SECTION II

ITEM 1.1

Structure de l'ADN et organisation

STRUCTURE DE L'ADN

STRUCTURE PRIMAIRE

L'ADN (acide désoxyribonucléique) (figure 1.1.1) est un polymère formé d'un enchaînement d'unités de base, les dNMP (désoxyribonucléosides 5' monophosphate). Chaque dNMP est composé d'une base azotée, d'un sucre (le désoxyribose) et d'un groupement phosphate. Les bases azotées sont l'adénine (A), la guanine (G), la cytosine (C) et la thymine (T). L'enchaînement nucléotidique se fait via une liaison phosphodiester entre le groupement hydroxyl en 3' d'un désoxyribose et le groupement 5'-phosphate du dNMP adjacent. Cet enchaînement conduit à une polarité dans la chaîne polynucléotidique avec une extrémité 5'-phosphate libre et une extrémité 3'-OH libre.

STRUCTURE SECONDAIRE

La structure secondaire de l'ADN a été révélée en 1953 grâce aux travaux de Watson, Crick et Franklin. L'ADN cellulaire est bicaténaire (double brin [db]), constitué par l'association de deux chaînes polynucléotidiques maintenues par des liaisons hydrogène (H) établies entre les bases azotées (2 liaisons H entre les bases A et T et 3 liaisons H entre les bases C et G). Dans le duplex d'ADN, les deux chaînes polynucléotidiques sont antiparallèles (figure 1.1.1). La taille d'un fragment d'ADN est généralement donnée en nombre de pb (paires de bases).

Les deux brins d'ADN s'enroulent l'un autour de l'autre, formant classiquement une hélice droite dans laquelle les plateaux de pb sont orientés vers l'intérieur de l'hélice, quasi perpendiculaires à l'axe de l'hélice (hélice de type B). Des enroulements hélicoïdaux alternatifs existent dans la cellule, notamment l'hélice Z avec un enroulement hélicoïdal gauche.

STRUCTURE TERTIAIRE

Le génome humain est constitué de l'ADN nucléaire et de l'ADN mitochondrial.

STRUCTURE TERTIAIRE DE L'ADN NUCLÉAIRE

L'ADN nucléaire est réparti sur 22 paires d'autosomes et 2 gonosomes (X et Y). Sa taille est de $3{,}05 \times 10^9$ pb par génome haploïde. Chaque chromosome correspond à une molécule d'ADN db linéaire. L'ADN est associé à des protéines, notamment aux histones, pour former la chromatine. Un premier niveau de structuration correspond à un enroulement gauche d'un segment d'ADN db de 146 pb autour d'un octamère d'histones (2 histones H2A, 2 histones H2B, 2 histones H3 et 2 histones H4) pour former le nucléosome, l'unité structurale de base de la chromatine. L'association de l'histone H1 au nucléosome conduit au chromatosome, impliquant l'enroulement d'un segment d'ADN de 166 pb. Des organisations structurales d'ordres supérieurs sont obtenues via des surenroulements et des repliements (figure 1.1.1). Celles-ci sont nécessaires à la compaction et à l'organisation de la chromatine, mais également importantes dans la fonction de l'ADN, notamment dans le contrôle de l'expression des gènes (voir item 2.3, « Promoteur, plaque tournante de la régulation transcriptionnelle »).

Dans le noyau, la chromatine présente des degrés variables de condensation. L'hétérochromatine correspond à des régions très denses avec une activité transcriptionnelle faible, voire nulle. L'euchromatine, beaucoup plus décondensée, correspond à la chromatine active dans laquelle se trouvent les gènes et les séquences impliquées dans leur expression et leur régulation. Le degré de condensation de l'ADN, et donc l'accessibilité de la chromatine, peut être modulé par des protéines de remodelage de la chromatine et des modifications épigénétiques affectant les bases (méthylation, notamment sur le carbone en position 5 de la cytosine [m^5C] dans les îlots CpG) ou les histones (voir item 2.3, « Promoteur, plaque tournante de la régulation transcriptionnelle »).

STRUCTURE TERTIAIRE DE L'ADN MITOCHONDRIAL (ADNmt)

▶ L'ADNmt est constitué d'un db circulaire. Sa taille est de 16 569 pb. Il est hérité de la mère. Il porte dans sa séquence l'information pour 37 gènes :

- 13 codant des sous-unités impliquées dans la phosphorylation oxydative ;
- 2 codant des ARNr (acides ribonucléiques ribosomiques) mitochondriaux ;
- 22 codant des ARNt (ARN de transfert).

▶ L'ADNmt est également associé à des protéines et notamment à la protéine TFAM (*Mitochondrial Transcription Factor A*) pour former une structure nucléoprotéique appelée nucléoïde.

Le nombre de copies de génomes dans une mitochondrie peut varier de plusieurs centaines à plusieurs milliers et dépend de l'activité cellulaire et des besoins énergétiques de la cellule.

TOPOLOGIE DE L'ADN ET TOPOISOMÉRASES

L'enroulement de l'ADN cellulaire en une longue double hélice et l'organisation de la chromatine avec la formation de grandes boucles génèrent des contraintes structurales qui, dans différentes situations, conduisent à une modification de l'enroulement de l'ADN et à l'apparition d'isomères topologiques (topoisomères) pouvant potentiellement mener à de crises topologiques et la mort cellulaire en cas de non-résolution. Ainsi, la réplication (voir item 1.3.1) ou la transcription de l'ADN (voir item 2) impliquent la séparation locale des deux brins d'ADN sur des distances plus ou moins longues. Cela génère à l'avant un excès d'enroulement de la double hélice (surenroulement positif) qui peut rapidement bloquer la fourche de réplication ou la bulle de transcription. Cette situation est évitée grâce à l'action d'enzymes, les topoisomérases (de type I ou II) qui gèrent la topologie de l'ADN en évitant son surenroulement et la formation de nœuds. Ces enzymes sont dotées d'une activité endonucléasique permet-

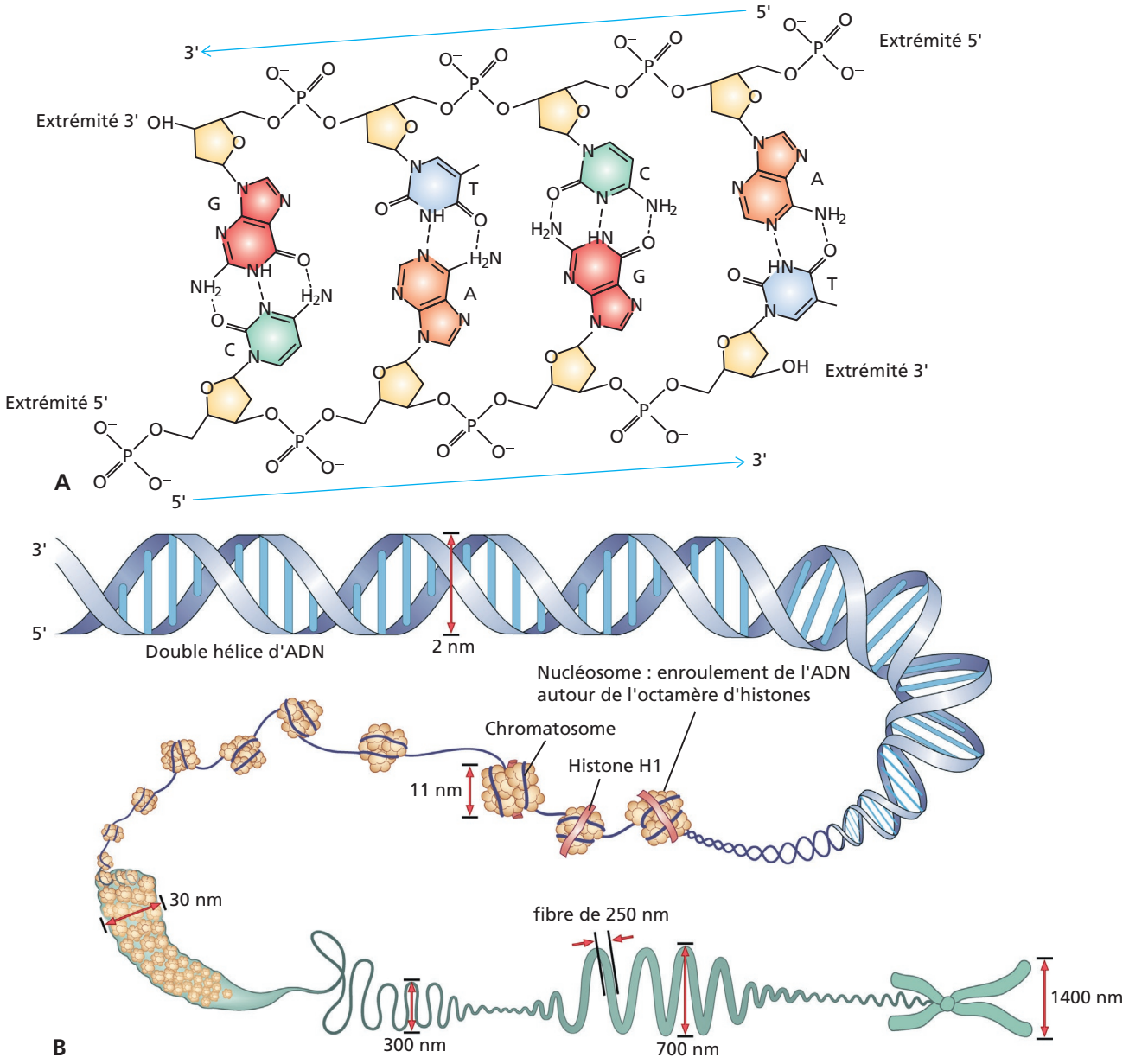

Figure 1.1.1 Structures primaire, secondaire et tertiaire de l'ADN
(A) Double-hélice d'ADN. Segment d'ADN de 4 pb montrant l'enchaînement des 4 dNMP, ainsi que les liaisons H entre les bases complémentaires. Chaque brin présente une polarité 5' et 3' et les deux brins sont antiparallèles ; (B) Organisation de l'ADN nucléaire. Formation de la chromatine nucléaire par association avec des protéines, notamment les histones. Formation du nucléosome par enroulement gauche de l'hélice d'ADN autour de l'octamère d'histones. La fixation de l'histone H1 conduit au chromatosome. Cette unité de base s'organise ensuite avec formation de fibres et de boucles. Le chromosome présenté correspond à l'étape ultime de condensation de la fibre de chromatine retrouvée lors de la mitose.
ADN : acide désoxyribonucléique ; dNMP : désoxyribonucléoside 5' monophosphate.
Source : Hormones, Fourth Edition, 978-0-323-90262-5, Gerald Litwack, © Elsevier Inc, 2022.

tant la coupure de la liaison 3'–5' phosphodiester et d'une activité ligase pour reformer cette liaison phosphodiester. Des médicaments bloquant l'action des topoisomérases I (irinotecan, topotecan…) ou II (agents intercalants tels les dérivés d'anthracyclines [daunorubicine, doxorubicine, épirubicine…], étoposide, actinomycine D, bléomycine, mitoxanthone…) sont utilisés comme agents cytotoxiques dans certaines chimiothérapies anticancéreuses.

SÉQUENÇAGE ET ANNOTATION DU GÉNOME HUMAIN

SÉQUENÇAGE ET ASSEMBLAGE DES SÉQUENCES

Le séquençage du génome humain fut officiellement lancé en 1990 dans le cadre d'un vaste programme scientifique international public : le HGP (*Human Genome Project*) avec

un budget de 3 milliards de dollars et sur une durée de quinze ans. Parallèlement, un projet privé (Celera) mené par Craig Venter vit le jour en 1998. Une première version de la séquence de l'euchromatine du génome humain, obtenue en utilisant la méthode de séquençage de Sanger, fut publiée en 2001 dans la revue *Nature* par le consortium public et dans *Science* pour le projet privé. La version publique a été progressivement complétée par le Genome Reference Consortium (qui a pris le relais du HGP) et a publié en 2022 une mise à jour de la séquence du génome humain de référence (assemblage GRCh38.p14) (www.ncbi.nlm.nih.gov/grc). Parallèlement, le consortium Telomere to Telomere (T2T) (www.ncbi.nlm.nih.gov/assembly/GCA_009914755.4) a publié, en 2022, une séquence alternative complète du génome humain (assemblage T2T-CHM13). La complétion de la séquence du génome humain a bénéficié de nouvelles techniques de séquençage NGS (*Next Generation Sequencing*) basées sur du séquençage massivement parallèle et la possibilité de séquencer avec beaucoup de précision de longs fragments d'ADN, ce qui facilite l'assemblage des séquences. Aujourd'hui, grâce à ces techniques, le séquençage d'un génome humain peut être fait en quelques jours, pour 1 000 dollars. Les méthodes de séquençage de Sanger et NGS seront expliquées et détaillées dans l'item 4.2, « Introduction aux méthodes de séquençage d'ADN » à « Technologies de séquençage haut débit de troisième génération (TGS) dites "SMRT" (*Single Molecule Real Time*) développées par Pacific Biosciences et "ONT" (Oxford Nanopore Technology) ».

ANNOTATION DU GÉNOME HUMAIN. COMPOSANTS MAJEURS DU GÉNOME

Une moitié environ du génome humain est représentée par des séquences uniques correspondant à la plupart des gènes, aux séquences associées (séquences régulatrices), à des séquences intergéniques et aux pseudogènes (Tableau 11.1). L'autre moitié est constituée de séquences répétées (figure 1.1.2).

GÈNES ET PSEUDOGÈNES

Les gènes correspondent à la partie de l'ADN pouvant être transcrite en ARN. Les gènes peuvent coder différents types d'ARN (codants ou non codants) ayant des fonctions diverses au sein de la cellule (voir item 2.1, « Généralités et acteurs clés » et « Différentes étapes de la transcription »). Le tableau 1.1.1 résume le nombre de gènes pour chacun de ces types dans le génome humain. Notons qu'un gène donné peut conduire à plusieurs transcrits de séquence différente. Cela est notamment le cas pour les gènes codant des protéines pour lesquels on compte en moyenne 4,3 transcrits par gène.

MORCELLEMENT DES GÈNES

Beaucoup de gènes humains sont morcelés, c'est-à-dire constitués d'exons et d'introns. Les exons correspondent aux séquences qui se retrouvent dans les ARN matures présents dans le cytoplasme cellulaire, comme l'ARNm (ARN messager) mais aussi les ARN non codants (voir item 2). Les introns sont les séquences placées entre deux exons. Ils sont (généralement) excisés après la transcription via une réaction d'excision-épissage (ou *splicing*) (voir item 2.1,

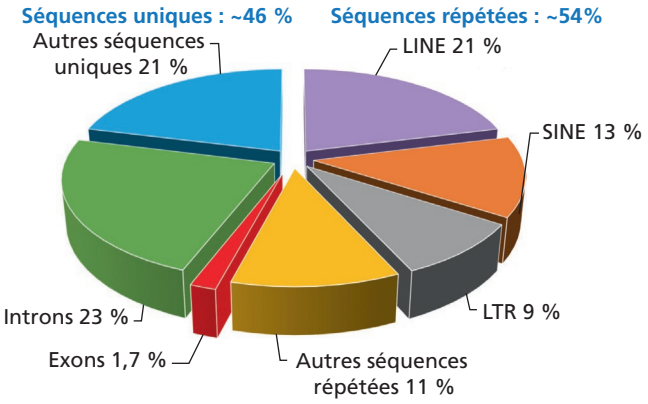

Figure 1.1.2 Composition du génome humain
Le génome humain est constitué de ~46 % de séquences uniques et de ~54 % de séquences répétées. La distribution des différents éléments de chaque catégorie de séquences est donnée en pourcentage.
Sources : À partir des données de Nuke et al. The complete sequence of a human genome. Science. 2022;376(6588); NCBI Homo sapiens Annotation Release GCF_000001405.40-RS_2023_03, disponible sur le site www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/GCF_000001405.40-RS_2023_03.

Tableau 1.1.1. Annotation du génome humain (assemblage T2T-CHM13v1.1) – Gènes et pseudogènes

Catégorie de gènes	Types de gènes	Nombre de gènes	Nombre de transcrits
Gènes codant des protéines	Gènes codant des protéines	**19 969**	**86 245**
Gènes ne codant pas des protéines	ARN longs non codants (LncARN)	17 482	
	Micro-ARN (miRNA)	2 045	
	ARN ribosomiques (ARNr)	1 007	
	ARN *small nuclear* (ARNsn)	1 886	
	Autres ARN non codants	4 566	
	Total	**27 002**	**57 196**
Pseudogènes	Pseudogènes dupliqués	3 667	
	Rétropseudogènes	11 322	
	Pseudogènes unitaires	236	
	Autres	574	
	Total	**15 799**	**15 997**

Longueur totale de la séquence du génome humain : 3 054 815 472 pb.
ARN : acide ribonucléique ; ARNr : ARN ribosomique ; LINE : *Long Interspersed Nuclear Elements* ; LTR : *Long Terminal Repeat* ; pb : paire de bases ; SINE : *Short Interspersed Nuclear Elements*.
Source : À partir des données de Nuke et al. The complete sequence of a human genome. Science. 2022;376(6588).

« Maturation du pré-ARNm et transport du noyau au cytoplasme »). Le nombre d'exons par transcrit est en moyenne de 11,89 et varie entre un (gène dit « monoexonique ») et 363 (gènes dits « pluriexoniques ») (www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/GCF_009914755.1-RS_2023_03/). L'ensemble des exons représente ~1,7 % du génome humain. Les exons codant des protéines représentent ~1,2 % du génome total.

PSEUDOGÈNES

Certaines séquences d'ADN ressemblent à un gène de la même espèce ou d'une autre espèce, mais ont perdu leur fonction d'origine suite à l'accumulation de mutations ou l'apparition de formes tronquées. La plupart de ces séquences présentent un défaut de transcription ou de traduction. La dernière annotation du génome (assemblage T2T CHM13v1.1) recense 15 799 pseudogènes parmi lesquels 1 633 (11,5 %) sont transcrits et 3 sont traduits.

On distingue :

▶ les pseudogènes dupliqués résultant d'une duplication d'un gène, suivie de mutations. Ils conservent leur structure morcelée, avec exons et introns, et se trouvent localisés à côté du gène d'origine ;

▶ les rétropseudogènes émanant d'une rétrotransposition d'un ARN cellulaire (voir « Rétrotransposons, éléments transposables de type I » ci-dessous). Ces pseudogènes sont dispersés dans le génome, n'ont pas de promoteur et ne possèdent (généralement) pas d'introns ;

▶ les pseudogènes unitaires résultant de mutations dans un gène, conduisant à la perte de sa fonction. Citons le gène *GULO* codant la L-gulonolactone oxydase impliquée dans la voie de biosynthèse de la vitamine C. Ce gène, fonctionnel chez la souris, ne l'est plus chez les primates (ψ*GULO*) qui ont donc besoin d'un apport en vitamine C.

GÈNES REGROUPÉS EN « CLUSTERS »

Certains gènes de structures et de fonctions proches sont regroupés en clusters. C'est le cas des gènes codant des protéines de la famille de l'α-globine, β-globine, récepteurs olfactifs, *Hox*, h*GH* (hormone de croissance humaine), ou *Pcdh* (protocadhérine). Ces gènes paralogues (c'est-à-dire des gènes similaires dans la même espèce) résultent de duplications en tandem (en file indienne) de segments d'ADN.

Des clusters sont également retrouvés pour les gènes codant les ARNr 45S et 5S. Ainsi, on trouve :

▶ ~100-250 répétitions (par génome haploïde) d'une séquence de ~45 000 pb codant le gène de l'ARNr 45S (S : Svedberg) qui porte les séquences des ARNr 18S, 5,8S et 28S, dispersées sur 5 chromosomes (5 clusters de ~7 à 13 répétitions sur les chromosomes acrocentriques 13, 14, 15, 21 et 22) ;

▶ plusieurs centaines de gènes pour l'ARNr 5S répartis en 3 clusters sur le chromosome 1.

SÉQUENCES RÉPÉTÉES DE L'ADN GÉNOMIQUE HUMAIN ET ÉLÉMENTS TRANSPOSABLES

ADN RÉPÉTÉ (TABLEAU 1.1.2 ET FIGURE 1.1.2)

Les séquences répétées (~54 % du génome humain) peuvent être codantes ou non codantes, de tailles variables, localisées ou disséminées dans le génome.

Tableau 1.1.2. Annotation du génome humain (assemblage T2T-CHM13v1.1) – Séquences répétées

Séquences répétées	Longueur totale (mégabases)	%
Total	**1 648**	**54**
LINE	631	20,7
SINE	390	12,8
Avec LTR	270	8,8
ARNr	1,71	0,05
Autres	348	11,4

Longueur totale de la séquence du génome humain : 3 054 815 472 pb.

ARN : acide ribonucléique ; ARNr : ARN ribosomique ; LINE : *Long Interspersed Nuclear Elements* ; LTR : *Long Terminal Repeat* ; pb : paire de bases ; SINE : *Short Interspersed Nuclear Elements*.

Source : À partir des données de Nuke et al. The complete sequence of a human genome. Science. 2022 ;376(6588).

SÉQUENCES RÉPÉTÉES NON CODANTES MAJORITAIREMENT

Parmi les séquences fortement répétées non codantes et localisées en tandem dans le génome, citons :

▶ les séquences télomériques, très courtes (-TTAGGG-) répétées entre 250 et 3 000 fois à l'extrémité des chromosomes (voir item 1.3.1) ;

▶ les séquences centromériques des chromosomes, courtes de 171 pb répétées sur 300 kilobases (kilobase [Kb] = 10^3 pb) à plusieurs mégabases (mégabase [Mb] = 10^6 pb).

Pour les séquences répétées disséminées, citons :

▶ les microsatellites STR (*Short Tandem Repeats*) : répétitions en tandem d'une très courte séquence de nucléotides (nt) (2-6 nt) ;

▶ les minisatellites VNTR (*Variable Number Tandem Repeats*) : répétitions en tandem d'une séquence de 7-100 nt sur plusieurs kilobases ;

▶ les SINE (*Short Interspersed Nuclear Elements*) de 80 à 400 pb ;

▶ les LINE (*Long Interspersed Nuclear Elements*) avec des tailles de plusieurs kilobases (7 Kb en moyenne).

SÉQUENCES RÉPÉTÉES CODANTES

La majorité des gènes cellulaires n'est pas répétée. On peut néanmoins citer quelques exceptions :

▶ gènes codant les ARNr (voir « Gènes regroupés en "clusters" » ci-dessus) ;

▶ gènes codant les histones : 61 gènes répartis sur 11 clusters distribués sur 7 chromosomes.

ÉLÉMENTS TRANSPOSABLES

Une partie de l'ADN répété est le fruit d'une activité de transposition liée aux éléments génétiques mobiles. Ces éléments sont dispersés dans le génome.

RÉTROTRANSPOSONS, ÉLÉMENTS TRANSPOSABLES DE TYPE I

Ce sont des segments d'ADN transposés via un intermédiaire ARN, c'est-à-dire que le segment d'ADN sera transcrit par une ARN polymérase puis rétrotransposé dans l'ADN à un endroit différent de sa position d'origine. Ce mécanisme permet une

amplification importante du segment d'ADN de départ. On distingue deux types de rétrotransposons selon l'absence ou la présence de LTR (*Long Terminal Repeat*). Les LTR sont des séquences présentes aux extrémités des génomes rétroviraux.

▶ Rétrotransposons avec LTR (~9 % du génome). Ils sont considérés comme les vestiges de l'infection de cellules germinales par des rétrovirus, comme les HERV (*Human Endogeneous Retrovirus*) H, W, K et L. Ces HERV sont des formes généralement tronquées des rétrovirus, non actives. Néanmoins certaines séquences de HERV jouent un rôle dans le fonctionnement cellulaire (par exemple celles codant les syncitines 1 et 2) ou dans le contrôle de l'expression des gènes et le remodelage du génome. D'autres semblent être à l'origine de pathologies suite à une réactivation de la production de protéines virales ou d'acides nucléiques viraux induisant une réaction inflammatoire ou ayant une toxicité pour les cellules (HERV-W dans la sclérose en plaques ou HERV-K dans la sclérose amyotrophique latérale).

▶ Rétrotransposons sans LTR (~33 % du génome).

- Les SINE (~13 % du génome nucléaire) sont des séquences de 80 à 400 pb, transcrites, mais ne codant pas de protéines. On dénombre environ 1 500 000 SINE par génome haploïde, dont 1 000 000 de séquences *Alu* (taille 300 pb).
- Les LINE (~21 % du génome nucléaire humain) sont des séquences entre 5 et 6 Kb, avec de 5 000 à plus de 500 000 copies. La plupart des éléments LINE-1 sont inactifs, seuls ~ 80-100 sont actifs/génome. Il existe 3 familles de LINE. L'une d'entre elles, LINE-1, contient deux ORF (*Open Reading Frames* ou cadres ouverts de lecture) codant les activités nécessaires à la rétrotransposition et notamment une activité endonucléasique permettant de couper l'ADN cible et une activité de transcriptase inverse permettant de rétrotranscrire l'ARN issu de la transcription de LINE-1 dans l'ADN aux sites préalablement coupés par l'endonucléase (figure 1.1.3).

La machinerie LINE-1 est active en cis, c'est-à-dire qu'elle rétrotranscrit son propre ARN, mais également en trans, c'est-à-dire qu'elle peut rétrotranscrire d'autres ARN, notamment les ARN issus de la transcription des SINE. Elle peut également rétrotranscrire d'autres ARN cellulaires pour former des rétropseudogènes.

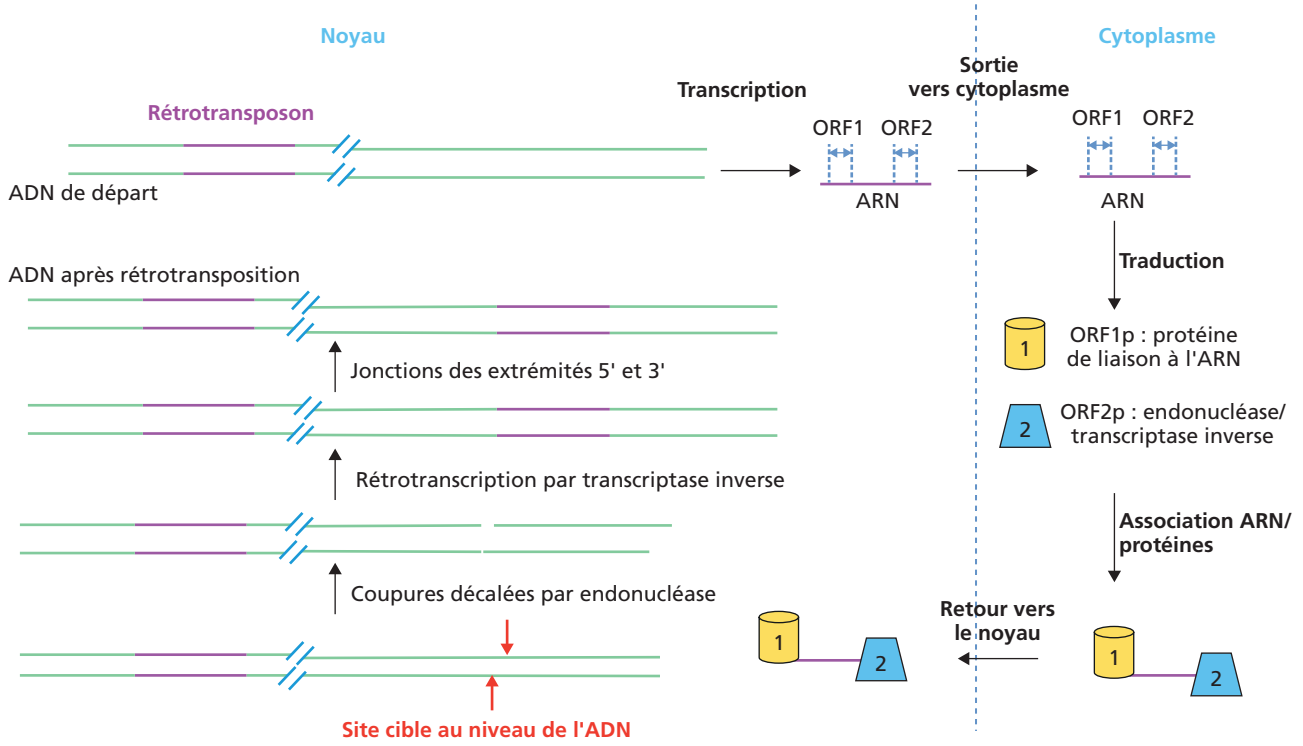

Figure 1.1.3 Mécanisme de la rétrotransposition d'un élément LINE-1
Le segment d'ADN correspondant à l'élément LINE-1 est transcrit. L'ARN correspondant polyadénylé passe dans le cytoplasme. Les 2 ORF sont traduits. L'ORF1 code une protéine (ORF1p) avec un motif de reconnaissance de l'ARN. L'ORF2 code une protéine (ORF2p) avec une activité nucléasique et une activité transcriptase inverse. Les deux protéines se lient à l'ARN pour former une ribonucléoprotéine qui retourne vers le noyau. L'activité nucléasique coupe l'ADN cible. L'ARN du rétrotransposon est rétrotranscrit par un mécanisme dit « TPRT ». Cette rétrotranscription est suivie de la synthèse du second brin d'ADN, puis de la mise en place des liaisons phosphodiesters par une ligase. Les segments d'ADN de part et d'autre des traits verticaux indiquent que le site de la rétrotransposition peut être sur le même chromosome ou sur des chromosomes différents.
ADN : acide désoxyribonucléique ; ARN : acide ribonucléique ; LINE : *Long Interspersed Nuclear Element* ; ORF : *Open Reading Frame* ; TPRT : *Target Primed Reverse Transcription*.
Source : D'après Pizarro JG, Cristofari G. Post-Transcriptional Control of LINE-1 Retrotransposition by Cellular Host Factors in Somatic Cells. Front Cell Dev Biol. 2016;4,article 14.

On estime que 30 % du génome humain dérive de l'activité de LINE-1.
La rétrotransposition peut être silencieuse, mais elle peut également induire une mutagenèse insertionnelle, une réaction d'excision/épissage aberrante, une perturbation de la transcription de l'élément qui héberge le rétrotransposon ou une modification de la régulation épigénétique.
Une centaine de rétrotranspositions liées à LINE-1 sont connues pour avoir donné lieu à des pathologies dont des cancers ou des pathologies monogéniques (hémophilies, thalassémies, maladie de Fabry…).

ÉLÉMENTS TRANSPOSABLES DE TYPE II

Les transposons à ADN (~3 % du génome) se transposent directement (sans intermédiaire ARN) généralement selon le modèle couper/coller faisant intervenir une transposase qui reconnaît les extrémités de l'élément, le coupe et l'insère à un autre endroit du génome. Leur activité de transposition semble très faible, voire nulle.

CONCLUSION

L'ADN est une chaîne polynucléotidique de structure linéaire dans le noyau cellulaire et circulaire dans les mitochondries. Sa structure est hautement organisée pour permettre sa condensation et sa fonction. Outre les séquences correspondant aux gènes codant les protéines ou les différents ARN non codants, environ la moitié de l'ADN nucléaire est composée de séquences répétées.