UMR Génétique 1 – Génomique, génétique formelle et épidémiologique – Génomique – Cours n°1 Mercredi 13 février 2019 Michel Vidaud – michel.vidaud@parisdescartes.fr	RT : Alice MARCOPOULOS RL : Louise LIMES

Structure, organisation, dynamique et polymorphisme du génome humain

Plan

Mot du professeur :

- *Concernant l'assiduité*

(Il y a une feuille de présence à signer à chaque cours.)
Le cours n°3 est fait par le président de l'Université Paris Descartes (Pr. Frédéric Dardel), le ***27 février*** *: si on est que 5 alors que la feuille de présence indique 160 inscrits, il partira tout de suite. Il rappelle que pendant ces cours, ce ne sont pas toujours des professeurs payés mais ce sont des chercheurs qui viennent nous transmettre leur savoir et que la moindre des choses est de se déplacer. Par ailleurs, le second cours est fait par Antonio RAUSELL : si des chercheurs savaient qu'il fait cette conférence, ils viendraient écouter sa conférence. Les cours sont de grande qualité. Tout le programme est très important et s'inscrit parfaitement dans ce qu'il se passe en ce moment en génomique.*

- *Concernant le master en génétique*

A propos du master de génétique proposé par les universités Paris V et Paris VII, le professeur explique que c'est un enseignement reconnu à l'international et extrêmement concurrentiel, un des meilleurs masters en Ile-de-France. Pour y candidater, il faut avoir validé les 3 UMR et le stage de 2 mois afin d'être auditionné, et éventuellement retenu pour ce master. C'est un master d'excellence que beaucoup d'étudiants en sciences aimeraient faire. Les masters ont répondus à un appel d'offre ministériel dans le cadre d'un projet qui s'appelle les EUR : Ecole Universitaire de Recherche. L'état a mis 360 millions d'euros pour aider à la pédagogie et stimuler la filière PHD en lien avec le master. Le master de génétique fait partie des deux entités retenues à l'échelle de l'Université de Paris ; ils ont obtenu 16 millions d'euros pour construire un parcours d'excellence M2-PHD ouvert à l'international.

- *Concernant le stage*

Le stage se fait dans un laboratoire labellisé, il existe des listes de lieu de stages ; c'est organisé notamment par Christine Guérin.
Les stages à l'étranger sont également possibles, mais le professeur conseille de rester en France, excepté si vous connaissez quelqu'un, que vous êtes attendu là-bas (par de la famille par exemple). Il faut alors contacter les responsables de stage.
Le stage donne lieu à la rédaction d'un mémoire écrit de 10 pages, suivi d'une soutenance orale de 13 minutes qui a lieu durant la deuxième quinzaine de septembre. Il y a une note d'assiduité au stage, une pour le rapport écrit et une dernière pour la soutenance du mémoire.

- *La personne à contacter*

Christine Guérin est la gestionnaire du parcours d'initiation à la recherche en génétique. Pour toute question concernant le parcours, le master ou le stage de génétique, vous pouvez la contacter à l'adresse suivante : christine.guerin@parisdescartes.fr.

Ce cours n'a rien à voir avec les autres de l'UMR 1 (=premier semestre de parcours de génétique, le semestre actuel). Cependant, il présente les bases nécessaires pour comprendre la génétique de demain.

I. Encadrement règlementaire des activités de génétique en France : Convention d'Oviedo, Lois de bioéthique

En 2019, on ne peut pas commencer un cours de génétique sans faire quelques minutes d'éthique.

La **convention d'Oviedo** est le seul instrument juridique international contraignant en matière de bioéthique. Il a été ratifié au niveau européen.

Les quatre articles de cette convention à l'échelle **planétaire** sont à apprendre en tant que référence :

- **Non-discrimination** : « Toute forme de discrimination à l'encontre d'une personne en raison de son patrimoine génétique est interdite. »
- **Tests génétiques prédictifs** : ils ne peuvent être faits sans consentement.
- **Interventions sur le génome humain** : on ne peut pas faire de modifications génétiques qui induiraient une modification de la descendance (thérapie génique germinale ou manipulation des gamètes par exemple)
- **Non-sélection du sexe** : il est interdit de sélectionner le sexe de l'enfant à naître.

Il existe actuellement des discussions pour déterminer si certaines de ces interdictions ont du sens ou non, notamment quant au fait de guérir un embryon en faisant une manipulation de type génome-editing.

Une disposition supplémentaire a été ajouté, elle est aujourd'hui essentielle : c'est la **primauté de l'être humain**. « **L'intérêt et le bien de l'être humain concerné par les tests génétiques visés par le présent Protocole doivent prévaloir sur le seul intérêt de la société ou de la science**. »

En France, depuis 1994, on a des **Lois de bioéthique**, normalement révisées tous les cinq ans (révisées en réalité en 2004 puis en 2011).

Le mot bioéthique est constitué de deux racines : « bio » signifiant « **vivant** » et « éthique » « **ce qui est bon pour l'homme** ». L'objectif est que les progrès faits soient bons pour l'homme.

En 2018, s'est engagée la révision des Lois de bioéthique. En tant qu'étudiant en santé, il n'est pas possible que nous ne nous intéressions pas à ces Lois.

Plusieurs documents sont **à lire** (disponibles en ligne) :

- Le **comité consultatif national d'éthique** rend des avis. Nous en sommes au **129e** qui abordent de nombreux sujets tels que le diagnostic préconceptionnel, le dépistage en population, la possibilité de tests sur les embryons…

- Il est obligatoire de lire le **rapport des Etats généraux de la bioéthique 2018**.

- En France, le conseil d'Etat est très important : il est le garant des aspects règlementaires. Leur rapport : « Révision de la loi de bioéthique : quelles options pour demain ? » est à lire.

- L'office parlementaire d'évaluation des choix scientifiques et technologiques (**OPECT**), qui est composé de membre de **l'assemblée nationale** et du **sénat**, a rédigé un rapport. Il est intéressant à lire car la position des conseillers d'état quant aux questions bioéthiques n'est pas toujours celle des parlementaires ni celle des comités nationaux d'éthique (il existe un grand débat à ces sujets).

- **L'agence de la biomédecine** (agence régulant la génétique, la procréation médicalement assistée, la greffe…) a fait un rapport sur ces lois de bioéthique. Il met en évidence la diversité internationale de ces lois. Aujourd'hui, il est nécessaire d'unifier les lois à un niveau européen. Par exemple, si des choix différents sont faits en France et en Espagne, il y aura des problèmes à Perpignan car il suffira de traverser la frontière pour changer de cadre règlementaire.

Les révisions de 2018 sont les plus importantes en raison des grands progrès en génomique. Nous sommes à un moment clé d'une évolution de la société, il faut impérativement faire ces lectures.

II. Génétique et séquençage

A. Généralités

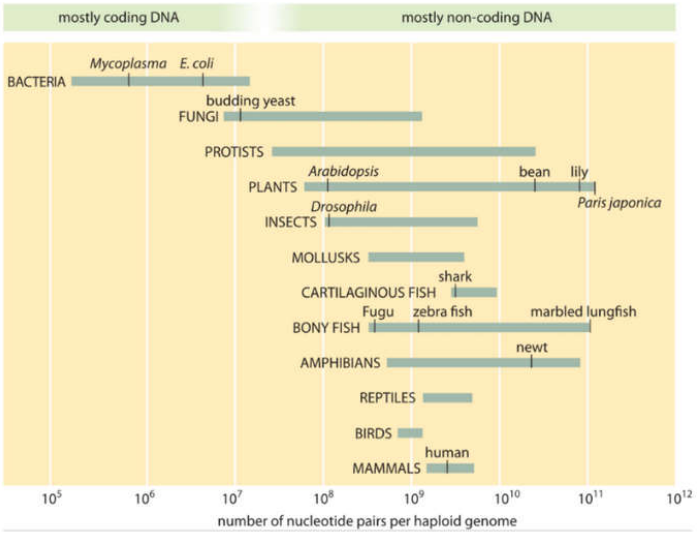
Le mot **génome** désigne l'ensemble de l'information héréditaire d'un organisme, présente en totalité dans chaque cellule, plus précisément contenue dans le noyau. Le support matériel de l'information génétique est l'**ADN** (A, T, C et G), plus rarement l'**ARN** (A, U, G et C).

La **valeur C** représente la taille d'un génome, qui peut être exprimée en en *millions de paires de bases* (Mb = Mégabase), ou en *pico-grammes* (pg=10^{-12} g) : **1 pg correspond à 978 Mb**.

Cette valeur est très variable en fonction des espèces :

- Virus : quelques milliers (HIV, HVC ~ 10 000 bases, HVB = 3 500 bases)
- Bactéries : quelques millions (colibacille = 4 millions)
- Mammifères : quelques milliards de paires de bases (ou 6 pg d'ADN par cellule chez l'humain)

Tout ce qui vit ou presque a été séquencé. On constate donc rapidement que **l'être humain n'a pas le plus gros génome**, mais en fait, **la taille du génome n'est pas corrélée avec la complexité d'un organisme**. C'est le paradoxe de la valeur C : on pourrait imaginer que plus le génome est grand et plus il va être complexe, mais ce n'est pas du tout le cas.



En effet, si l'on place la taille des génomes en fonction du positionnement des espèces, l'humain a un génome petit par rapport à certains organismes *(Protopterus aethiopicus : 130 000 Mb)* ou certaines plantes *(Paris japonica (plante): 149 000 Mb).* Le génome humain quant à lui fait 3 milliards de paires de bases, c'est-à-dire 3 000 Mb.

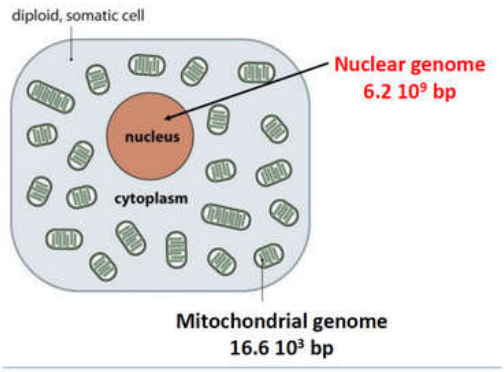
Par ailleurs, le nombre de gènes varie moins que la taille des génomes.

La différence du nombre de gènes est relativement homogène entre les espèces.

Mot de la RT : Durant le cours, le prof ne fait pas de réelle distinction entre base et paire de bases (cela dépend juste du fait que l'on considère 1 brin ou les 2), et on peut aussi bien parler de nucléotides. Par ailleurs, je tiens à clarifier la taille du génome. Le prof dit qu'il fait 3 milliards de pb, mais aussi 6 000 Mb. En fait, les cellules humaines étant diploïdes, elles contiennent notre information génétique en double. Donc on a une information génétique d'une taille de 3 Mb, mais un contenu substantiel en information génétique de 6Mb. J'espère que c'est plus clair ^^

B. Le génome nucléaire et mitochondrial

■ Le génome humain est dans le **noyau** et fait 6,2 milliards de paires de bases *(en effet, l'information contenue est en double dans le noyau, donc en termes de contenu informatif, on a 3 milliards de pb, mais en substantiel, cela revient à compter 6 milliards de pb)*. Il est composé de 23 paires de chromosomes, qui sont des molécules linéaires. **Le génome nucléaire n'est contenu qu'une fois par cellule**, puisque l'on a qu'un seul noyau.
Les chromosomes ont des **tailles différentes**, de 48Mb pour le chromosome 21 à 249 pour le chromosome 1.



■ Il ne faut cependant pas oublier que l'**on possède un génome dans la mitochondrie**, plus petit de l'ordre de 16 600 pb. Ce génome est **circulaire** (pas besoin de protéger les télomères), fait d'une seule molécule, et il est présent par centaines ou par milliers de copies dans chaque cellule. En effet, il y a **entre 2 et 10 copies d'ADN mitochondrial par mitochondrie** et il y a **plusieurs mitochondries** par cellules (la quantité de mitochondrie dépend du type cellulaire). La quantité d'ADN mitochondrial par cellule est finement régulée.

■ Le génome mitochondrial code seulement **13 protéines** (principalement dans la chaîne respiratoire) alors qu'il a plus d'un millier de protéines dans les mitochondries. Les nombreuses autres protéines sont codées par les gènes du noyau et adressées à la mitochondrie.
Si la mitochondrie est malade, c'est soit à cause de mutations du génome mitochondrial (très rare), soit à cause de mutations du génome nucléaire.

De plus, il y a **24 gènes codant des ARN**. Certains codent des **ARNt,** ces derniers étant spécifiques de la mitochondrie. En effet, le code génétique de la mitochondrie est différent du code génétique du noyau, les anticodons sont différents. Les autres ARN sont des **ARNr**.

Enfin, on rappelle que l'hérédité du génome mitochondrial est **exclusivement maternelle** car les mitochondries du spermatozoïde sont détruites. En revanche, quand la maladie est liée aux gènes nucléaires, on aura l'hérédité mendélienne basique.

C. L'histoire du séquençage

✓ Le premier séquençage date de **1968** et concerne le phage filamenteux λ. On a pu séquencer **12 bases** nucléotidiques.

✓ Le deuxième épisode de séquençage est publié quelques années plus tard, en 1979 : des français (dont Francis Galibert, Patrick Charnay et Pierre Tiollais) ont séquencé le **VHB** (3 182 pb), moment fort de l'histoire du séquençage, car c'est le **premier virus séquencé en entier.**

✓ La méthode de Sanger *(à connaitre par cœur ad vitam aeternam)* : découverte par Frédéric Sanger (qui a reçu 2 prix Nobels) à partir des didésoxyribonucléotides en 1977, cette méthode a été automatisée par Leroy Hood en 1986, grâce à des molécules fluorescentes accrochées sur les ddNTP. **C'est cette méthode qui a permis de séquencer le génome humain pour la première fois en 1989.**
On a séquencé les 3 milliards de paires de bases par morceaux de 1 000 bases en parallèles dans plusieurs séquenceurs. La première version du génome complet a été publiée en **2001**, suivie de la version définitive en **2004**. Il aura donc fallu **13 ans**, 2800 scientifiques, 16 instituts, 6 pays et **2,7 milliards de dollars** pour obtenir le génome de référence. En effet, il fallait séquencer par petits morceaux, puis tout remettre dans l'ordre avant d'obtenir la version finale.

Quelques noms sont à retenir : **Eric Lander**, leader incroyable de la génomique américaine et **Francis Collins**, qui a dirigé le projet génome américain et qui est à l'origine du premier séquençage du génome humain.

✓ Avec la révolution du NGS, on a pu séquencer de façon symbolique le génome de son inventeur en 2008 (James Watson), par séquençage parallèle massif. Cela aura pris 4 mois et demi et 1,5 millions de dollars.

✓ Aujourd'hui, séquencer le génome va beaucoup plus vite car on n'est plus obligé de l'aligner manuellement pour le reconstituer. L'alignement est fait automatiquement par rapport à un génome de référence (la carte génomique étant désormais déjà faite). On a déjà dépassé les 15 000 génomes complets et les 130 000 exomes. **En 2018, on peut séquencer un génome humain complet en moins d'un jour, pour moins de 1 000 dollars.** Le record actuel de séquençage du génome complet (avec interprétation des variants) date du 12 février 2018 et correspond à un temps inférieur à 24h (19,5h). On appelle ça le RWGS *(rapide whole genome sequencing).* Il est désormais raisonnable de séquencer un génome et de l'interpréter en une semaine.

L'ampleur de la découverte du séquençage est comparable à celui de la découverte du microscope.

Tableau récapitulatif de l'évolution du séquençage

Genome sequenced	HGP (2003)	Waston (2008)	2018
Time taken (start to finish)	13 years	4.5 months	1 day
Number of scientists listed as authors	> 2 800	26	-
Cost of sequencing (start to finish)	2.7 billion $	# 1.5 million $	# 1 000 $
Number of institute involved	16	2	-
Number of countries involved	6	1	-

D. L'évolution technologique du séquençage

Le 7 avril 1964, IMB annonce la mise sur le marché des ordinateurs de la série **IBM 360**. Cette annonce marque un tournant radical dans le monde de l'informatique et est perçue comme une véritable révolution du concept d'ordinateur. C'est le premier ordinateur jamais construit, qui représente un point clé dans l'évolution du séquençage génétique, ouvrant la possibilité d'automatiser les traitements via l'outil informatique.

Pour bien se rendre compte de l'impact évolutif du séquençage, on peut faire un parallèle entre l'annonce de l'Ipad en 2010, après la sortie d'IBM 360 en 1964, et le fait qu'en 2018 on séquence 10^{12} bases par jour alors qu'on en était à peine à 10^3 en 1987.

On est rentré dans le monde du *Big Data*. Séquencer n'est plus un problème, autant du point de vue technique que du point de vue financier (le prix du séquençage est littéralement tombé). La difficulté réside dans le stockage et l'interprétation de ces données. Aujourd'hui, il est impossible d'analyser un génome sans une aide informatique.

Maintenant, puisqu'on a pu séquencer de nombreux génomes, on a réussi à définir assez précisément de quoi se composait le génome humain, et c'est précisément l'objet de ce cours.

III. Les caractéristiques du génome humain

Définir un gène est aujourd'hui une question complexe. La définition n'a rien à voir avec ce qu'était un gène il y a dix ans. Un gène n'est pas un caractère, on ne le définit plus tel que « un gène code une protéine ».

Définition : Un gène est un ensemble de **séquences génomiques** codant un **ensemble cohérent** de **produits fonctionnels** potentiellement chevauchant.

Quelques chiffres à connaitre :
• Le génome humain fait **6 milliards de paires de bases** (10^6bp)
• Il comporte **près de 60 000 gènes** (~58 720)
• Un peu **moins de 20 000 gènes codent pour les protéines**, dont les exons occupent **1,2%** du génome.

A. Les gènes codant pour les protéines

Un gène est transcrit en ARN messager qui passe dans le cytoplasme pour être traduit en protéine. Cet ARNm a pour seul but d'être transformer en protéine. La fonction est assurée par la protéine. On peut modéliser et réfléchir à beaucoup de choses quant à ces gènes car on connait le code génétique et les protéines grâce aux nombreuses recherches faites en biochimie depuis 50ans.

Exon : partie du gène qui **persiste** lors de la maturation (épissage) du transcrit primaire en ARNm. Les exons **ne sont pas tous codants.**

Intron : partie du gène située entre deux exons et qui est **excisée** lors de la maturation (épissage) du transcrit primaire en ARNm.

UTR (UnTranslated Region) : régions transcrites **non traduites** (5'UTR et 3'UTR).

Séquences consensus : séquences nucléotidiques impliquées dans des **fonctions semblables** et ne présentant entre elles que quelques variations. (ex : jonctions intron/exon, séquence de Kozak, signal de polyadénylation, TATA box etc…)

• Si les gènes sont en nombre assez constant d'une espèce à l'autre, chez l'Homme leur taille est extrêmement variable (SRY < 1 kb vs. dystrophin = 2,4 Mb).

• La plupart des gènes sont morcelés en **exons**, et il y a plus de **220 000 exons** dans un génome (= **exome**), séparés par des **introns** dont le nombre est très variable d'un gène à l'autre. Quelques gènes n'ont pas d'exons ni d'intron. Le gène le plus riche en exon, la titine, en possède 315 (mis dans le bon ordre lors de l'épissage).

• Un exon a un message de traduction de protéine, mais il a aussi en lui un message qui lui permet d'être reconnu en tant qu'exon via des séquences spécifiques qui fixent des protéines **enhancer** d'épissage (qui garde l'exon) ou **silencer** d'épissage (qui l'exclut). Le principe est similaire à celui des protéines qui se fixent à l'ADN du promoteur pour permettre ou non la transcription.
Une mutation d'un exon, même si elle ne change pas la nature des acides aminés codés par cet exon, pourra modifier un site enhancer d'épissage et l'exon pourra alors être exclu.
Les conséquences des mutations des exons **ne doivent plus être interprétées à la seule vue du code génétique !**

• Tous les gènes (ou presque, soit plus de 95%) sont soumis à **transcription**, **épissage et maturation 3' alternatifs**. Cela veut dire que pour un gène donné on peut avoir :

- Des promoteurs différents (donc on inclut des exons différents)
- Des épissages différents (donc on garde ou on élimine des exons différents, selon un mécanisme très complexe d'*enhancer* et *silencer* d'épissage)
- Des polyadénylations alternatives, c'est-à-dire utiliser des signaux pour maturer différemment les extrémités 3' (par exemple, la sécrétion ou la localisation membranaires des immunoglobulines dépend de la queue polyA alternative).

• Les gènes qui codent pour des protéines vont être transcrits en **ARN messagers**, **coiffés**, **épissés** et **maturés**, mais qui n'ont pas de structure particulière. Ici, **c'est la protéine qui acquiert une structure spécifique, qui lui confère sa fonction.**

• Les gènes ne sont pas toujours en copie unique, ils peuvent être **répétés** : les uns à côtés des autres (en cluster) ou dispersés.

• Certains gènes sont exprimés dans toutes les cellules (**house keeping**) tandis que d'autres sont uniquement dans certains types cellulaires (**tissus-spécifiques**).

B. Les ARN non-codants

La proportion de gêne non codant augmente avec la complexité des espèces. De plus, on a **plus de gènes qui produisent des ARN** (> 23 000) **que de gènes qui produisent des protéines** (< 20 000). On en déduit que **la fonction de ces ARN va dépendre de leur structure**.

Ainsi, les petits ARN non-codants (< 200 bases, on en compte 7 500) ont une structure connue (cf. ARNt) alors que pour les grands ARN non-codants (15 800 comptés à ce jour, de l'ordre de taille d'une protéine), on n'a pas encore déterminé de structure. Lors d'une mutation, il faut se demander si la mutation perturbe ou non sa structure, il n'est pas question de code génétique.

La découverte de ces ARN constitue la deuxième révolution en génétique après le séquençage. Andrew Fire et Craig Mello sont à l'origine de cette révolution et ont reçu un prix Nobel pour la découverte des ARN interférents.

Parmi les ARN non-codants on distingue :

- Les **ncRNA** (ARN non-codants classiques) : tRNA (ARN de transfert), rRNA (ARN ribosomal)

- Les **sRNA (**petits ARN non-codants < 200 nt) : snRNA *(small nuclea RNA)*, siRNA *(small interfering RNA)*, piRNA *(piwi-interacting RNA)*, miRNA *(microRNA)*, snoRNA *(small nucleolar RNA)*

- Les **lcnRNA** (grands ARN non-codants < 200 nt), il y a quatre catégories : les intergéniques (entre deux gènes codants des protéines), les introniques (dans l'intron d'un gène), les antisens (est transcrit à l'envers par rapport gènes codants les protéines) et les divergents.

Ils ont divers rôles majeurs, comme XIST qui inactive l'un des deux chromosomes X. Cependant, nous ne savons pas bien interpréter tous leurs rôles, ni même l'impact que peut avoir un variant dans un ARN non codant.

C. Les pseudogènes

On trouve ensuite environ **15 000 pseudogènes**, c'est presque autant que le nombre de gène codant des protéines. Un pseudogène peut être transcrit. Il existe plusieurs catégories de pseudogènes.

• **Les pseudogènes dupliqués** : c'est la copie conforme du gène, juste à côté. L'un des gènes reçoit une mutation dans son promoteur et n'est plus transcrit. S'il n'est plus transcrit, il va pouvoir accumuler beaucoup de mutations. Le pseudogène (comme un cancre à côté d'un bon élève) pourra nuire à sa copie transcrite en créant des délétions et des duplications suite à des crossing over ectopique voir même faire de la conversion génique.

• **Processed pseudogene** (=rétropseudogènes) : un gène est transcrit en ARN, l'ARN est transformé en cDNA qui s'insère dans le génome. Il ne possède **pas de promoteur** (car il est issu d'un ARN) et va ainsi accumuler des mutations.

• **Retrogènes** : un gène est transcrit en ARN, l'ARN se transforme en cDNA, cependant, cette fois ci, le cDNA s'insère dans un endroit où un promoteur est présent. Ce gène, sans intron, **pourra alors être exprimé.** Ces gènes rétroprocessés actifs sont appelés rétrogènes.

Quelques exemples mettent en évidence l'importance des rétrogènes :
- Certains spermatozoïdes (les mâles) n'ont pas de chromosome X, or des protéines capitales de la chaîne respiratoire sont codées par le X. Il existe des rétrogènes de ces protéines sur d'autres chromosomes qui s'activent alors.
- Plus un organisme est grand, plus il a de cellules en division, plus il a une grande probabilité de développer un cancer. Pourtant l'éléphant et la baleine ont une résistance au cancer(=paradoxe de Peto). L'éléphant, grâce à plusieurs rétrogènes issus de P53, est protégé.
- PTEN est un gène majeur en cancérologie. Il possède un pseudogène qui est transcrit. S'il est délété, il y a des pathologies. Une fois transcrit, il intervient dans la transcription de PTEN par rapport aux micro-ARN. En l'enlevant, le tout est déstabilisé.

Il existe donc des pathologies liées aux anomalies des pseudogènes.

D. Les séquences répétées (= the human mobilhome)

45% du génome est composé de séquences répétées, représentées en majorité par des éléments transposables, c'est-à-dire mobiles (SINE, LINE, HERV). Par exemple 42% du chromosome 22 correspondent à des séquences répétées. Il ne faut pas en faire de l'ADN poubelle ! Elles ont un rôle majeur dans le génome et on en réalise aujourd'hui l'importance. *On doit la découverte de ces éléments génétiques mobiles à Barbara McClintock, qui a par ailleurs été la première femme à recevoir un prix Nobel seule pour ses travaux.*

On distingue ainsi :

⇢ Les ADN transposons

⇢ Les rétrotransposons

⇢ Les HERV

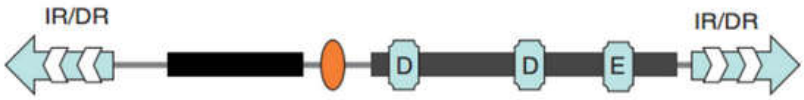
⇢ Les séquences répétées tandem

⇢ Les séquences non-codantes hautement conservées

i. Les ADN transposons

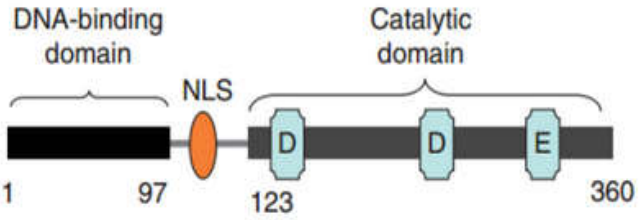
Occupant près de **3%** du génome, ils sont inactifs depuis plus de 40 millions d'années chez l'Homme (mais ils fonctionnent encore chez certaines espèces). Ils codaient pour une enzyme, la **transposase**, qui reconnait les séquences répétées inversées dites **ITR**, disposées de part et d'autre du transposon (gène d'intérêt). La transposase va donc couper au niveau de ces séquences, pour déplacer ce qui était encadré et l'intégrer ailleurs dans le génome. C'est une mécanique de type couper-coller (transposition **conservative**) du transposon, qui est alors **mobile**. Il existe également des versions de transposition **réplicative** : mécanisme de type copier-coller. Selon les transposons, un mécanisme ou l'autre, ou les deux sont employés.

Structure d'un transposon



Certains gènes sont dérivés de transposons et codent pour des protéines de l'immunité tels que les gènes *RAG1* et *RAG2*, impliqués dans la recombinaison des gènes d'immunoglobulines. Ces gènes sont à l'origine même de l'apparition de l'immunité, mécanisme nécessaire à la survie de l'Homme. Ils sont apparus dans le génome il y a plus de 500 millions d'années ! Ils jouent un rôle tout aussi important que l'endocytosymbiose des mitochondries et des chloroplastes dans la cellule eucaryote.

Détail du contenu d'un transposon



Récemment, des scientifiques ont pensé que si ces transposons n'étaient plus actifs dans le génome humain, on pouvait peut-être faire renaitre artificiellement un transposon fonctionnel, appelé en conséquence « ***Sleeping beauty*** ». Il est flanqué de **séquences IR/DR** (équivalents des séquences ITR), et code pour un **domaine de liaison à l'ADN** ainsi qu'un **signal de localisation nucléaire**, en plus d'un **domaine catalytique**…

Cette transposase artificielle peut ainsi être utilisée pour manipuler le génome. Elle permet par exemple d'intégrer des gènes par un vecteur non viral (plasmide) dans le cadre d'une thérapie génique. Cela permet aussi de faire de la mutagénèse intentionnelle non aléatoire et d'observer si le fait de bousiller ce gène provoque l'apparition d'un phénotype.

Les transposons du génome humain sont en fait des vestiges de transposons autrefois fonctionnels, qui ont accumulé un nombre de mutations tel que l'enzyme n'est plus fonctionnelle.
C'est un peu comme un volcan éteint dans les Vosges. Mais malgré tout ça, on peut se demander s'il existe encore dans le génome une activité transposase ? Et bien oui ! Il existe des enzymes qui ont une activité transposase-like. Elles ne sont pas codées par des transposons, mais elles ont tellement évolué que leur fonction reprend le mécanisme de la transposase. On a ainsi montré que le génome de certains rhabdomyosarcomes (tumeurs malignes des muscles striés) se trouve remanié par l'activité de ces enzymes transposase-like, qui ne sont pas des transposons ! Au final, les transposons humains donnent des transposases tellement mutées qu'elles ont perdu leur activité, et inversement, certains gènes ont tellement évolué qu'on ne leur reconnait plus une structure de transposon à proprement parler, même s'ils ont pu l'être un jour, mais qui ont gardé une activité transposase-like.

ii. Les ADN rétrotransposons

Ce sont des éléments mobiles via un intermédiaire ARN : un élément est transcrit en ARN puis il est transformé en cDNA et s'insère dans le génome. C'est un mécanisme « copier/coller ».

Cela explique qu'ils représentent la majorité des séquences répétées du génome humain. On en compte deux catégories :

⇢ Les rétrotransposons non-autonomes

⇢ Les rétrotransposons autonomes non LTR

a) Les rétrotransposons non-autonomes (SINEs = Alu, SVA, et rétro-pseudogènes)

■ Les **SINEs** *(« Short INterspersed Elements »)* sont très nombreux et un peu parasite mais ils peuvent avoir des fonctions. Ce sont des éléments de petite taille dispersés (quelques centaines de bases). On peut individualiser deux types de séquences au sein des SINE :

⇨ La séquence **Alu** : **dérive de l'ARN non codant 7SL** dupliqué et remanié (élément de la ribonucléoprotéine SRP responsable du signal de reconnaissance peptidique = signal que la protéine doit être sécrétée). **La séquence Alu code donc pour un petit ARN**, qui par conséquent possède un **promoteur interne**, reconnu par la Polymérase III. Dans le génome humain, on compte **plus d'un million de copies** de la séquence Alu (**10%** du génome).

Une séquence Alu est en fait une séquence consensus de **280 pb**. On peut décrire plus de **200 familles.** Elle est apparue il y a plus de 65 millions d'années. Les séquences Alu ont envahi le génome (il y en aurait une toutes les 5kb). Elles sont spécifiques des primates, et certaines familles sont ethnie-dépendantes. **Elles ne codent rien** (pas de phase ouverte de lecture).

Elles sont encadrées par des **séquences répétées directes** (c'est-à-dire qu'en s'insérant dans le génome, la séquence dans laquelle elles s'insèrent est automatiquement dupliquée). Les séquences Alu sont toujours actives et continuent de se déplacer. On estime un déplacement toutes les 20 naissances vivantes. Il y aurait **850 copies actives**. Elles peuvent s'insérer dans les exons des gènes et donner des maladies par exemple en cassant les exons. La neurofibromatose de type 1 est par exemple expliquée par l'insertion d'une séquence Alu dans un intron.

⚠ **Alu est dépendante d'une rétrotranscription puisque c'est un ARN**, donc il doit être rétrotranscrit pour pouvoir réintégrer l'ADN ! On comprend donc pourquoi ces séquences sont dites non-autonomes.

⇨ La **SVA** : cousine de la séquence Alu, plus petite (< 2kb), environ 2 700 copies, apparue il y a 25 millions d'années à partir d'ARN, dépendant de la polymérase II, encadrée par des séquences répétées directes, apparaissant toutes les 1000 naissances vivantes.

■ Les **rétro-pseudogènes** sont aussi des rétrotransposons non autonome.

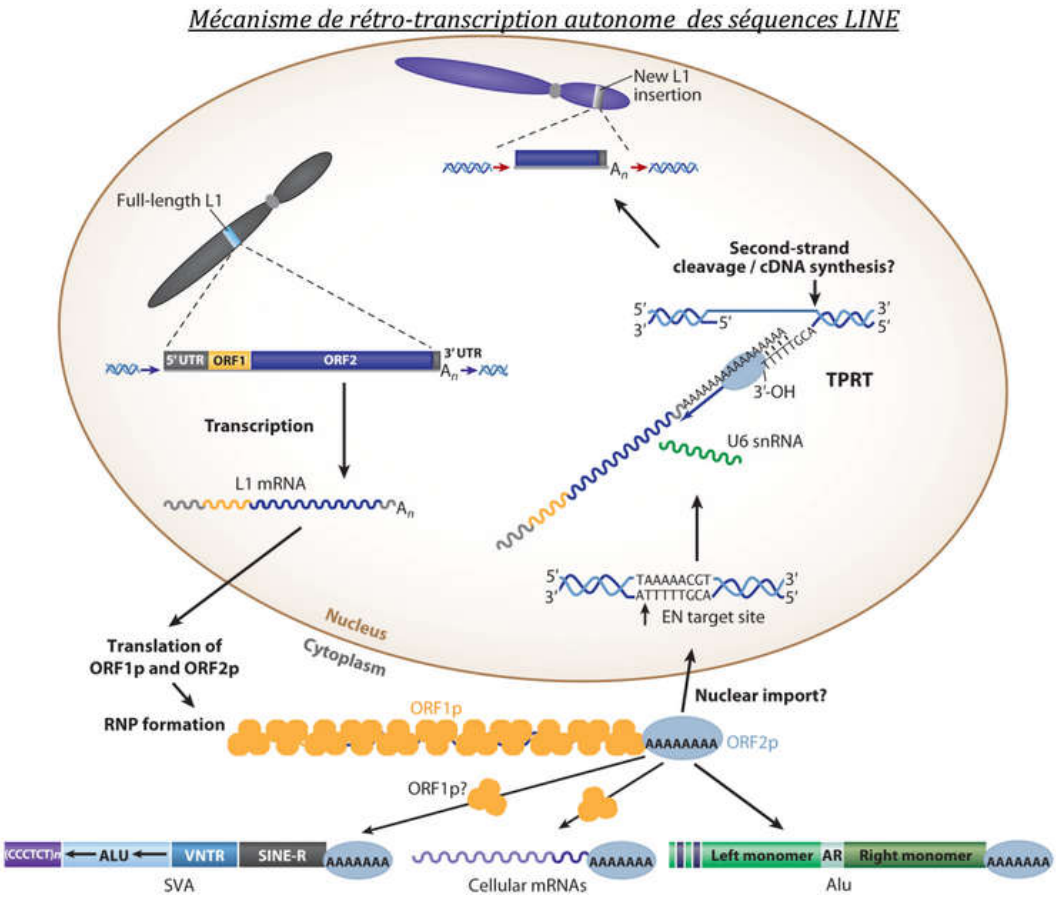
b) Les rétrotransposons autonomes non-LTR

☞« Non-LTR » sous-entend que ce ne sont pas des rétrovirus car ils ne possèdent pas de séquence *long terminal repeat*.

Ce sont les LINE *(Long Interspersed Element)*, des séquences longues et **autonomes** de plusieurs milliers de pb correspondant à **21%** du génome humain (1/5ème). Au sein des séquences LINE, on peut individualiser les séquences L1.

Elles sont au nombre de **500 000 copies**, la plupart incomplètes et tronquées en 5'. Les L1 à elles seules représentent près de **17%** du génome. On compte 40 à 50 éléments actifs dont 6 «superactivés» (les hotline). L'élément complet fait à peu près **6,1 kb** (c'est grand). Elles possèdent un **promoteur intrinsèque** le même que celui de la Polymérase II (le même que celui qui transcrit les gènes codants pour les protéines).

Les LINE sont **spécifiques des mammifères**, et seraient actives depuis 150 millions d'années, ce qui est peu par rapport à l'histoire de l'humanité. **Elles codent des phases ouvertes de lecture** donc des protéines : trois protéines ORF (ORF 1 et 2 communes aux mammifères, ORF 2 codant une *reverse transcriptase*, et ORF 0 antisens spécifique des primates). Ces séquences sont **encadrées par des séquences répétées directes** (car toute séquence issue d'une rétrotransposition est obligatoirement entourée de séquences répétées directes, cela fait partie du mécanisme d'insertion). Il apparaitrait une séquence LINE toutes les 100 naissances vivantes.



Dans le génome, la séquence LINE complète (donc active) est transcrite en ARNm, qui va être maturé et passer dans le cytoplasme. Ici, les protéines qu'il va traduire vont s'associer à lui pour former une **ribonucléoprotéine**. Cet ARN, encadré par ses protéines, va retourner dans le noyau. Ici, une endonucléase va couper l'ADN, l'ARN va s'hybrider à l'ADN, générant une extrémité 3'OH libre à partir de laquelle la *reverse transcriptase* codée par la séquence LINE va transformer l'ARN en *cDNA*. On va ensuite le recopier puis réparer l'ADN pour insérer le rétrotransposon. C'est ce mécanisme de rétrotranscriptase autonome des LINE qui prendra en charge les séquences ALU et SVA et les rétro-ARN cellulaires, pour les intégrer dans le génome.
(Il existe deux reverse transcriptase dans le génome : celle codée par les éléments LINE et celle de la télomérase.)

<u>Mobilité des séquences LINE en somatique et en germinal</u> :

• La première mobilité des séquences LINE est dans le zygote. Si elles sont transcrites, elles peuvent se rétro-intégrer. Elles créent des rétro-transpositions *de novo* qui amènent des maladies.

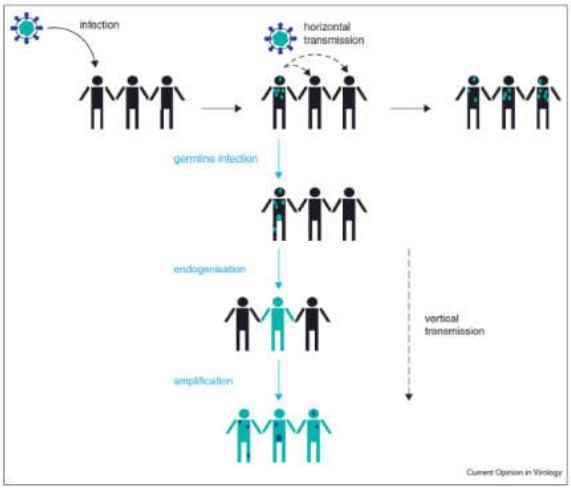
• Leur mobilité a également été démontrée dans le cerveau, pour expliquer la variabilité neuronale. On savait que les lymphocytes remaniaient leur génome pendant leur différenciation. On a pensé que les neurones pouvaient faire la même chose, et en observant l'expression des gènes au cours de la différenciation neuronale, on a constaté qu'il y avait une séquence très surexprimée, à savoir la séquence LINE. **On a donc montré que les LINE étaient capables de mobilité=de rétrotransposition dans les neurones** (notamment dans l'hippocampe)**.** La répression transcriptionnelle des LINEs va être levée dans les neurones, permettant la transcription de la séquence LINE et donc sa rétro-insertion autonome. Elle peut s'insérer dans un endroit neutre, dans un endroit provoquant la mort neuronale, ou encore dans un endroit stratégique qui participe à la diversité neuronale.

• Aujourd'hui, il est primordial de comprendre comment sont régulées ces séquences. Récemment, on a utilisé l'outil **CRISPR/Cas9** pour inhiber les promoteurs des séquences LINE afin de démontrer leur fonction et la régulation de leur expression. Ces séquences sont par exemple très mobiles en cancérologie. Une tumeur qui en apparence ne va pas accumuler beaucoup de mutation va en fait mobiliser ses rétrotransposons (ceci va permettre l'apparition de néo-antigènes et donc le ciblage par immunothérapie). Il est à noter que le séquençage par NGS des rétrotransposons impose des protocoles particuliers.

Globalement, on a augmenté notre génome de 8 millions de pb grâce à ces séquences au cours des 6 derniers millions d'années (+2 000 L1, +7 000 Alu et +1 000 SVA). On a identifié 92 maladies héréditaires pour lesquelles la mutation causale est une insertion d'une de ces trois séquences. Ce sont des éléments majeurs du polymorphisme du génome humain et elles constituent **un élément majeur de l'évolution des génomes.** Ainsi, **les éléments transposables contribuent finalement au polymorphisme du génome humain** (par leur présence ou leur absence) et font partie des variants structuraux. Elles ont un impact sur l'expression et la régulation des gènes, car elles arrivent avec des promoteurs, des signaux, etc.

<u>iii. Les HERV</u>

L'acronyme HERV correspond à *Human Endogenous RetroVirus.* Ce sont des rétrovirus endogènes humains, <u>acquis</u> : leur ARN a été transformé en cDNA puis s'est intégré au génome. Ils représentent **8%** du génome. **Ce sont des vestiges d'infections rétrovirales anciennes dans les gamètes**. Les rétrovirus peuvent en effet infecter les gamètes en s'intégrant sous une forme provirale.



Ici, la transmission d'un rétrovirus peut être horizontale, comme le VIH, mais ce virus n'infecte pas les gamètes donc il ne peut pas être endogène. Mais la transmission peut aussi se faire de façon verticale : le virus peut infecter les gamètes et s'intégrer dans le génome sous forme d'un pro-virus. Il va donc être transmis verticalement et devenir endogène. Ainsi, **les HERV du génome humain sont les témoins d'infections rétrovirales anciennes <u>au niveau des gamètes</u>.**

On a montré que le placenta utilisait des HERV. En effet, dans ce tissu, le passage de la forme cytotrophoblastique à la forme syncytiotrophoblastique, c'est à dire la fusion des membranes, est liée à l'expression d'un HERV qui va coder une protéine d'enveloppe, la syncytine.

De plus, il a récemment été démontré que le gène **Arc permettant le bourgeonnement synaptique n'est autre qu'un rétrovirus.**

De très nombreux rétrovirus sont utilisés et nécessaire à notre existence. Cependant, il est attendu qu'une partie de l'auto-immunité de notre génome soit due aux rétrovirus.

↳ Sommes-nous les seuls à avoir des HERV ?

Non, toute espèce infectée au niveau des gamètes par un rétrovirus va en avoir dans sa lignée. Il y a quelques années, on a pensé à faire de la **xénogreffe**, en raison de la pénurie d'organe. Or, on s'est dit que l'on pourrait greffer le cœur de cochon sur les humains.

Néanmoins, le porc possède ses propres rétrovirus endogènes, les PERV *(porcine endogenous retrovirus)*, ce dernier a appris à les maitriser. Hélas, l'homme ne serait peut-être pas capable de domestiquer les rétrovirus de cochon. La xénogreffe a donc été arrêtée.

Cependant, en 2015, on a montré qu'on était capables d'infecter des cellules humaines par des PERV, donc on a proposé d'utiliser CRISPR/Cas9 pour inactiver les PERV en ciblant la *reverse transcriptase*. En cassant puis en réparant la séquence, on accumule des mutations qui permettent de l'inactiver. On a réussi à le faire dans un modèle cellulaire qui a permis par transfert nucléaire d'obtenir des cochons vivants !

Ainsi, on peut imaginer qu'un jour on pourra faire des batteries d'élevage de cochons PERV inactivés, et humanisés pour être immunocompatibles, permettant ensuite de récupérer leurs cœurs dans le cadre de xénogreffe chez l'Homme.

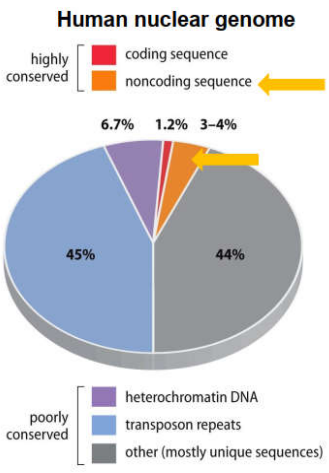
iv. Les séquences répétées en tandem

Ce sont des séquences répétées le plus souvent concentrés dans des régions chromosomiques. Selon la taille de **l'unité de répétition** et la taille de la **région concernée**, on distingue :

- ✓ ADN satellite (centromères)
- ✓ ADN minisatellite (juxta-télomérique)
- ✓ ADN microsatellite (partout dans le génome)
- ✓ ADN télomérique ($TTAGGG_n$).

v. Les séquences non-codantes hautement conservées

Entre les différentes espèces, si l'on compare les séquences, on trouve des gènes communs, mais aussi des séquences non codantes très conservées, presque plus que ne le sont les exons. Elles ont donc probablement un rôle dans l'expression du génome que l'on n'a pas encore découvert (il y a là peut-être des *enhancer* ou des *silencer*). Elles représenteraient **3 à 4%** du génome nucléaire.



IV. Le polymorphisme du génome humain

A. Diversité du génome

Le concept de **polymorphisme génétique** (du grec « poly » = plusieurs et « morphe » = forme) désigne la coexistence de plusieurs allèles pour un gène ou un locus donné, dans une population donnée.

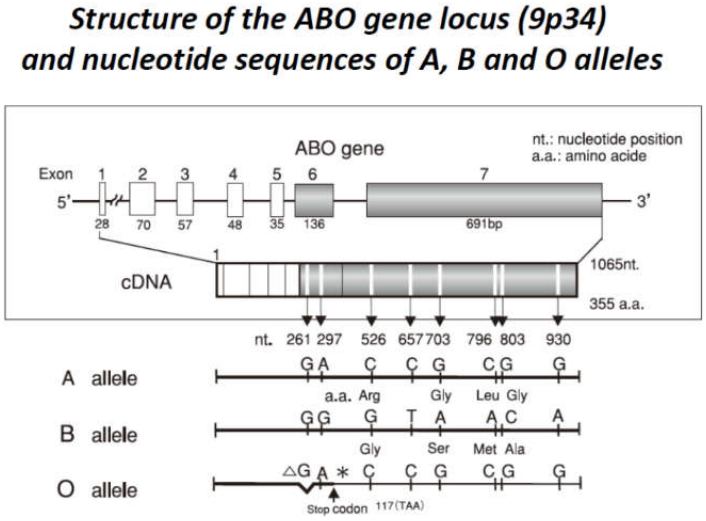
Ces polymorphismes peuvent être explorés :

→ Par analyse des caractères (**polymorphismes phénotypiques**) : la couleur des yeux en est un exemple, c'est un polymorphisme dans un intron.
→ Par analyse du produit protéique du gène (**polymorphismes protéiniques**) : les groupes sanguins et les HLA en sont deux exemples.
→ Par analyse des chromosomes (**polymorphismes chromosomiques**)
→ Par analyse de l'ADN (**polymorphismes nucléotidiques**)
NB : Les polymorphismes chromosomiques et nucléotidiques n'ont pas besoin de siéger dans des séquences codantes pour être détectés.

i. Deux exemples de polymorphisme protéiques

■ Les groupes sanguins

On doit la découverte du système ABO à Karl Landsteiner en 1900 et leur base moléculaire a été expliquée 90 ans après par Fumi-ichiro Yamamoto. Il a montré qu'une enzyme va maturer la **substance H** en additionnant des sucres, et qu'il existe **un seul gène ABO** qui code pour cette enzyme (5 exons non-codants, 2 exons codants). Ce gène présente des polymorphismes selon lesquels l'enzyme codée va maturer la substance H en antigène A ou B. Si l'on est O, cela veut dire que le gène est délété d'une guanine et ne peut pas produire l'enzyme, donc on reste avec la substance H non maturée. On comprend donc pourquoi on ne peut être O que si l'on est homozygote pour ce polymorphisme



■ Les HLA et le don de moelle osseuse

C'est Jean Dausset qui a découvert le complexe majeur d'histocompatibilité en 1958. On rappelle que **la probabilité de trouver un donneur et un receveur HLA-compatibles pour le don de moelle osseuse est de 1 pour 1 million**. On prend donc conscience qu'il s'agit là d'un polymorphisme très important, et qu'il est très important de donner sa moelle osseuse. Cela consiste à devenir un « Veilleur de Vie » à l'agence de la Biomédecine : on réalise une prise de sang pour déterminer notre HLA, puis on est génotypé et enregistré dans une banque. On sera contacté si un jour un patient en attente de greffe de moelle osseuse est HLA-compatible. Donner sa moelle, c'est d'abord donner son sang, avant de pouvoir sauver des vies.

ii. Les polymorphismes nucléotidiques

✓ Le premier polymorphisme du génome humain était un RFLP, découvert en 1978 (variation de séquence dans le site de coupure d'une enzyme de restriction). C'est à partir de là qu'on s'est dit que le génome était polymorphe.

✓ Ensuite en 1985, on découvre les **minisatellites** et on invente le concept **d'empreinte génétique à des fins judiciaires.** Ces minisatellites étaient utilisés par la police scientifique pour identifier un coupable à partir d'une liste de suspect en comparant leurs séquences, mais ce n'était pas assez précis, alors on a été chercher les **microsatellites**. Avec la PCR, on s'est aperçu qu'en amplifiant une même séquence, on obtenait des fragments de tailles variables.

Cette variation était due aux microsatellites, petites séquences répétées en tandem (on les appelle donc des **STR** = *short tandem repeat*), de toute petite unité (1 à quelques nucléotides) et qui s'étend sur une centaine de nucléotides. Elles sont régulièrement réparties dans tout le génome (1 STR toutes les 30 kb d'euchromatine) et sont des **polymorphismes** multi-alléliques.

Les STR dont l'unité de répétition n'est pas multiple de 3 sont cependant absents des exons codants car ils décalent le cadre de lecture. On peut avoir des STR mono-nucléotidiques (A_n, T_n), di-nucléotidiques (CA_n), tri-nucléotidiques (AAT_n), etc. Ce sont ces microsatellites qui constituent l'empreinte génétique aujourd'hui utilisés par la police scientifique.
Ces séquences sont **spécifiques d'un individu** (la probabilité de trouver deux individus identiques est de l'ordre de 1/1025).

Le fichier FNAEG (fichier national des empreintes génétiques) rassemble plus de **3 millions** d'empreintes génétiques (uniquement non codante en Europe).
Aux Etats-Unis, un criminel a déjà été retrouvé à cause d'un test génétique fait par sa sœur afin de déterminer son ethnie.
Suite à l'accumulation de données génétiques dues à ce genre de test, un papier de science a montré que l'on pouvait parfaitement retrouver presque tout le monde à partir du moment où on a son génome. En effet, un polymorphisme non codant peut renseigner sur ce qui est autour de lui pour peu qu'il y ait des déséquilibres de liaisons.

En France, l'identification d'une personne par ses empreintes génétique ne peut être recherchée que dans le cadre d'une **enquête policière ou d'une procédure judiciaire**, à des **fins médicales**, de **recherche scientifique** ou aux fins d'établir, lorsqu'elle est inconnue, l'identité de **personnes décédées**. C'est interdit de le faire par jeu.

✓ Les substitutions nucléotidiques (*SNPs : Single Nucleotide Polymorphisms*) sont des **variations de séquence par changement de base.** On sait que les **transitions** (purine⇔purine ou pyrimidine⇔pyrimidine) sont beaucoup plus fréquentes que les **transversions** (purine⇔pyrimidine). Un polymorphisme de type SNP veut que l'allèle le plus rare, dans un système à deux allèles, ait au moins une fréquence de 1%.

Ces polymorphismes sont donc le plus souvent bi-alléliques (on rappelle qu'en génétique épidémiologique, dans un système bi-allélique, le terme de polymorphisme est réservé à l'allèle le plus rare qui a au moins 1% de fréquence). Ces SNP peuvent définir des **haplotypes**, c'est-à-dire la succession des allèles d'un gène ou d'un locus sur une région chromosomique de petite taille. (*Le projet HapMap avait pour but de répertorier tous les haplotypes.)*

Du fait de l'absence de recombinaison, les associations d'allèles de plusieurs polymorphismes très proches sont transmis généralement en bloc, comme un seul marqueur. Les SNP sont référencés : rs (référence SNP) suivi de [n° de localisation et fréquence du polymorphisme]. On en trouve partout dans le génome. On peut classer ces SNP en fonction de la MAF (séquence de l'allèle le plus rare = minor allele fequency).

Suite au séquençage du premier génome, le projet 1000 génomes a été lancé. Le but était de séquencer avec grande profondeur des donneurs volontaires dans des populations particulières. Pour la première fois, on a eu une vision de ce qu'est le polymorphisme. Sur 2500 individus, on a identifié 85 millions de SNP. Certains polymorphismes étaient fréquents, d'autres moins, voir spécifiques d'un seul individu.

Les SNPs représentent 85 millions de variants répartis régulièrement sur l'ensemble du génome. La fréquence allélique de l'allèle le plus rare (MAF pour Minor Allele Frequency) se distingue en :

- MAF > 5% : variant commun retrouvé chez plus de 5% de la population -> 8 millions
- 0.5 < MAF < 5% -> 12 millions
- MAF < 0.5% : SNV -> 65 millions

Les fréquences sont analysables individuellement (PCR) ou globalement (puces à ADN, NGS).

✓ La variation structurale (SV) : en fonction de la taille de la SV, on distingue :
⇢ Les **INDELs**, qui sont de petites insertion/délétion de moins de 50 nucléotides, sont très nombreux.
⇢ Les autres variations de structures qui correspondent à des variations de plus de 50 nucléotides sont appelés vrais variants structuraux. Ces derniers sont classés en groupes (CNVs, larges deletions, inversions, MEI)

✓ Les CNV (*copy number variant*) sont les **variations du nombre de copies de segment d'ADN d'une taille allant de 50 à plus de 3 millions de nucléotides et pouvant inclure des gènes.** Ces polymorphismes peuvent être **bi- ou multi-alléliques**, puisque ce sont des répétitions et non des changements de bases. On en a répertorié plus de 10 000 analysables par CGH Array et par NGS.

Aujourd'hui un génome typique présente 4 à 5 millions de différences par rapport à un génome de référence. La plupart sont des SNV ou des INDELs. D'autres, moins nombreux mais plus grands donc ayant plus d'impact sont des variants structuraux. On est aujourd'hui capable de faire des médianes de variation du génome.

Quand on séquence tous les exons d'un génome, on est confronté à **20 000 variations**. On en compte plus de 1.6 millions dans les introns.

⇨ **Le génome est donc extrêmement variable, polymorphe, et c'est cette diversité qui est à la base même de sa complexité.**

La base ExAC répertorie toutes les variations trouvées dans les exons et la base gnomAD rassemble toutes les variations obtenues dans 123 000 exomes et dans 15 000 génomes.

B. Nomenclature et classification

On essaie de retirer le mot « **mutation** » du langage scientifique.

En génétique médicale :

- Une mutation est une variation génomique associé à un effet pathogène.
- Un polymorphisme serait une variation génomique qui *a priori* ne serait pas associée à un effet pathogène.

En génétique des populations :

- Un polymorphisme est une variation génétique présente à une fréquence de plus de 1% de la population générale.

Mais attention, ici on ne prédit pas de la pathogénicité du variant ! Certaines variations sont présentes à plus de 1% et sont pathogènes, comme la ΔF508 de la mucoviscidose.

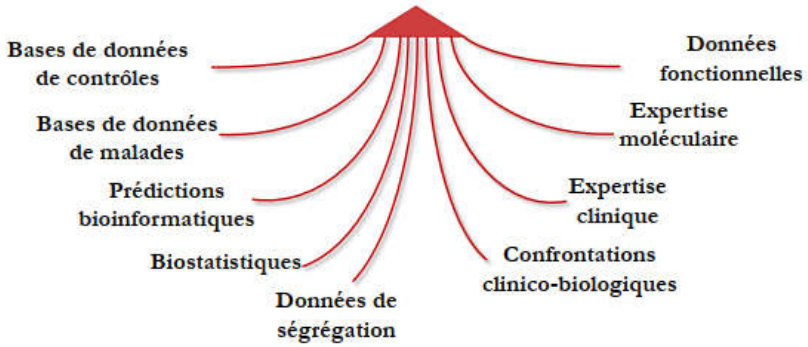
En médecine génomique, on propose donc de remplacer le terme « ~~mutation~~ » par celui de « **variant** » ou « **variation de séquence** », ceux-ci ayant un sens plus neutre, qui ne préjuge pas de l'existence ou non d'un effet pathogène. Au passage, des variants pathogènes peuvent être présents chez un patient sans pour autant être responsable de sa pathologie.

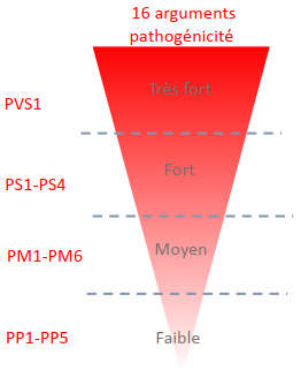
Il faut bien distinguer la présence de variants pathogènes, impliqués dans la pathologie du patient, et les variants considérés comme pathogènes en ce qui concerne l'effet sur la fonctionnalité du gène mais qui ne sont pas responsables de la maladie (exemple du gène ABO : quand on est O on ne métabolise pas la substance H et on n'est pas malade pour autant). On ne parle plus de SNP mais bien de SNV et de CNV.

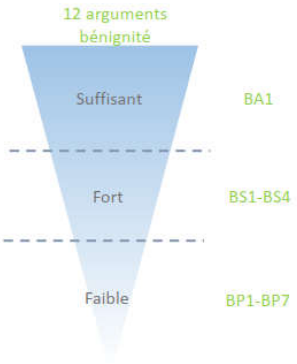
Le défi de la génomique médicale est l'interprétation des conséquences des variations de séquence nucléotidique du génome humain. On a donc créé une classification des variants, en fonction de leur probabilité d'être bénin ou malin.

CLASSES	DESCRIPTION	
1	BENIN BENIGN (B)	
2	PROBABLEMENT BENIN LIKELY BENIGN (LB)	La classe 2 repose sur une confiance de 90% que le variant soit bénin.
3	SIGNIFICATION INCONNUE UNCERTAIN SIGNIFICANCE (VUS)	On ne sait pas vraiment catégoriser ce variant.
4	PROBABLEMENT PATHOGENE LIKELY PATHOGENIC (LP)	La classe 4 repose sur une probabilité de 90% : on a 90% de confiance quant au fait que le variant soit pathogène.
5	PATHOGENE PATHOGENIC (P)	

On dispose **de 28 critères** pour interpréter ces variants avec des bases de données de population contrôle ou de patients, biostatistiques et prédictions informatiques, etc.







16 critères peuvent classer un variant comme pathogène, 12 comme bénin, et la situation de contradiction correspond au VUS. On classe les critères eux-mêmes en fonction de leur force. A partir de ça, on utilise des algorithmes qui permettent de classer les variants.

La diapo à connaître :

Caractéristiques du génome humain

~ 20 000 protein-coding genes qui occupent 1.2 % du génome
Taille très variable : 1 kbp pour le gène *HBB* à plus de 2 Mbp pour le gène *DMD*
Morcelés en exons (~ 220 000) séparés par des introns dont le nombre est très variable d'un gène à l'autre
> 95% des gènes soumis à épissage alternatif

~ 25 000 gènes produisant des ARN non codants
nc RNA classiques : tRNA, rRNA
Small ncRNA (< 200 nt) : snRNA (small nuclear RNA), siRNA (small interfering RNA), piRNA (piwi-interacting RNA), miRNA (microRNA), snoRNA (small nucleolar RNA).
Long ncRNA (> 200 nt) : intronic, exonic, overlapping and intergenic

~ 50% du génome est constitué de séquences répétées représentées en majorité par des éléments transposables (SINEs, LINEs, HERV)

~ 4 à 5 millions de variations comparativement à un génome de référence
En majorité des SNPs (~ 3.5 millions) et des INDELS (~ 500 000)
2 000 à 2 500 variations de structure > 50 bp affectant plus de 20 millions bp

~ 80% du génome humain est fonctionnel et plus de 60% est transcrit en ARN

Fiche récapitulative : Structure, organisation, dynamique et polymorphisme du génome humain

I. Encadrement réglementaire des activités de génétique en France

→ convention d'Oviedo : → non-discrimination
→ tests génétiques prédictifs avec consentement
→ interventions sur le génome humain réglementées
→ non sélection du sexe
→ primauté de l'être humain
→ lois de Bioéthique 1994, révisées en 2004 puis 2011 et 2018

II. Génétique et séquençage

A. Généralités

→ Génome = ensemble de l'information héréditaire d'un organisme présente en totalité dans chaque cellule. Support matériel de l'information génétique : ADN et ARN.
→Valeur C = taille d'un génome, variable en fonction des espèces : ne corrèle en rien avec la complexité d'un organisme : l'être humain n'a pas le plus gros génome (3000 Mb)

B. Le génome nucléaire et mitochondrial

→ Génome humain : 3000 Mb, donc 6000 Mb dans le noyau (diploïdie). Le génome nucléaire n'est contenu qu'une fois par cellule, et est composé de 46 chromosomes linéaires.
→ Génome mitochondrial : plus petit, circulaire, centaines ou milliers dans chaque cellule.

C. L'histoire du séquençage

→ Premier en 1968 (12 bases séquencées).
→ Deuxième par la méthode de Maxam et Gilbert concernant le VHB (1er virus séquencé en entier).
→ Méthode de Sanger : 1er séquençage génome humain en 1989 => 2004 : génome de référence.
→ Révolution du NGS : séquençage parallèle massif du génome de son inventeur en 4 mois.
→Aujourd'hui : séquençage beaucoup plus rapide car alignement automatique par rapport au génome de référence. En 2018, on peut séquencer un génome humain complet en moins d'un jour pour moins de 1 000 dollars (RWGS).

D. L'évolution technologique du séquençage

→1964 : ordinateurs IBM 360 mis sur le marché.
→ Big Data : séquencer n'est plus un problème, la difficulté est le stockage et l'interprétation.
→ Diagnostic : on utilise des approches pangénomiques avec les microarrays. Différentes approches.

III. Les caractéristiques du génome humain

A. Les gènes codant pour des protéines

→ 6.10^6bp (diploïdie), 60 000 gènes, 1,2% (un peu moins 20 000 gènes) codent pour des protéines.
→ La taille des gènes est très variable. La plupart sont morcelés en exons (>220 000 dans l'exome) séparés par des introns dont le nombre est variable. Tous (presque) les gènes sont soumis à transcription + épissage + maturation 3' alternatifs.
→ Gènes qui codent pour des protéines : transcrits en ARNm coiffés, épissés et maturés sans structure particulière. C'est la protéine qui acquiert une structure spécifique (=> fonction).

B. Les ARN non-codants

→ Gènes produisant ARN (23 000) > gènes produisant protéines (20 000) : la fonction de ces ARN non-codants (pas de traduction) va dépendre de leur structure.
→Petits ARN non-codants (7 500, <200 bases) : bien structurés
→Grands ARN non-codants (15 800) : pas encore de structure déterminée.
→On distingue parmi les ARN non-codants : ncRNA (classiques), sRNA (petits), lcnRNA (grands)

C. Les pseudogènes

→15 000, en majorité des rétro-pseudogènes (transcrit => ADNc). Ne codent pour rien mais peuvent s'exprimer sous la forme de transcrits : conséquences fonctionnelles importantes sur gène codant.

D. Les séquences répétées

→50% du génome, en majorité des éléments transposables, rôle majeur dans le génome.

i. Les ADN transposons

→ 3 génomes, plus fonctionnels depuis plus de 40 millions d'années chez l'homme.
→ Codaient pour la transposase : enzyme qui reconnait les séquences répétées inversées ITR et qui coupe à ce niveau pour l'intégrer ailleurs dans le génome. Le transposon est mobile.
→ Certains gènes : dérivés de transposons et codent pour des protéines de l'immunité RAG1 et RAG2 = origine même immunité.
→ Scientifiques : possible de faire renaitre artificiellement un transposon fonctionnel flanqué de séquences IR/DR, code pour un domaine de liaison à l'ADN, un signal de localisation nucléaire et un domaine catalytique => transposase artificielle utilisée pour manipuler le génome

ii. Les ADN rétrotransposons (majorité des séquences répétées du génome humain)

a) Rétrotransposons non-autonomes

→SINE = très nombreux, deux types : séquence Alu (ne code rien) et SVA ;
→rétro-pseudogènes

b) Rétrotransposons autonomes non-LTR

→ LINE = séquences longues et autonomes (21% génome) dont L1, spécifiques des mammifères, codent des phases ouvertes de lecture.
→LINE => ARNm : maturation + passage dans cytoplasme => association avec les protéines qu'il traduit => ribonucléoprotéine : retour dans noyau => endonucléase : coupe ADN + hybridation
ARN-ADN => reverse transcriptase : ARN devient ADNc => réparation ADN pour insérer le rétrotransposon
→LINE : capables de mobilité => rétro-transposition dans zygotes et neurones

iii. Les HERV *(-K, -W, -R)* : rétrovirus endogènes acquis, 8 génomes = vestiges infections rétrovirales anciennes dans gamètes. Transmission virale verticale => pro-virus dans gamète : endogène

iv. Les séquences répétées en tandem

→ hétérochromatine composée de séquences répétées différentes
→ ADN satellite, ADN minisatellite, ADN microsatellite, ADN télomérique

v. Les séquences répétées hautement conservées

→ surement rôle dans expression génome, 3-4%

IV. Le polymorphisme du génome humain

A. Diversité du génome

→Polymorphisme génétique = coexistence plusieurs allèles pour un gène ou un locus donné => polymorphismes phénotypiques, protéiniques, chromosomiques, nucléotidiques

i. Deux exemples de polymorphismes protéiques

a) groupes sanguins : un seul gène ABO codant pour enzyme qui mature la substance H en Ag A ou B en fonction des polymorphismes. Absence d'enzyme => O.
b) HLA et le don de moelle osseuse : polymorphisme très important

ii. Les polymorphismes nucléotidiques

→ premier découvert = RFLP, puis découverte des minisatellites et microsatellites (STR) spécifiques d'un individu ; substitutions nucléotidiques ; variation structurale (INDELS <50 nucléotides), CNV

B. Nomenclature et classification

→Génome typique : 4-5 millions de différences avec génome de référence => polymorphe
→Génétique médicale : mutation = pathogène, polymorphisme = à priori non pathogène
→Génétique des populations : polymorphisme = variation génétique avec fréquence >1%
→Classification variants : 1=bénin ; 2=probablement bénin ; 3=VUS ; 4=probablement malin ; 5=malin
→ 28 critères dont la MAF (variant très fréquent surement bénin, de novo surement pathogène)