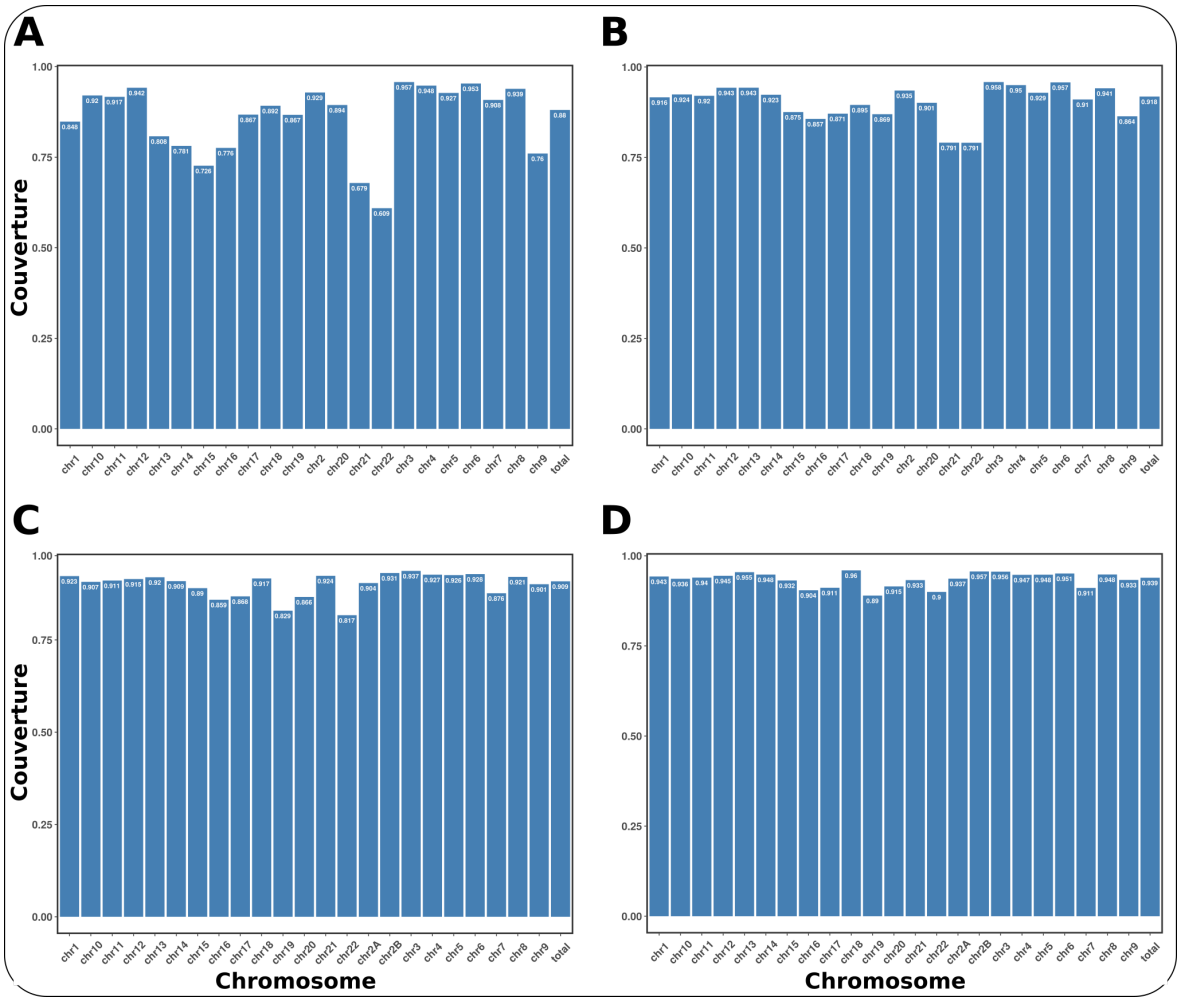


FIGURE A.4 – **Couverture chromosome par chromosome des génomes de l'humain et du chimpanzé par les intervalles alignés entre les deux espèces.** Les panneaux A (humain) et C (chimpanzé) représentent la couverture des chromosomes complets, indépendamment de si les séquences sont connues ou non. Les panneaux B (humain) et D (chimpanzé) représentent la couverture des chromosomes en se concentrant sur les zones dont les séquences sont connues (en retirant les bases N de l'analyse).

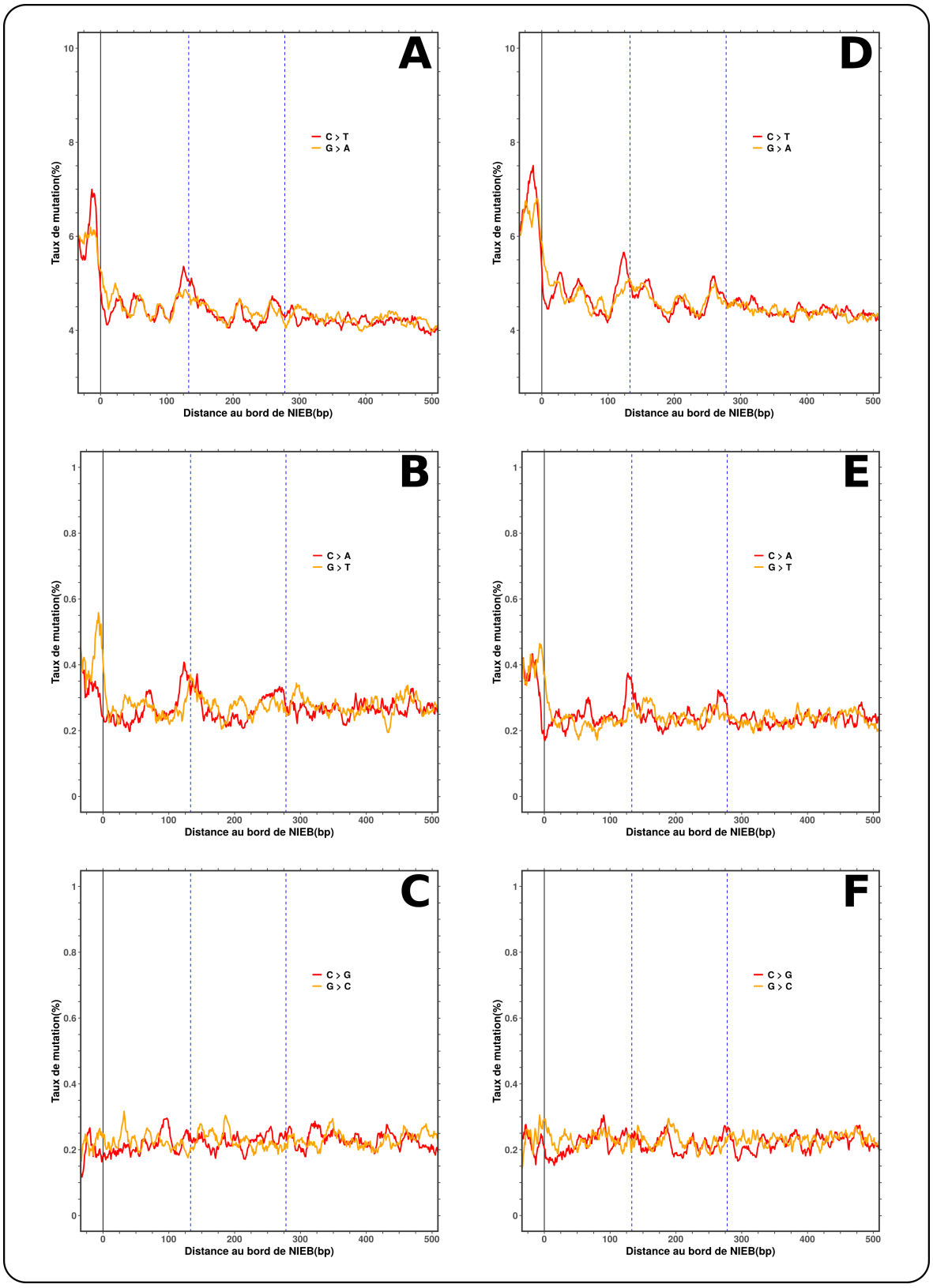


FIGURE A.5 – **Taux de substitutions des bases C en contexte CpG aux bords des barrières nucléosomales chez l'humain et le chimpanzé. A/D** Taux de substitutions C>T (rouge) et G>A (orange) chez l'humain (A) et le chimpanzé (D). **B/E** Taux de substitutions C>A (rouge) et G>T (orange) chez l'humain (B) et le chimpanzé (E). **C/F** Taux de substitutions C>G (rouge) et G>C (orange) chez l'humain (C) et le chimpanzé (F). Les lignes verticales noires correspondent à l'abscisse 0. Les lignes verticales pointillées bleues correspondent aux abscisses 133 et 278 (minima du %GC au bord des NIEBs de l'humain). Les courbes sont lissées sur 10 pb, et représentent les moyennes des taux obtenus à droite des NIEBs (bord 3') et du complémentaire inverse des substitutions à gauche des NIEBs (bord 5').

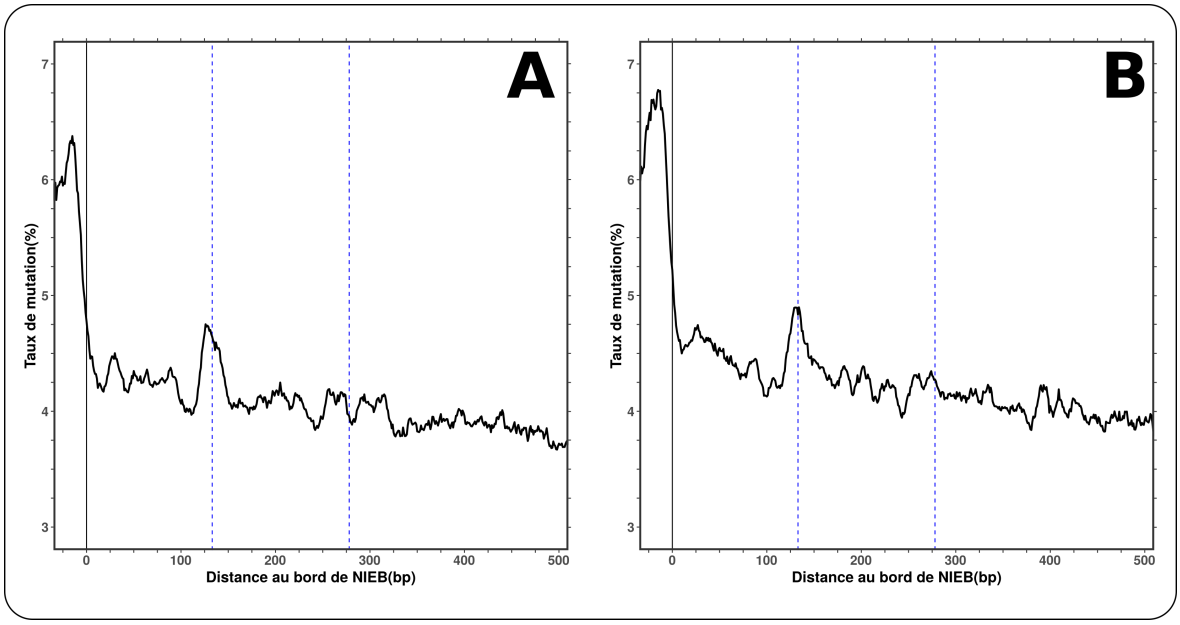


FIGURE A.6 – **Taux de substitutions C vers T et G vers A en contexte CpG aux bords des barrières nucléosomales chez l'humain et le chimpanzé en l'absence d'élément Alu.** La courbe correspond à la moyenne des deux taux de substitutions complémentaires. **A** Taux de substitutions chez l'humain. **B** Taux de substitutions chez le chimpanzé. La ligne verticale noire correspond à l'abscisse 0 pb. Les lignes verticales pointillées bleues correspondent aux abscisses 133 et 278 (minima du %GC au bord des NIEBs de l'humain). Les courbes sont lissées sur 10 pb, et représentent les moyennes des taux obtenus à droite des NIEBs (bord 3') et du complémentaire inverse des substitutions à gauche des NIEBs (bord 5').

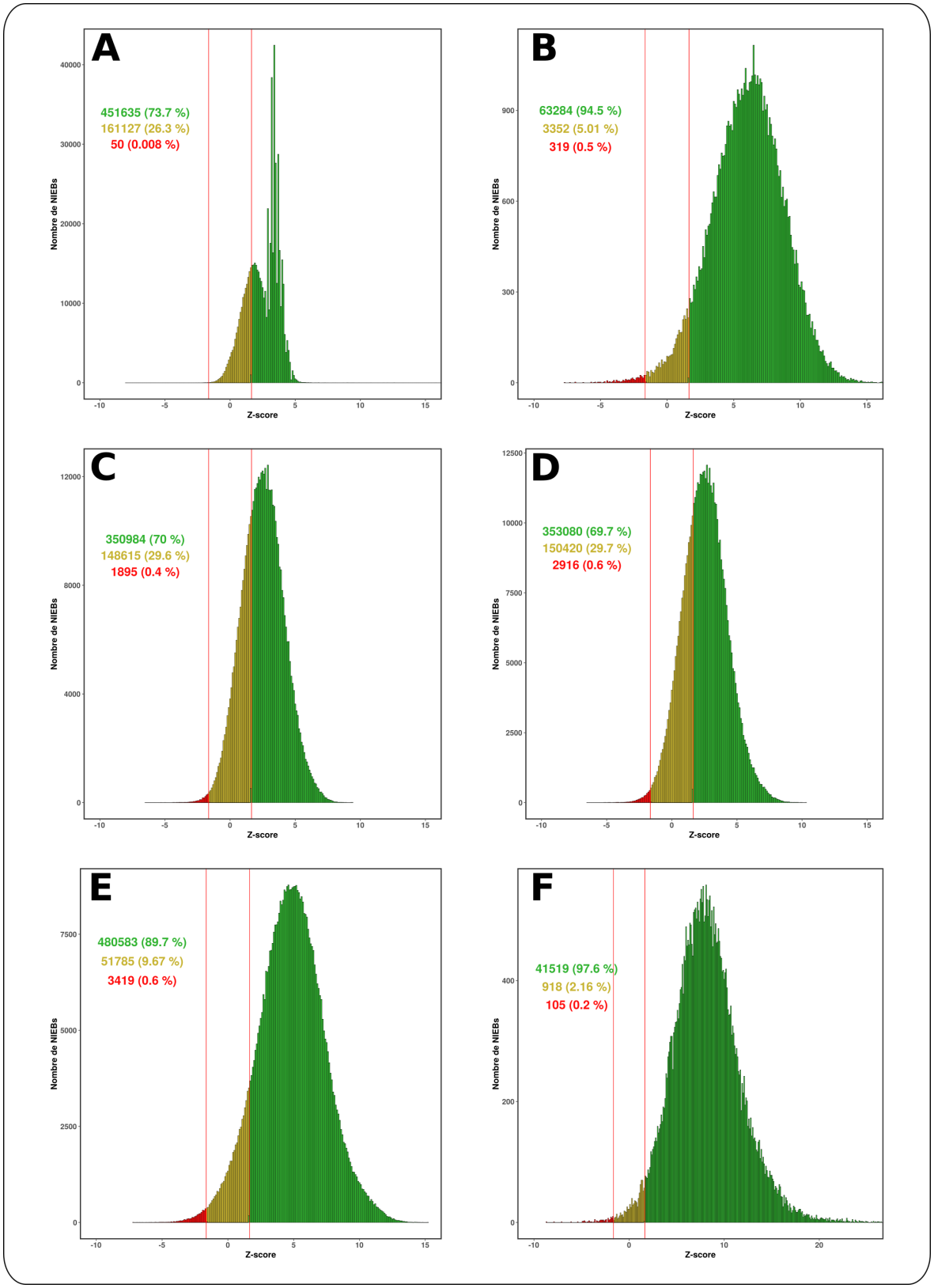


FIGURE A.7 – **Distributions des zscores des barrières nucléosomales chez la souris, la drosophile, le poisson-zèbre et l'arabette.** Sont représentés les résultats obtenus chez la souris (Mieczkowski - A), la drosophile (Chereji - B), le poisson-zèbre (Zhang 1 - C, Zhang 3 - D, Zhang 4 - E) et l'arabette (Pass 2 - F). Les données utilisées pour chaque espèce sont résumées dans le **Tableau 3.2**. Les z-scores ont été obtenus en utilisant le pipeline décrit en **Partie 3.2.1**.

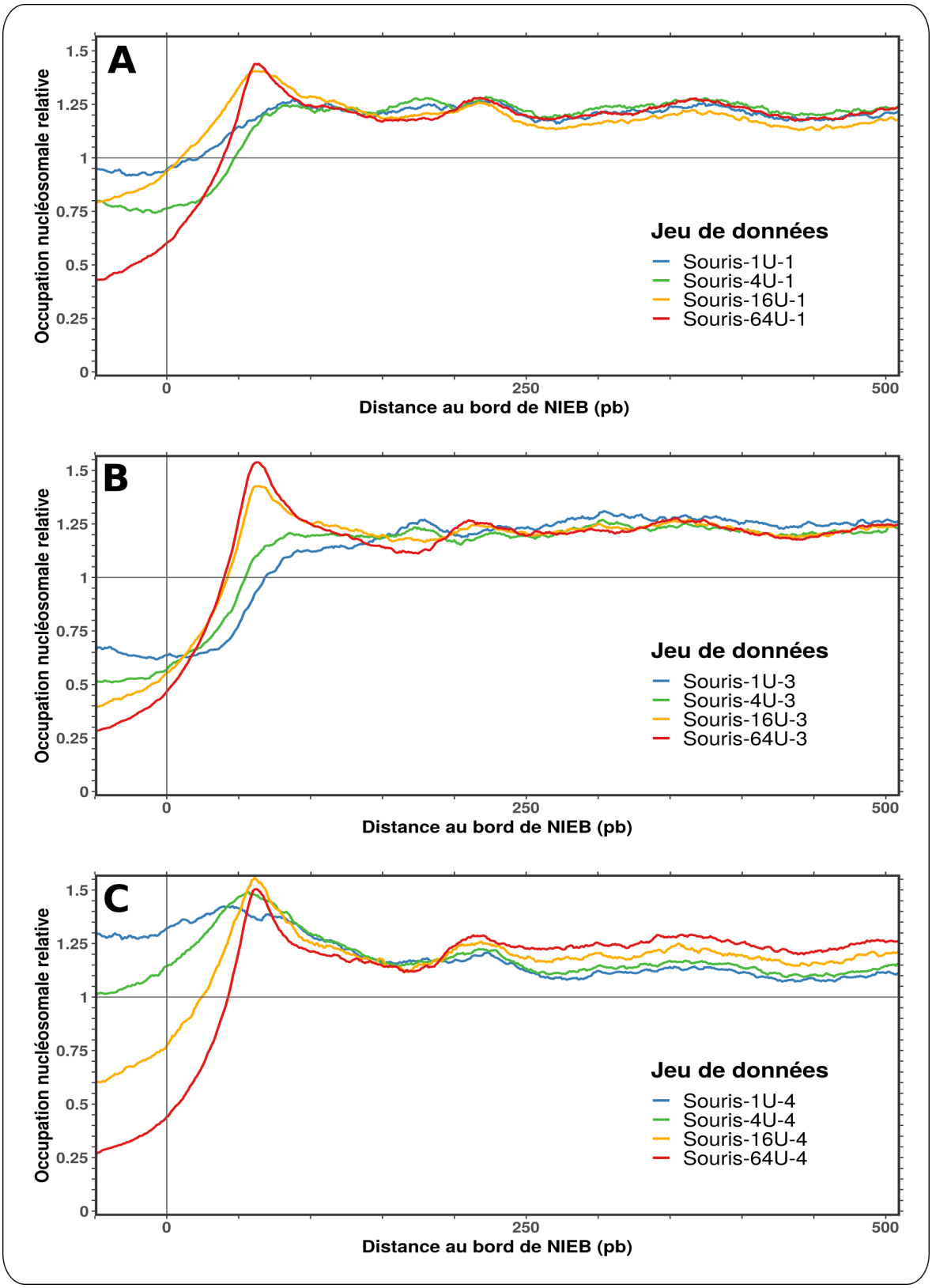


FIGURE A.8 – **Profils d'occupation en nucléosomes aux bords des barrières nucléosomales dans trois lignées cellulaires de souris selon le niveau de digestion de la chromatine.** Sont représentés ici les résultats obtenus avec les lignées J1 ESC (A), NPC (B) et eNPC (C) (voir **Tableau 3.3** pour le détail des données utilisées). Les profils obtenus en utilisant le pipeline décrit en **Partie 3.2.1** sont normalisés par la moyenne génomique. Les courbes rouges, jaunes, vertes et bleues correspondent sur chaque graphique respectivement aux digestions complètes, moyennes, légères et très légères. Les paramètres MACC, coefficients de la régression linéaire calculés selon la méthode décrite en **Partie 3.3.1** sont, respectivement pour les positions -35, 65 et 450 : (A) -0.064 ; 0.045 ; -0.00072 ; (B) -0.053 ; 0.101 ; -0.008 ; (C) -0.161 ; 0.026 ; 0.017.

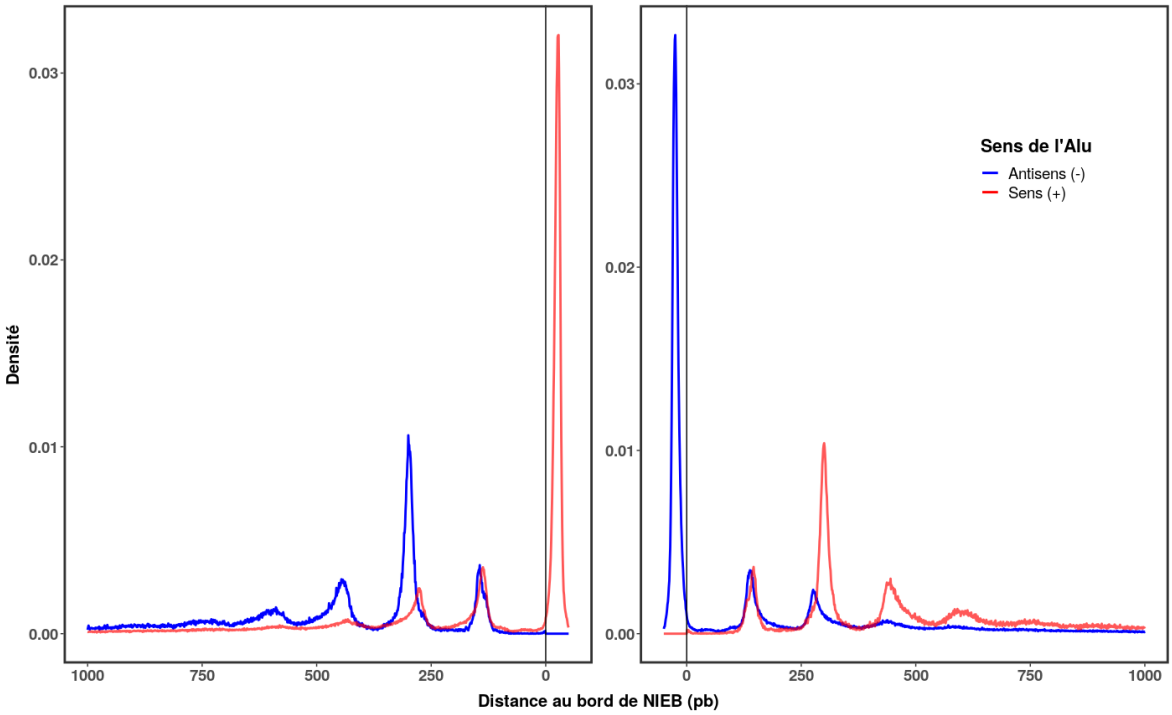


FIGURE A.9 – **Distribution des** 1 059 895 **éléments Alu des autosomes du chimpanzé aux bords droits et gauches des barrières nucléosomales.** Le 0 de la partie gauche (resp. droite) de la figure représente le bord gauche (resp. droit) des barrières nucléosomales. La courbe rouge de la partie gauche (reps. droite) représente les 387 225 (reps. 140 098) Alu identifiés sur le brin sens à gauche (resp. droite) d'un NIEB. La courbe bleue de la partie gauche (reps. droite) représente les 141 224 (reps. 391 348) Alu identifiés sur le brin antisens à gauche (reps. droite) d'un NIEB. Chaque courbe est normalisée par le nombre d'Alu de la catégorie correspondantes.

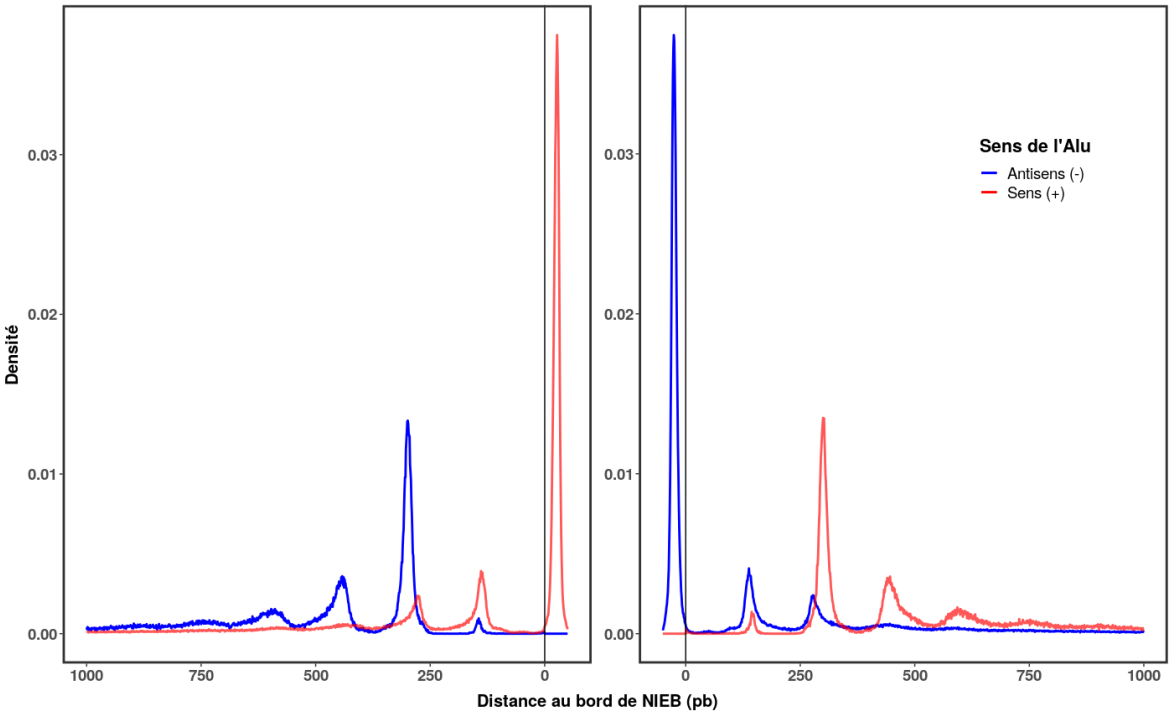


FIGURE A.10 – **Distribution des** 844 057 **éléments Alu de tailles supérieures à** 250 **pb des autosomes humains aux bords droits et gauches des barrières nucléosomales.** Le 0 de la partie gauche (resp. droite) de la figure représente le bord gauche (resp. droit) des barrières nucléosomales. La courbe rouge de la partie gauche (reps. droite) représente les 319 147 (reps. 102 747) Alu identifiés sur le brin sens à gauche (resp. droite) d'un NIEB. La courbe bleue de la partie gauche (reps. droite) représente les 102 306 (reps. 319 857) Alu identifiés sur le brin antisens à gauche (reps. droite) d'un NIEB. Chaque courbe est normalisée par le nombre d'Alu de la catégorie correspondantes.

A.3 Conférences

A.3.1 Communications orales

Réunion du groupement de recherche "Architecture et Dynamique du Noyau et des Génomes"

— Septembre 2021 - Millau

— https://indico.in2p3.fr/event/21937/

— Nucleosome positioning by nucleosome inhibitory energy barriers is mediated by Alu element transposition

Réunion du groupement de recherche MobilET

— 14 Septembre 2021 - Paris

— Alu transposition success and sequence-encoded nucleosomal positioning

Rencontres ALignement et PHYlogénie (ALPHY)

— Février 2019 - Paris

— https://lbbe-dmz.univ-lyon1.fr/spip_alphy/spip.php?article70

— Alu transposable elements and their interactions with the DNA sequence-encoded nucleosome inhibitory energy barriers

Congrès National sur les Eléments Transposables (CNET)

— Juillet 2018 - Clermont-Ferrand

— https://www.mobil-et.eu/en/french-meeting/cnet-2018/

— Coupling between Alu transposable element insertions and sequence-encoded nucleosome inhibitory barriers

A.3.2 Poster

5th Uppsala Transposon Symposium

— Octobre 2021 Visio-conférence

— https://transposonsymposium.wordpress.com/

— Coupling between Alu transposable element insertions and sequence-encoded nucleosome inhibitory barriers

Liste complète des références

C. Ade, A. M. Roy-Engel & P. L. Deininger (2013). Alu elements : an intrinsic source of human genome instability. *Current Opinion in Virology* **3**, 639–645. 44, 45

M. Ahmed, W. Li & P. Liang (2013). Identification of three new Alu Yb subfamilies by source tracking of recently integrated Alu Yb elements. *Mobile DNA* **4**, 25. 42

R. Ammar, D. Torti, K. Tsui, M. Gebbia, T. Durbic, G. D. Bader, G. Giaever & C. Nislow (2012). Chromatin is an ancient innovation conserved between Archaea and Eukarya. *eLife* **1**, e00078. 2

G. Annibalini, P. Bielli, M. De Santi, D. Agostini, M. Guescini, D. Sisti, S. Contarelli, G. Brandi et al. (2016). MIR retroposon exonization promotes evolutionary variability and generates species-specific expression of IGF-1 splice variants. *Biochimica Et Biophysica Acta* **1859**, 757–768. 39

A. Antonaki, C. Demetriades, A. Polyzos, A. Banos, G. Vatsellas, M. D. Lavigne, E. Apostolou, E. Mantouvalou et al. (2011). Genomic analysis reveals a novel nuclear factor-κB (NF-κB)-binding site in Alu-repetitive elements. *The Journal of Biological Chemistry* **286**, 38768–38782. 47

J. Barau, A. Teissandier, N. Zamudio, S. Roy, V. Nalesso, Y. Hérault, F. Guillou & D. Bourc'his (2016). The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science (New York, N.Y.)* **354**, 909–912. 40

M. Barbi, J. Mozziconacci, H. Wong & J.-M. Victor (2014). DNA topology in chromosomes : a quantitative survey and its physiological implications. *Journal of Mathematical Biology* **68**, 145–179. 3

J. Barbier, C. Vaillant, J.-N. Volff, F. G. Brunet & B. Audit (2021). Coupling between Sequence-Mediated Nucleosome Organization and Genome Evolution. *Genes* **12**, 851. 4, 5, 29, 41, 48, 76, 86, 148

T. Barnes & P. Korber (2021). The Active Mechanism of Nucleosome Depletion by Poly(dA :dT) Tracts In Vivo. *International Journal of Molecular Sciences* **22**, 8233. 48

M. A. Batzer & P. L. Deininger (2002). Alu repeats and human genomic diversity. *Nature Reviews. Genetics* **3**, 370–379. 41

M. Ben Maamar, D. Beck, E. Nilsson, J. R. McCarrey & M. K. Skinner (2020). Developmental origins of transgenerational sperm histone retention following ancestral exposures. *Developmental Biology* **465**, 31–45. 4

E. A. Bennett, H. Keller, R. E. Mills, S. Schmidt, J. V. Moran, O. Weichenrieder & S. E. Devine (2008). Active Alu retrotransposons in the human genome. *Genome Research* **18**, 1875–1883. 42

D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59. 89

D. A. Beshnova, A. G. Cherstvy, Y. Vainshtein & V. B. Teif (2014). Regulation of the nucleosome repeat length in vivo by the DNA sequence, protein concentrations and long-range interactions. *PLoS computational biology* **10**, e1003698. 65

A. P. Bird (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research* **8**, 1499–1504. 78, 84

C. Biémont (2010). A Brief History of the Status of Transposable Elements : From Junk DNA to Major Players in Evolution. *Genetics* **186**, 1085–1093. 32, 37

C. Biémont & C. Vieira (2006). Junk DNA as an evolutionary force. *Nature* **443**, 521–524. Number : 7111 Publisher : Nature Publishing Group. 32, 50

G. Bourque (2009). Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Current Opinion in Genetics & Development* **19**, 607–612. 40

G. Bourque, K. H. Burns, M. Gehring, V. Gorbunova, A. Seluanov, M. Hammell, M. Imbeault, Z. Izsvák et al. (2018). Ten things you should know about transposable elements. *Genome Biology* **19**, 199. 35

J. Brosius (1991). Retroposons–seeds of evolution. *Science (New York, N.Y.)* **251**, 753. 33

F. G. Brunet, B. Audit, G. Drillon, F. Argoul, J.-N. Volff & A. Arneodo (2018). Evidence for DNA Sequence Encoding of an Accessible Nucleosomal Array across Vertebrates. *Biophysical Journal* . 30, 48, 52, 61, 63, 108, 120, 125, 126, 152

W. D. Burke, C. C. Calalang & T. H. Eickbush (1987). The site-specific ribosomal insertion element type II of Bombyx mori (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme. *Molecular and Cellular Biology* **7**, 2221–2230. 36

K. H. Burns (2020). Our Conflict with Transposable Elements and Its Implications for Human Disease. *Annual Review of Pathology* **15**, 51–70. 45

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode C. elegans : a platform for investigating biology. *Science (New York, N.Y.)* **282**, 2012–2018. 33

P. A. Callinan, J. Wang, S. W. Herke, R. K. Garber, P. Liang & M. A. Batzer (2005). Alu retrotransposition-mediated deletion. *Journal of Molecular Biology* **348**, 791–800. 47

J. Cappello, K. Handelsman & H. F. Lodish (1985). Sequence of Dictyostelium DIRS-1 : an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. *Cell* **43**, 105–115. 35

B. R. Carone, J.-H. Hung, S. J. Hainer, M.-T. Chou, D. M. Carone, Z. Weng, T. G. Fazzio & O. J. Rando (2014). High resolution mapping of chromatin packaging in mouse ES cells and sperm. *Developmental cell* **30**, 11–22. 99, 100

I. L. Cartwright & S. C. Elgin (1986). Nucleosomal instability and induction of new upstream protein-DNA associations accompany activation of four small heat shock protein genes in Drosophila melanogaster. *Molecular and Cellular Biology* **6**, 779–791. 100

D. Chalopin, M. Naville, F. Plard, D. Galiana & J.-N. Volff (2015). Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates. *Genome Biology and Evolution* **7**, 567–580. 41

J. Chen, E. Li & J. Lai (2017a). The coupled effect of nucleosome organization on gene transcription level and transcriptional plasticity. *Nucleus (Austin, Tex.)* **8**, 605–612. 76

J. Chen, E. Li, X. Zhang, X. Dong, L. Lei, W. Song, H. Zhao & J. Lai (2017b). Genome-wide Nucleosome Occupancy and Organization Modulates the Plasticity of Gene Transcriptional Status in Maize. *Molecular Plant* **10**, 962–974. 76

X. Chen, Z. Chen, H. Chen, Z. Su, J. Yang, F. Lin, S. Shi & X. He (2012). Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. *Science (New York, N.Y.)* **335**, 1235–1238. 4, 76

R. V. Chereji, J. Ocampo & D. J. Clark (2017). MNase-Sensitive Complexes in Yeast : Nucleosomes and Non-histone Barriers. *Molecular Cell* **65**, 565–577.e3. 104, 106

R. V. Chereji, T. D. Bryson & S. Henikoff (2019). Quantitative MNase-seq accurately maps nucleosome occupancy levels. *Genome Biology* **20**, 198. 99, 100, 104

G. Chevereau, A. Arneodo & C. Vaillant (2011). Influence of the genomic sequence on the primary structure of chromatin. *Frontiers in Life Science* **5**, 29–68. 29, 50, 73

J. Choi, D. B. Lyons, M. Y. Kim, J. D. Moore & D. Zilberman (2020). DNA Methylation and Histone H1 Jointly Repress Transposable Elements and Aberrant Intragenic Transcripts. *Molecular Cell* **77**, 310–323.e7. 100

D. M. Church, V. A. Schneider, K. M. Steinberg, M. C. Schatz, A. R. Quinlan, C.-S. Chin, P. A. Kitts, B. Aken et al. (2015). Extending reference assembly models. *Genome Biology* **16**, 13. 83

B. Chénais, A. Caruso, S. Hiard & N. Casse (2012). The impact of transposable elements on eukaryotic genomes : from genome size increase to genetic adaptation to stressful environments. *Gene* **509**, 7–15. 41

M. Cobb (2017). 60 years ago, Francis Crick changed the logic of biology. *PLoS Biology* **15**, e2003243. 32

C. K. Collings & J. N. Anderson (2017). Links between DNA methylation and nucleosome occupancy in the human genome. *Epigenetics & Chromatin* **10**, 18. 86

R. Cordaux, D. J. Hedges, S. W. Herke & M. A. Batzer (2006). Estimating the retrotransposition rate of human Alu elements. *Gene* **373**, 134–137. 42

D. Cotnoir-White, D. Laperrière & S. Mader (2011). Evolution of the repertoire of nuclear receptor binding sites in genomes. *Molecular and Cellular Endocrinology* **334**, 76–82. 47

C. Coulondre, J. H. Miller, P. J. Farabaugh & W. Gilbert (1978). Molecular basis of base substitution hotspots in Escherichia coli. *Nature* **274**, 775–780. 78, 84

F. H. Crick (1958). On protein synthesis. *Symposia of the Society for Experimental Biology* **12**, 138–163. 32

A. R. Cutter & J. J. Hayes (2015). A brief review of nucleosome structure. *FEBS Letters* **589**, 2914–2922. 2

P. Deininger (2011). Alu elements : know the SINEs. *Genome biology* **12**, 236. 33, 34, 43, 44, 45, 46, 125, 129, 133

M. Dewannieux, C. Esnault & T. Heidmann (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nature Genetics* **35**, 41–48. 34, 43

P. M. Diesinger & D. W. Heermann (2009). Depletion effects massively change chromatin properties and influence genome folding. *Biophysical Journal* **97**, 2146–2153. 65

C. Dingwall, G. P. Lomonossoff & R. A. Laskey (1981). High sequence specificity of micrococcal nuclease. *Nucleic Acids Research* **9**, 2659–2673. 104

S. Dridi (2012). Alu mobile elements : from junk DNA to genomic gems. *Scientifica* **2012**, 545328. 153

G. Drillon, B. Audit, F. Argoul & A. Arneodo (2015). Ubiquitous human 'master' origins of replication are encoded in the DNA sequence via a local enrichment in nucleosome excluding energy barriers. *Journal of Physics : Condensed Matter* **27**, 064102. 29, 89, 108

G. Drillon, B. Audit, F. Argoul & A. Arneodo (2016). Evidence of selection for an accessible nucleosomal array in human. *BMC Genomics* **17**. 29, 30, 31, 48, 50, 52, 53, 54, 55, 56, 59, 63, 65, 76, 78, 81, 83, 84, 86, 88, 89, 93, 94, 97, 98, 108, 149, 152

R. M. Durbin, D. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, F. S. Collins et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073. Number : 7319 Publisher : Nature Publishing Group. 31, 83, 126, 131, 149

E. Englander & H. Bruce (1995). Nucleosome positioning by human Alu elements in chromatin. *The journal of biological chemistry* **270**, 10091–10096. 48

E. W. Englander, A. P. Wolffe & B. H. Howard (1993). Nucleosome interactions with a human Alu element. Transcriptional repression and effects of template methylation. *The Journal of Biological Chemistry* **268**, 19565–19573. 48

E. Etchegaray, M. Naville, J.-N. Volff & Z. Haftek-Terreau (2021). Transposable element-derived sequences in vertebrate development. *Mobile DNA* **12**, 1. 38, 39

R. Everaers & H. Schiessel (2015). The physics of chromatin. *Journal of Physics. Condensed Matter : An Institute of Physics Journal* **27**, 060301. 65

M. B. Evgen'ev & I. R. Arkhipova (2005). Penelope-like elements–a new class of retroelements : distribution, function and possible evolutionary significance. *Cytogenetic and Genome Research* **110**, 510–521. 35

M. B. Evgen'ev, H. Zelentsova, N. Shostak, M. Kozitsina, V. Barskyi, D. H. Lankenau & V. G. Corces (1997). Penelope, a new family of transposable elements and its possible role in hybrid dysgenesis in Drosophila virilis. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 196–201. 35

Y. Field, N. Kaplan, Y. Fondufe-Mittendorf, I. K. Moore, E. Sharon, Y. Lubling, J. Widom & E. Segal (2008). Distinct Modes of Regulation by Chromatin Encoded through Nucleosome Positioning Signals. *PLoS Computational Biology* **4**. 50, 127, 133

Y. Field, Y. Fondufe-Mittendorf, I. K. Moore, P. Mieczkowski, N. Kaplan, Y. Lubling, J. D. Lieb, J. Widom et al. (2009). Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. *Nature Genetics* **41**, 438–445. 4, 76

A. Ganguly, T. Dunbar, P. Chen, L. Godmilow & T. Ganguly (2003). Exon skipping caused by an intronic insertion of a young Alu Yb9 element leads to severe hemophilia A. *Human Genetics* **113**, 348–352. 46

N. Gilbert, S. Boyle, H. Fiegler, K. Woodfine, N. P. Carter & W. A. Bickmore (2004). Chromatin architecture of the human genome : gene-rich domains are enriched in open chromatin fibers. *Cell* **118**, 555–566. 151

A. Gonzalez-Perez, R. Sabarinathan & N. Lopez-Bigas (2019). Local Determinants of the Mutational Landscape of the Human Genome. *Cell* **177**, 101–114. 84

T. J. D. Goodwin & R. T. M. Poulter (2004). A new group of tyrosine recombinase-encoding retrotransposons. *Molecular Biology and Evolution* **21**, 746–759. 35

I. M. Greenblatt & R. A. Brink (1962). Twin Mutations in Medium Variegated Pericarp Maize. *Genetics* **47**, 489–501. 36

M. A. M. Groenen, A. L. Archibald, H. Uenishi, C. K. Tuggle, Y. Takeuchi, M. F. Rothschild, C. Rogel-Gaillard, C. Park et al. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491**, 393–398. 153

D. Gussakovsky & S. A. McKenna (2021). Alu RNA and their roles in human disease states. *RNA biology* **18**, 574–585. 45

J. S. Han & J. D. Boeke (2005). LINE-1 retrotransposons : modulators of quantity and quality of mammalian gene expression? *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology* **27**, 775–784. 34

D. C. Hancks & H. H. Kazazian (2016). Roles for retrotransposon insertions in human disease. *Mobile DNA* **7**, 9. 37, 38

S. Henikoff, J. G. Henikoff, A. Sakai, G. B. Loeb & K. Ahmad (2009). Genome-wide profiling of salt fractions maps physical properties of chromatin. *Genome Research* **19**, 460–469. 111

C. M. Houck, F. P. Rinehart & C. W. Schmid (1979). A ubiquitous family of repeated DNA sequences in the human genome. *Journal of Molecular Biology* **132**, 289–306. 41

H. Ishii, J. T. Kadonaga & B. Ren (2015). MPE-seq, a new method for the genome-wide analysis of chromatin structure. *Proceedings of the National Academy of Sciences* **112**, E3457–E3465. 92

P.- Jacques, J. Jeyakani & G. Bourque (2013). The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS genetics* **9**, e1003504. 40

H. S. Jang, W. J. Shin, J. E. Lee & J. T. Do (2017). CpG and Non-CpG Methylation in Epigenetic Gene Regulation and Brain Function. *Genes* **8**, 148. 150

N. Jansz (2019). DNA methylation dynamics at transposable elements in mammals. *Essays in Biochemistry* **63**, 677–689. 40

M. Kajikawa & N. Okada (2002). LINEs mobilize SINEs in the eel through a shared 3' sequence. *Cell* **111**, 433–444. 34

T. Kaneko-Ishino & F. Ishino (2012). The role of genes domesticated from LTR retrotransposons and retroviruses in mammals. *Frontiers in Microbiology* **3**, 262. 38

V. V. Kapitonov & J. Jurka (2001). Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 8714–8719. 36

V. V. Kapitonov & J. Jurka (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews. Genetics* **9**, 411–412 ; author reply 414. 33

H. H. Kazazian (1998). Mobile elements and disease. *Current Opinion in Genetics & Development* **8**, 343–350. 37

N. Kepper, D. Foethke, R. Stehr, G. Wedemann & K. Rippe (2008). Nucleosome geometry and internucleosomal interactions control the chromatin fiber conformation. *Biophysical Journal* **95**, 3692–3705. 65

M. G. Kidwell (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**, 49–63. 41

S. Kim, C.-S. Cho, K. Han & J. Lee (2016). Structural Variation of Alu Element and Human Disease. *Genomics & Informatics* **14**, 70–77. 38, 45, 46, 47

Y.-J. Kim, J. Lee & K. Han (2012). Transposable Elements : No More 'Junk DNA'. *Genomics & Informatics* **10**, 226–233. 47

B. Knight, S. Kubik, B. Ghosh, M. J. Bruzzone, M. Geertz, V. Martin, N. Dénervaud, P. Jacquet et al. (2014). Two distinct promoter architectures centered on dynamic nucleosomes control ribosomal protein gene transcription. *Genes & Development* **28**, 1695–1709. 104, 106

R. D. Kornberg (1974). Chromatin structure : a repeating unit of histones and DNA. *Science (New York, N.Y.)* **184**, 868–871. 2

R. D. Kornberg (1977). Structure of chromatin. *Annual Review of Biochemistry* **46**, 931–954. 2

R. D. Kornberg & L. Stryer (1988). Statistical distributions of nucleosomes : nonrandom locations by a stochastic mechanism. *Nucleic Acids Research* **16**, 6677–6690. 29

D. A. Kramerov & N. S. Vassetzky (2005). Short retroposons in eukaryotic genomes. *International Review of Cytology* **247**, 165–221. 34

J. O. Kriegs, G. Churakov, J. Jurka, J. Brosius & J. Schmitz (2007). Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends in genetics : TIG* **23**, 158–161. 42

A. Kumar & J. L. Bennetzen (1999). Plant retrotransposons. *Annual Review of Genetics* **33**, 479–532. 34

D. Labuda, D. Sinnett, C. Richer, J. M. Deragon & G. Striker (1991). Evolution of mouse B1 repeats : 7SL RNA folding pattern conserved. *Journal of Molecular Evolution* **32**, 405–414. 153

E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921. 45

B. Langmead & S. L. Salzberg (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359. Bandiera_abtest : a Cg_type : Nature Research Journals Number : 4 Primary_atype : Research Publisher : Nature Publishing Group Subject_term : Bioinformatics ;Genomics ;Sequencing Subject_term_id : bioinformatics ;genomics ;sequencing. 91

B. Langmead, C. Trapnell, M. Pop & S. L. Salzberg (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25. 91

A. Le Rouzic & P. Capy (2005). The First Steps of Transposable Elements Invasion. *Genetics* **169**, 1033–1043. 32

A. Lesne & J.-M. Victor (2006). Chromatin fiber functional organization : some plausible models. *The European Physical Journal. E, Soft Matter* **19**, 279–290. 65

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079. 91

D. Lisch (2009). Epigenetic regulation of transposable elements in plants. *Annual Review of Plant Biology* **60**, 43–66. 40

G. E. Liu, C. Alkan, L. Jiang, S. Zhao & E. E. Eichler (2009). Comparative analysis of Alu repeats in primate genomes. *Genome Research* **19**, 876–885. 47

P. T. Lowary & J. Widom (1998). New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *Journal of Molecular Biology* **276**, 19–42. 48

D. D. Luan, M. H. Korman, J. L. Jakubczak & T. H. Eickbush (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site : a mechanism for non-LTR retrotransposition. *Cell* **72**, 595–605. 44

K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent & T. J. Richmond (1997). core particle at 2.8 A resolution. *Nature* **389**, 18. 2, 3

W. Makałowski, G. A. Mitchell & D. Labuda (1994). Alu sequences in the coding regions of mRNA : a source of protein variability. *Trends in genetics : TIG* **10**, 188–193. 33

W. Makałowski, V. Gotea, A. Pande & I. Makałowska (2019). Transposable Elements : Classification, Identification, and Their Use As a Tool For Comparative Genomics. *Methods in Molecular Biology (Clifton, N.J.)* **1910**, 177–207. 33, 34

K. D. Makova & R. C. Hardison (2015). The effects of chromatin organization on variation in mutation rates in the genome. *Nature Reviews. Genetics* **16**, 213–223. 4, 84, 85, 149

M. H. Malamy, M. Fiandt & W. Szybalski (1972). Electron microscopy of polar insertions in the lac operon of Escherichia coli. *Molecular & general genetics : MGG* **119**, 207–222. 32

D. Marnetto, F. Mantica, I. Molineris, E. Grassi, I. Pesando & P. Provero (2018). Evolutionary Rewiring of Human Regulatory Networks by Waves of Genome Expansion. *American Journal of Human Genetics* **102**, 207–218. 40

L. Martinez-Gomez, F. Abascal, I. Jungreis, F. Pozo, M. Kellis, J. M. Mudge & M. L. Tress (2020). Few SINEs of life : Alu elements have little evidence for biological relevance despite elevated translation. *NAR genomics and bioinformatics* **2**, lqz023. 42

T. N. Mavrich, C. Jiang, I. P. Ioshikhes, X. Li, B. J. Venters, S. J. Zanton, L. P. Tomsho, J. Qi et al. (2008). Nucleosome organization in the Drosophila genome. *Nature* **453**, 358–362. 52

E. McArthur & J. A. Capra (2021). Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability. *American Journal of Human Genetics* **108**, 269–283. 32

B. McClintock (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences* **36**, 344–355. 32

J. D. McGhee & G. Felsenfeld (1980). Nucleosome structure. *Annual Review of Biochemistry* **49**, 1115–1156. 2

R. K. McGinty & S. Tan (2015). Nucleosome Structure and Function. *Chemical Reviews* **115**, 2255–2273. 2, 3, 4

K. J. McKernan, H. E. Peckham, G. L. Costa, S. F. McLaughlin, Y. Fu, E. F. Tsung, C. R. Clouser, C. Duncan et al. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Research* **19**, 1527–1541. Company : Cold Spring Harbor Laboratory Press Distributor : Cold Spring Harbor Laboratory Press Institution : Cold Spring Harbor Laboratory Press Label : Cold Spring Harbor Laboratory Press Publisher : Cold Spring Harbor Lab. 89

J. Mieczkowski, A. Cook, S. K. Bowman, B. Mueller, B. H. Alver, S. Kundu, A. M. Deaton, J. A. Urban et al. (2016). MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nature Communications* **7**. 91, 94, 99, 100, 104, 105, 106, 107, 111, 114, 141, 143, 150

V. Miele, C. Vaillant, Y. d'Aubenton Carafa, C. Thermes & T. Grange (2008). DNA physical properties determine nucleosome occupancy from yeast to fly. *Nucleic Acids Research* **36**, 3746–3756. 29, 50

P. Milani, G. Chevereau, C. Vaillant, B. Audit, Z. Haftek-Terreau, M. Marilley, P. Bouvet, F. Argoul et al. (2009). Nucleosome positioning by genomic excluding-energy barriers. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 22257–22262. 29, 52

G. A. Mitchell, D. Labuda, G. Fontaine, J. M. Saudubray, J. P. Bonnefont, S. Lyonnet, L. C. Brody, G. Steel et al. (1991). Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase : a role for Alu elements in human mutation. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 815–819. 46

L. D. Moore, T. Le & G. Fan (2013). DNA Methylation and Its Basic Function. *Neuropsychopharmacology* **38**, 23–38. Number : 1 Publisher : Nature Publishing Group. 78, 84, 150

S. Morganella, L. B. Alexandrov, D. Glodzik, X. Zou, H. Davies, J. Staaf, A. M. Sieuwerts, A. B. Brinkman et al. (2016). The topography of mutational processes in breast cancer genomes. *Nature Communications* **7**, 11383. Number : 1 Publisher : Nature Publishing Group. 76

N. Nassif, J. Penney, S. Pal, W. R. Engels & G. B. Gloor (1994). Efficient copying of nonhomologous sequences from ectopic sites via P-element-induced gap repair. *Molecular and Cellular Biology* **14**, 1613–1625. 36

M. Naville, S. Henriet, I. Warren, S. Sumic, M. Reeve, J.-N. Volff & D. Chourrout (2019). Massive Changes of Genome Size Driven by Expansions of Non-autonomous Transposable Elements. *Current biology : CB* . 41

J. Norris, D. Fan, C. Aleman, J. R. Marks, P. A. Futreal, R. W. Wiseman, J. D. Iglehart, P. L. Deininger et al. (1995). Identification of a new subclass of Alu DNA repeats which can function as estrogen receptor-dependent transcriptional enhancers. *The Journal of Biological Chemistry* **270**, 22777–22782. 46

H. O'Geen, L. Echipare & P. J. Farnham (2011). Using ChIP-seq technology to generate high-resolution profiles of histone modifications. *Methods in Molecular Biology (Clifton, N.J.)* **791**, 265–286. 152

S. Ohno (1972). So much "junk" DNA in our genome. *Brookhaven Symposia in Biology* **23**, 366–370. 32, 37

A. V. Onufriev & H. Schiessel (2019). The nucleosome : from structure to function through physics. *Current Opinion in Structural Biology* **56**, 119–130. 3, 4

M.-L. Pardue & P. G. DeBaryshe (2003). Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annual Review of Genetics* **37**, 485–511. 38

D. A. Pass, E. Sornay, A. Marchbank, M. R. Crawford, K. Paszkiewicz, N. A. Kent & J. A. H. Murray (2017). Genome-wide chromatin mapping with size resolution reveals a dynamic sub-nucleosomal landscape in Arabidopsis. *PLoS Genetics* **13**. 99, 100

J. Piriyapongsa, M. T. Rutledge, S. Patel, M. Borodovsky & I. K. Jordan (2007). Evaluating the protein coding potential of exonized transposable element sequences. *Biology Direct* **2**, 31. 39

P. Polak & E. Domany (2006). Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC genomics* **7**, 133. 46

E. Y. Popova, S. A. Grigoryev, Y. Fan, A. I. Skoultchi, S. S. Zhang & C. J. Barnstable (2013). Developmentally Regulated Linker Histone H1c Promotes Heterochromatin Condensation and Mediates Structural Integrity of Rod Photoreceptors in Mouse Retina *. *Journal of Biological Chemistry* **288**, 17895–17907. Publisher : Elsevier. 100

L. A. Pray (2008). Eukaryotic Genome Complexity | Learn Science at Scitable. *Nature Education* **1**. Cg_cat : Eukaryotic Genome Complexity Cg_level : MED Cg_topic : Eukaryotic Genome Complexity. 116

A. L. Price, E. Eskin & P. A. Pevzner (2004). Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome research, Genome Research* **14, 14**, 2245, 2245–2252. 42, 43

L. Quadrana, A. Bortolini Silveira, G. F. Mayhew, C. LeBlanc, R. A. Martienssen, J. A. Jeddeloh & V. Colot (2016). The Arabidopsis thaliana mobilome and its impact at the species level. *eLife* **5**, e15716. 36

Y. Quentin (1992). Fusion of a free left Alu monomer and a free right Alu monomer at the origin of the Alu family in the primate genomes. *Nucleic Acids Research* **20**, 487–493. 42

Y. Quentin (1994). A master sequence related to a free left Alu monomer (FLAM) at the origin of the B1 family in rodent genomes. *Nucleic Acids Research* **22**, 2222–2227. 153

O. Raymond, J. Gouzy, J. Just, H. Badouin, M. Verdenaud, A. Lemainque, P. Vergne, S. Moja et al. (2018). The Rosa genome provides new insights into the domestication of modern roses. *Nature Genetics* **50**, 772–777. II

R. Rebollo, K. Miceli-Royer, Y. Zhang, S. Farivar, L. Gagnier & D. L. Mager (2012a). Epigenetic interplay between mouse endogenous retroviruses and host genes. *Genome Biology* **13**, R89. 40

R. Rebollo, M. T. Romanish & D. L. Mager (2012b). Transposable elements : an abundant and natural source of regulatory sequences for host genes. *Annual Review of Genetics* **46**, 21–42. 50

P. J. Robinson & D. Rhodes (2006). Structure of the '30nm' chromatin fibre : A key role for the linker histone. *Current Opinion in Structural Biology* **16**, 336–343. 3

A. M. Roy-Engel, A.-H. Salem, O. O. Oyeniran, L. Deininger, D. J. Hedges, G. E. Kilroy, M. A. Batzer & P. L. Deininger (2002). Active Alu element "A-tails" : size does matter. *Genome Research* **12**, 1333–1344. 43, 133

F. Sabot & A. H. Schulman (2006). Parasitism and the retrotransposon life cycle in plants : a hitchhiker's guide to the genome. *Heredity* **97**, 381–388. 34

P. SanMiguel, A. Tikhonov, Y. K. Jin, N. Motchoulskaia, D. Zakharov, A. Melake-Berhan, P. S. Springer, K. J. Edwards et al. (1996). Nested retrotransposons in the intergenic regions of the maize genome. *Science (New York, N.Y.)* **274**, 765–768. 33, 34

S. C. Satchwell, H. R. Drew & A. A. Travers (1986). Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology* **191**, 659–675. 127, 133

D. E. Schones, K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, G. Wei & K. Zhao (2008). Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell* **132**, 887–898. 88, 93

L. Schrader & J. Schmitz (2018). The impact of transposable elements in adaptive evolution. *Molecular Ecology* . 38, 41

O. Seberg & G. Petersen (2009). A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nature Reviews. Genetics* **10**, 276. 33

E. Segal & J. Widom (2009a). Poly(dA :dT) tracts : major determinants of nucleosome organization. *Current Opinion in Structural Biology* **19**, 65–71. 48, 127, 133, 152

E. Segal & J. Widom (2009b). What controls nucleosome positions? *Trends in genetics : TIG* **25**, 335–343. 5, 31, 50, 127, 148, 152

N. Sela, B. Mersch, N. Gal-Mark, G. Lev-Maor, A. Hotz-Wagenblatt & G. Ast (2007). Comparative analysis of transposed element insertion within human and mouse genomes reveals Alu's unique role in shaping the human transcriptome. *Genome Biology* **8**, R127. 39

S. K. Sen, K. Han, J. Wang, J. Lee, H. Wang, P. A. Callinan, M. Dyer, R. Cordaux et al. (2006). Human genomic deletions mediated by recombination between Alu elements. *American Journal of Human Genetics* **79**, 41–53. 47

C. Sessegolo, N. Burlet & A. Haudry (2016). Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. *Biology Letters* **12**, 20160407. 41

E. Smith & A. Shilatifard (2014). Enhancer biology and enhanceropathies. *Nature Structural & Molecular Biology* **21**, 210–219. Number : 3 Publisher : Nature Publishing Group. 32

D. Srikanta, S. K. Sen, C. T. Huang, E. M. Conlin, R. M. Rhodes & M. A. Batzer (2009). An alternative pathway for Alu retrotransposition suggests a role in DNA double-strand break repair. *Genomics* **93**, 205–212. 47

L. Stenz (2021). The L1-dependant and Pol III transcribed Alu retrotransposon, from its discovery to innate immunity. *Molecular Biology Reports* **48**, 2775–2789. 41, 42, 44

K. Struhl (1985). Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **82**, 8419–8423. 127, 133

K. Struhl & E. Segal (2013). Determinants of nucleosome positioning. *Nature Structural & Molecular Biology* **20**, 267–273. 5, 50, 127

T. Sultana, A. Zamborlini, G. Cristofari & P. Lesage (2017). Integration site selection by retroviruses and transposable elements in eukaryotes. *Nature Reviews. Genetics* **18**, 292–308. 36, 37, 149

T. Sultana, D. van Essen, O. Siol, M. Bailly-Bechet, C. Philippe, A. Zine El Aabidine, L. Pioger, P. Nigumann et al. (2019). The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. *Molecular Cell* . 36

V. Sundaram & J. Wysocka (2020). Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **375**, 20190347. 40

M. V. Suntsova & A. A. Buzdin (2020). Differences between human and chimpanzee genomes and their implications in gene expression, protein functions and biochemical properties of the two species. *BMC genomics* **21**, 535. 65

E. Szenker, D. Ray-Gallet & G. Almouzni (2011). The double face of the histone variant H3.3. *Cell Research* **21**, 421–434. Number : 3 Publisher : Nature Publishing Group. 4, 111, 151

Y. Tanaka, R. Yamashita, Y. Suzuki & K. Nakai (2010). Effects of Alu elements on global nucleosome positioning in the human genome. *BMC Genomics* **11**, 309. 48, 124, 127

W. Tang & P. Liang (2019). Comparative Genomics Analysis Reveals High Levels of Differential Retrotransposition among Primates from the Hominidae and the Cercopithecidae Families. *Genome Biology and Evolution* **11**, 3309–3325. 128, 129, 134

W. Tang & P. Liang (2020). Alu master copies serve as the drivers of differential SINE transposition in recent primate genomes. *Analytical Biochemistry* **606**, 113825. 128, 129, 134

B. C. Taylor & N. L. Young (2021). Combinations of histone post-translational modifications. *The Biochemical Journal* **478**, 511–532. 4

V. B. Teif (2016). Nucleosome positioning : resources and tools online. *Briefings in Bioinformatics* **17**, 745–757. 88, 99

J. J. Tena & J. M. Santos-Pereira (2021). Topologically Associating Domains and Regulatory Landscapes in Development, Evolution and Disease. *Frontiers in Cell and Developmental Biology* **9**, 702787. 32

M. Y. Tolstorukov, P. V. Kharchenko & P. J. Park (2010). Analysis of primary structure of chromatin with next-generation sequencing. *Epigenomics* **2**, 187–197. 88

M. Tompitak, C. Vaillant & H. Schiessel (2017). Genomes of Multicellular Organisms Have Evolved to Attract Nucleosomes to Promoter Regions. *Biophysical Journal* **112**, 505–511. 54, 55

A. Tsankov, Y. Yanagisawa, N. Rhind, A. Regev & O. J. Rando (2011). Evolutionary divergence of intrinsic and trans-regulated nucleosome positioning sequences reveals plastic rules for chromatin organization. *Genome Research* **21**, 1851–1862. 50

A. M. Tsankov, D. A. Thompson, A. Socha, A. Regev & O. J. Rando (2010). The role of nucleosome positioning in the evolution of gene regulation. *PLoS biology* **8**, e1000414. 4, 54, 76

J. Ule (2013). Alu elements : at the crossroads between disease and evolution. *Biochemical Society Transactions* **41**, 1532–1535. 46

C. Vaillant, B. Audit & A. Arneodo (2007). Experiments confirm the influence of genome long-range correlations on nucleosome positioning. *Physical Review Letters* **99**, 218103. 88

C. Vaillant, L. Palmeira, G. Chevereau, B. Audit, Y. d'Aubenton Carafa, C. Thermes & A. Arneodo (2010). A novel strategy of transcription regulation by intragenic nucleosome ordering. *Genome Research* **20**, 59–67. 29, 50

A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J. A. Malek et al. (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Research* **18**, 1051–1063. 89

A. Valouev, S. M. Johnson, S. D. Boyd, C. L. Smith, A. Z. Fire & A. Sidow (2011). Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516–520. 65, 88, 91, 93, 94, 97, 100, 127, 141

G. Vansant & W. F. Reynolds (1995). The consensus sequence of a major Alu subfamily contains a functional retinoic acid response element. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 8229–8233. 46

A. E. Van't Hof, P. Campagne, D. J. Rigden, C. J. Yung, J. Lingley, M. A. Quail, N. Hall, A. C. Darby et al. (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nature* **534**, 102–105. 41

C. Vitte & O. Panaud (2005). LTR retrotransposons and flowering plant genome size : emergence of the increase/decrease model. *Cytogenetic and Genome Research* **110**, 91–107. 41

J. Vogt, K. Bengesser, K. B. Claes, K. Wimmer, V.-F. Mautner, R. van Minkelen, E. Legius, H. Brems et al. (2014). SVA retrotransposon insertion-associated deletion represents a novel mutational mechanism underlying large genomic copy number changes with non-recurrent breakpoints. *Genome Biology* **15**, R80. 38

B. J. Wagstaff, D. J. Hedges, R. S. Derbes, R. Campos Sanchez, F. Chiaromonte, K. D. Makova & A. M. Roy-Engel (2012). Rescuing Alu : recovery of new inserts shows LINE-1 preserves Alu activity through A-tail expansion. *PLoS genetics* **8**, e1002842. 36

T. Wang, J. Zeng, C. B. Lowe, R. G. Sellers, S. R. Salama, M. Yang, S. M. Burgess, R. K. Brachmann et al. (2007). Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 18613–18618. 40

A. Weiner, A. Hughes, M. Yassour, O. J. Rando & N. Friedman (2010). High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Research* **20**, 90–100. 104

T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy et al. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* **8**, 973. 33, 34, 35

J. Wu, M. McKeague & S. J. Sturla (2018). Nucleotide-Resolution Genome-Wide Mapping of Oxidative DNA Damage by Click-Code-Seq. *Journal of the American Chemical Society* **140**, 9783–9787. 4, 76

S.-F. Wu, H. Zhang & B. R. Cairns (2011). Genes for embryo development are packaged in blocks of multivalent chromatin in zebrafish sperm. *Genome Research* **21**, 578–589. 100

Y. Xi, J. Yao, R. Chen, W. Li & X. He (2011). Nucleosome fragility reveals novel functional states of chromatin and poises genes for activation. *Genome Research* **21**, 718–724. 104, 105, 106, 108

M. Xie, C. Hong, B. Zhang, R. F. Lowdon, X. Xing, D. Li, X. Zhou, H. J. Lee et al. (2013). DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nature Genetics* **45**, 836–841. 40

W.-S. Yong, F.-M. Hsu & P.-Y. Chen (2016). Profiling genome-wide DNA methylation. *Epigenetics & Chromatin* **9**, 26. 86

G.-C. Yuan, Y.-J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler & O. J. Rando (2005). Genome-scale identification of nucleosome positions in S. cerevisiae. *Science (New York, N.Y.)* **309**, 626–630. 88

T. Zemojtel, S. M. Kielbasa, P. F. Arndt, H.-R. Chung & M. Vingron (2009). Methylation and deamination of CpGs generate p53-binding sites on a genomic scale. *Trends in genetics : TIG* **25**, 63–66. 47

Y. Zhang, N. L. Vastenhouw, J. Feng, K. Fu, C. Wang, Y. Ge, A. Pauli, P. van Hummelen et al. (2014). Canonical nucleosome organization at promoters forms during genome activation. *Genome Research* **24**, 260–266. 99, 100

Z. Zhang & B. F. Pugh (2011). High resolution genome-wide mapping of the primary structure of chromatin. *Cell* **144**, 175–186. 104, 106