ORGANISATION DU GENOME HUMAIN

Pr Dali-Sahi Majda

Les génomes des organismes vivants ne sont pas constitués uniquement de séquences codantes, mais possèdent aussi des fragments dits intergéniques, situés entre les différents gènes. Ces zones non-codantes recèlent une richesse inouïe en termes de diversité de séquences, de mécanismes moléculaires et de fonctions pour l'organisme.

Il y a bien sûr la présence de régions géniques et de région intergéniques au niveau du génome Humain. L'idée est qu'on puisse apporter une définition correcte des régions dites génique et des régions dites intergéniques.

La question qu'on doit se poser est : Quelles sont les séquences qui répondent à la définition de séquences géniques ?

Il y a bien sûr la présence des introns, des exons, des promoteurs et des sites de régulation, des enhancer qui permettent l'expression des gènes, les séquences proches des gènes, fragments de gène et des pseudogènes fragments de gènes et les séquences uniques.

Une autre question s'impose que contient la portion intergénique ?

Il y a bien sur les répétitions dispersées, les microsatellites, les LINE, les LTR, Les SINE et les transposon d'ADN.

Mais quelles sont les éléments de distinction que devons nous retenir pour donnez une définition de l'organisation du génome?

1- En fonction de la cinétique dénaturation de l'ADN

On peut répondre d'emblé que tout **génome** eucaryote peut être divisé en trois **catégories** de séquences désoxyribonucléiques : Chacune de ces **catégories** est définie par une cinétique de renaturation différente de **l'ADN** en solution après dénaturation à haute température (Britten et Kohne, 1968).
RQ : Les séquences **hautement** répétées se renaturant le plus rapidement.

A- La première comprenant les séquences hautement répétées
B- La deuxième les séquences moyennement répétées
C- La troisième les séquences uniques

2- En fonction de données de la Moléculaire de l'organisation du génome

Avec le développement **de la** biologie moléculaire, dans les années 80, et notamment les techniques de séquençage, il a été possible d'obtenir de nouvelles connaissances sur la structure, la nature et l'origine des séquences répétées. La première distinction a **été** de diviser **les** séquences répétitives **en deux** catégories majeures selon le type de distribution dans les génomes.

1- **La première catégorie représente l'ADN répété en tandem, ou encore ADN dit satellite** (John et **Miklos, 1979). Ce sont** en général des motifs simples regroupés en bloc de répétition.

Les ADN satellites composent la majorité de **I'ADN** répété des **génomes** eucaryotes. Ils peuvent représenter jusqu'à 30% de la masse génomique, ne sont en général ni transcrits ni traduits.
a-Sont constitués de motifs peu complexes, et sont retrouvés généralement dans les régions centromériques et télomérique des chromosomes.
b- Ils forment ainsi la majeure partie de I'hétérochromatine constitutive.
c- De façon générale, les **ADN** satellites sont regroupés **en** tandem (John et Miklos, 1979).

La répétition d'une même séquence peut être trouvée à différents locus d'un même génome.

Dans les **génomes.** Les ADN satellites **sont** représentés par les télomères, les " satellites", les minisatellites et les microsatellites.
Les télomères sont constitues, chez les vertébrés, d'un motif de 6 nucléotides(TTAGGG) hautement répété (taille pouvant atteindre 30 Kb) (Morin, 1989).

Ils sont situés à chacune des extrémités des chromosomes. La synthèse des télomères est **effectuée** par une transcriptase inverse spécifique, la télomérase (Greider et lackburn, 1987). Plusieurs fonctions sont attribuées aux télomères : protection vis-à-vis de la dégradation par les nucléases, maintien de la longueur des **chromosomes** lors de la réplication, rôle dans l'organisation structurale via un attachement à la membrane nucléaire (Blackburn, 199 **1).**
Les "grands" satellites **sont** des séquences regroupées en un ou plusieurs **blocs** généralement situés dans les régions centromériques **(Willard,** 199 1). **Les** ADN alpha satellites peuvent constituer de 3 à *5 8* de chacun des chromosornes (Willard. 1991)

Chez l'humain il existe d'autres ADN satellites : les satellites 1.dont l'unité de répétition est **de 42** nucléotides. Et les satellites **II et III** qui sont à l'origine **issus** de la répétition basale ATTCC.
Les minisatellites. ou VNTR (Variable Number Tandem Repeat) :sont des répétitions définies par **un** motif central dont la taille peut varier de 10 à 60 nucléotides.

Les microsatellites, ou SSR (Simple Sequence Repeats), sont des répétitions en tandem de un à cinq nucléotides. Ces motifs sont en général dispersés par petits blocs, **d'un maximum** de 100 paires de bases. a différents loci (Tautz, 1989). Le motif principal chez l'humain est la répétition (CA), - (GT), pour les dinucléotides et (A). - (T). Pour les mononucléotides, séquences provenant essentiellement des terminaisons poly-adénylée (poly.4) des séquences répétées des familles de rétroposons ***A h*** et L1 (Arcot et al., 1995:
Comme les minisatellites. Ces répétitions sont polymorphes et peuvent servir d'outils pour l'étude des populations **(Tautz,** l989), de marqueur génétique pour la cartographie (Weissenbach ***rr*** al.. 1992) et les empreintes génétiques (Econornou ***el*** al.,1990). La variabilité de répétition microsatellite trinucléotidique est un facteur responsable de l'induction **d'un** certain nombre de maladies génétiques récessives. Comme I'ataxie spinocérébrale et I ‘ataxie de Friedreich (Orr ***et al.,*** 1993; **Campuzano *er*** al., 1996) ou dominantes, comme la dystrophie **myotonique (Hunter** et ***al., 1992).***

2-La seconde représente l'ADN répété dispersé intergénique (Rogers, 1985; Schmid).

A l'intérieur de c**ette** catégorie existe une grande hétérogénéité structurale des séquences et celles-ci sont généralement distribuées de façon aléatoire dans le génome. La diversité structurale de toutes ces séquences a permis de définir différents mécanismes responsables de l'amplification de **I'ADN** répété dispersé, tels que la transposition, la duplication, le réarrangement chromosomique et l'intégration de génomes viraux **(Finnegan,** 1 ***989).***

Cependant, les **mécanismes** biologiques dominants responsables de la formation des familles majeures des séquences répétées dispersées sont la rétrotransposition (Boeke et ***al.*.** 1985; Varmus et Brown. 1989) et la rétroposition (Rogers,1985).

Ces **deux** mécanismes font appel à la transcription, **a la** transcription inverse et l 'intégration, qui permettent l'amplification des éléments appelés rétrotransposons (Boeke ***er*** ai.. 1985) et rétroposons (Rogers. 1983). La rétrotransposition est un mécanisme bien défini qui présente de fortes homologies avec l'intégration des rétrovirus (Varmus et **Brown.** 1989)

Les séquences répétées dispersées sont très variées et de structures généralement complexes. Elles représentent plus de 35% de la masse génomique nucléaire chez l'humain.

Elles appartiennent à l'ensemble des séquences moyennement répétées, mais **une** seule famille peut représenter dans certains cas plus de 10% du génome, **comme les** éléments ***ALU***du génome humain (Smit, 1996).

Ces motifs sont distribués tout le long du génome de façon aléatoire entre des séquences uniques de l'euchromatine, ou entre des séquences de I'hétérochromatine (Rogers, **1985).**

Plusieurs de ces séquences ont été étudiées dans différents organismes allant de la **bactérie aux eucaryotes** supérieurs. Ces études ont permis d'établir que la majorité de ces séquences **sont ou ont** été mobiles. Elles ont donc été appelées éléments transposables. La diversité de leurs arrangements et structures a conduit à les classer en deux groupes différents selon leur mode de propagation. Le premier est constitué d'éléments mobilisés par un mécanisme de transposition ADN-ADN, les transposons. Le second groupe se propage à l'aide d'intermédiaires **ARN** et **ADNc, l**es retroelérnents.

Les transposons (fragment D'ADN).

Les transposons **sont** les éléments mobiles qui ont été trouvés dans tous les phvla procaryotes ou eucaryotes. Malgré leur grande variété, il est possible de donner des caractéristiques structurales générales.

1-Leur taille varie entre 500 et **5000** nucléotides.

2-Chaque extrémité de la séquence possède des répétitions de taille variable, de 13 à 1250 pb, en orientation inverse, appelés répétitions terminales inverses.

3-Lors de l'insertion de l'élément dans l'ADN cible, une duplication du site d'insertion de chaque côté du transposon est créée, appelée répétition directe. La taille de ces répétitions, est aussi variable, de quelques nucléotides à une vingtaine.

3-Les transposons possèdent, en général, dans leur structure la protéine nécessaire à leur déplacement, **la transposase** ou **l'intégrase**. En effet, une seule **enzyme** est nécessaire et suffisante, ***Ni vitro,*** pour induire le mécanisme de transposition (Kaufman et Rio, 1992).

L**es** transposases et les intégrases **possèdent** un site catalytique fortement conservé qui **est** responsable de la transphosphorylation nécessaire pour la coupure de l'ADN et le transfert pendant la transposition. Le site catalytique a pour particularité de rassembler deux résidus aspartate et un glutamate (Labrador et Corces, 1997). La transposition peut être soit **réplicative,** auquel cas **une** nouvelle copie est donnée à **I'ADN,** soit conservative, le transposon est déplacé **d'un** site de **l'ADN** à un autre. Dans tous les cas de figure, le transposon ne se trouve jamais à l'état libre dans la cellule. **De façon générale, la transposition est un événement rare. Chez l'humain les transposons représentent moins de 2% du génome** (Smit. 1996) et semblent avoir perdu toute activité de mobilité **(Oosumi** et *o/.* , **1995;** Smit *et* Riggs. 1996).

Les rétroéléments(fragment intermédiaire D'ARN).

Les rétroéléments ont des structures très variées. Ils utilisent divers mécanismes de propagation. Cependant tous nécessitent l'activité d'une transcriptase inverse, impliquant donc un intermédiaire **ARN** .

Suivant les structures des éléments et leurs identités avec les rétrovirus, les rétroéléments ont été séparés en **deux** groupes majeurs : la famille des rétroéléments de type viral et la famille des rétroéléments de type non viral. Les rétrotransposons sont très faiblement représentés, et sans doute **même** absents, dans le **génome** humain.

Les rétroposons(fragment d'ARN)

Les LINE"(pleine longueur **):**

ce sont des rétroposons actifs de grandes tailles. Le terme actif est associé directement au fait que les LINE possèdent au moins un cadre de lecture qui code pour les activités nécessaires à leur amplification.

II a été proposé a la **fin** des années 80. que les éléments LINE "pleine longueur" possèdent un promoteur interne reconnu par I'ARN polymérase **Il,** notamment par le fait qu'ils sont constitués de séquences riches en **G+C**

Les SINE ("Short Interspersed Element")

A l'inverse des éléments L M . les **SNE** ("Short Interspersed Element") sont des éléments de petite taille (entre 100 et **100** nucléotides), définis comme étant des rétroposons passifs. Ils ne codent pour aucune protéine. Et sont donc dépendants de l'activité retropositionnelle d'éléments actifs comme par exemple les LINE ou les rétrovirus.

Les séquences géniques

Les gènes

Nous choisissons d'appeler gène la fraction d'ADN capable de diriger la synthèse d'une protéine ou d'un ARN non codant.

On peut choisir de parler plutôt d'allèles ce qui est plus correcte puisque nous sommes mis d'accord dans les cours précédent que le gène physique n'existe pas.

Les allèles sont les deux séquences qui occupent le même site sur deux chromosomes homologues. L'un d'eux représente l'information venue du père, et l'autre, celle venue de la mère.

C'est en particulier le cas des ARN ribosomiques (ARNr) qui représentent près de 80% du transcriptome. La famille des ARNr est composée de 4 ARN chez les eucaryotes qui sont les ARN 28 S, 5,8 S et 5 S qui forment l'ossature sur laquelle s'assemble la grande sous-unité ribosomique 60 S et l'ARN 18 S sur lequel s'assemble la petite sous-unité 40 S.

Si le rôle attribué aux ARNr a longtemps été vu comme purement structural, il est apparu plus récemment avec les premières données cristallographiques sur le ribosome obtenues en 2000 que le cœur catalytique du ribosome qui assure la formation de la liaison peptidique est en fait porté par l'ARN (Moore *et al*, 2011). Ainsi, non seulement, la classe la plus abondante d'ARN ne code pas une protéine mais elle fournit aussi l'activité catalytique au cours de la traduction.
Les ARNr et ARNt ne sont d'ailleurs que les premiers d'une longue liste d'ARN non codant découverts par la suite comme les petits ARN nucléaires (ou snRNA pour « small nuclear RNA ») impliqués dans la machinerie d'épissage ou les petits ARN nucléolaires (ou snoARN) qui participent à la maturation des ARNr.
Au-delà des ARNr et des différents types de petits ARN, qui pour l'essentiel sont transcrits par des ARN polymérases spécialisées (Pol I et Pol III), des ARN non codants d'une taille et d'une organisation similaire à celle des ARN messagers, transcrits par la Pol II, ont progressivement été décrits. Le fait qu'ils étaient souvent associés à des phénomènes de suppression de l'expression, comme Xist pour l'inactivation de l'X ou H19 pour l'empreinte parentale, suggérait un rôle régulateur dispensé aux ARN.non codants.

Séquences proches des gènes

Les pseudogènes

Il y a dans le génome humain environ 20 000 pseudogènes [1]. On a aussi montré que leur similitude avec les gènes fonctionnels pouvait être à l'origine de réarrangements et de conversions géniques, surtout quand gène et pseudogènes restaient voisins, ce qui n'est pas toujours le cas. La délocalisation vers un autre chromosome n'est, en effet, pas exceptionnelle. Mis à part ces fonctions indirectes.
La transcription des pseudogènes semblait peu probable ; un certain nombre de transcrits ont cependant été identifiés, parfois des épissages alternatifs, ce que peut expliquer soit le maintien des séquences régulatrices, soit l'utilisation d'un autre promoteur de voisinage. Du fait de mutations de la séquence codante, la traduction en une protéine n'était pas possible.
Le dogme de la non-fonctionnalité de ces pseudogènes se trouvait cependant partiellement remis en question par l'identification de ces transcrits, et de fait, le rôle fonctionnel d'un pseudogène vient d'être mis en évidence. Leur rôle serait de réguler l'expression des gènes .
Ainsi, ces pseudogènes entrent en compétition avec les gènes fonctionnels pour l'interaction avec les miRNA : les chercheurs les ont surnommés *competitive endogenous RNA* ou ceRNA.

Introns.

Les introns sont définis comme des segments de gènes transcrits en ARN mais retirés par un processus d'épissage lors de la maturation de ce dernier. Ils permettent de contribuer à la diversité des transcrits produits à partir d'un même gène. Les introns représentent la moitié du génome humain (UCSC, GCRh38) et leur taille varie de quelques dizaines de nucléotides à la mégabase. Le plus petit intron dont l'épissage a été montré expérimentalement possède une longueur de 43 nt [230]. Les introns des gènes codants ont une longueur médiane d'1,5 kb et les introns des gènes non-codants de 1,7 kb.
Aujourd'hui, de nombreuses fonctions ont été associées aux introns et ils sont impliqués dans différentes étapes allant de la transcription à la traduction pour les gènes codants. Les séquences introniques influencent ainsi l'épissage alternatif [247], elles peuvent amplifier

l'expression des gènes [242], contrôler le transport de l'ARN [265] et les introns des extrémités 5' et 3' des gènes affectent la dégradation des ARNs non-sens [153]. Les introns sont également impliqués dans diverses fonctions de façon indirecte : la longueur des introns est importante dans l'évolution des génomes et ils peuvent contenir de nombreux gènes non-codants [116]

Exons

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

UTRs. Les régions non traduites situées aux extrémités 3' et 5' d'un gène sont les UTRs (Untranslated transcribed region, UTR). On distingue les 5'UTRs à l'extrémité 5' des 3'UTRs à l'extrémité 3'.

Le 5'UTR est défini comme la région située entre le TSS (qui contient la majorité des sites de fixation pour le recrutement de l'ARN polymérase) et le codon d'initiation de la traduction (ATG au niveau de l'ADN, AUG au niveau de l'ARN transcrit). Il contient des éléments régulateurs et joue un rôle important dans le contrôle de l'expression des gènes. Lors de la transcription, l'extrémité 5' des ARNm est coiffée ce qui permet de le protéger et de le stabiliser. La proximité de la coiffe au codon initiateur, la composition nucléotidique et la structure secondaire du 5'UTR influencent l'initiation de la traduction [203].

Les 5'UTRs peuvent également contenir des cadres de lecture amont ou uORF (Upstream open reading frame, uORF). Les uORFs possèdent leur propre codon initiateur (Upstream AUG,) en amont du site principal associé au gène et leur propre codon stop [24]. Chez l'homme, ils sont présents dans environ 50% des 5'UTRs et peuvent être traduits. Leur présence corrèle avec une diminution de l'expression à la fois au niveau de la quantité d'ARNs produite et de protéines.

La région 3'UTR est localisée en aval du site de terminaison de la traduction. Cette région est impliquée dans de nombreux processus de régulation comme le clivage de l'extrémité 3' de l'ARN transcrit, la stabilité de l'ARNm et sa polyadénylation, ou encore la traduction et la localisation de l'ARNm [172]. Chez l'homme, une grande portion des gènes codants pour des protéines utilisent des sites de clivage et de polyadénylation alternatifs pour générer des 3'UTRs alternatifs ayant un impact sur le devenir de l'ARN [259, 65]

Les séquences proches des gènes ou SEQUENCES UNIQUES

Ce sont des séquences très conservées, non codantes, réparties sur tout le génome.

Elles sont souvent situées près des régions régulatrices de gènes des fonctions essentielles de la cellule (gènes du développement, ARNr...). Elles ont un rôle dans la régulation de l'expression des gènes et dans l'appariement correct des chromatides lors de la division cellulaire.

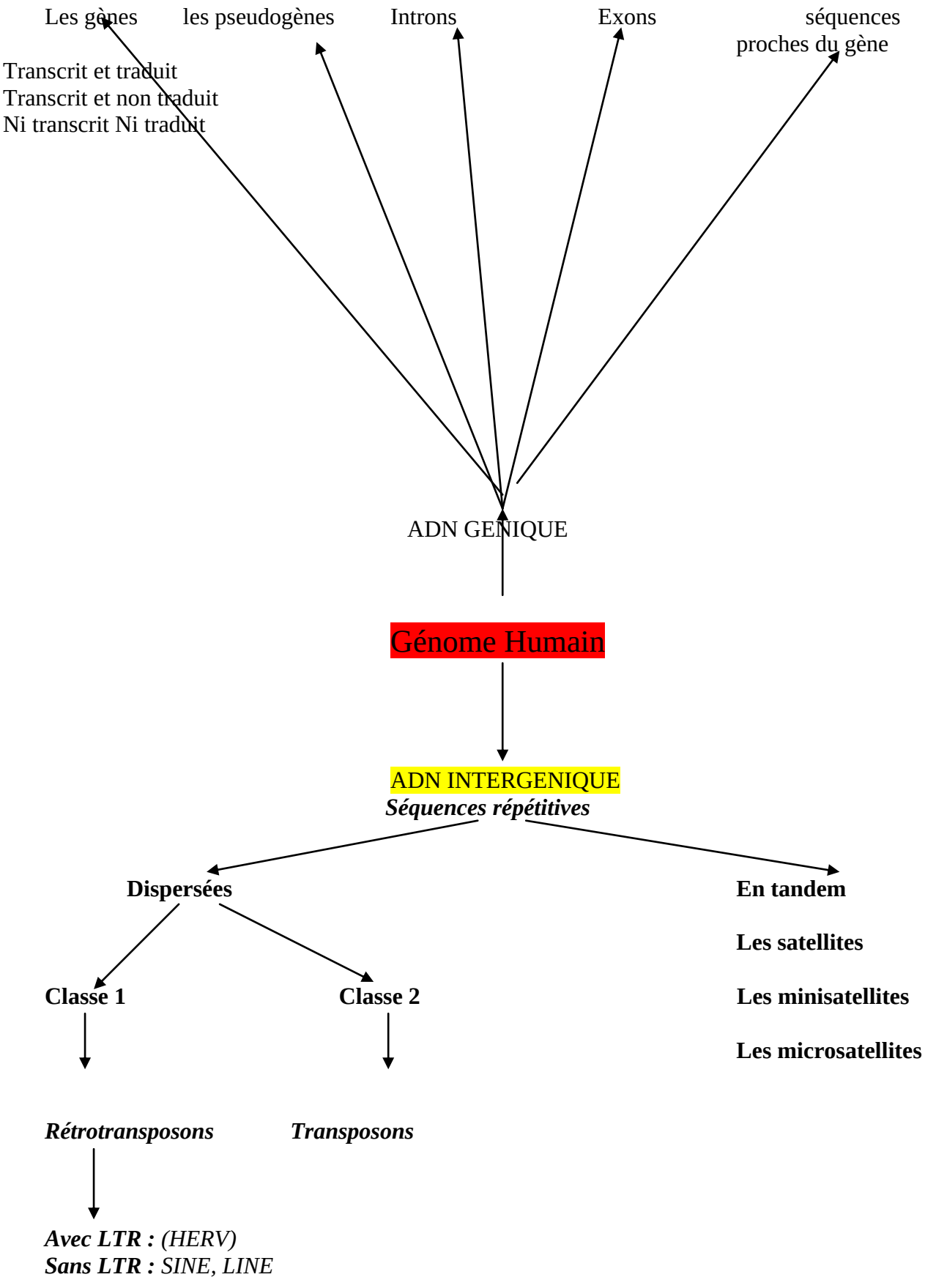
Pour conclure

Les séquences répétées Il existe dans le génome des mammifères des séquences répétées, dont le plus souvent on ignore complètement le rôle. Certaines de ces séquences sont regroupées dans les régions centromériques et télomérique.
D'autres sont dispersées dans tout le génome et peuvent même se groupées et se suivent en tandem.

On peut citer :
Les satellites Les microsatellites Les minisatellites ou VNTR *(variable number of tandem repeats)* Les séquences SINE *(short interspersed elements)* Les séquences LINE *(long interspersed elements)*

Par proportion nous pouvons résumer ainsi

Comme nous l'avons déjà mentionne, les exons des gènes codants représentent 2% seulement du génome.
Les 98% restants correspondent a L 'ADN non codant et environ 50% du génome est représenté par les éléments répètes qui sont des séquences que l'on retrouve en plusieurs copies.


Les gènes
les pseudogènes
Introns
Exons
séquences proches du gène
Transcrit et traduit
Transcrit et non traduit
Ni transcrit Ni traduit
ADN GENIQUE
Génome Humain
ADN INTERGENIQUE
Séquences répétitives
Dispersées
En tandem
Les satellites
Les minisatellites
Les microsatellites
Classe 1
Classe 2
Rétrotransposons
Transposons
Avec LTR : (HERV)
Sans LTR : SINE, LINE