# CHAPITRE 1 – GÉNOMES ET ÉTUDES D'ASSOCIATION PANGÉNOMIQUES

## 1.1 Composition et structure du génome humain

Le génome humain diploïde est composé de 22 paires de chromosomes dits non sexuels nommés autosomes ainsi que d'une paire de chromosomes sexuels (gonosomes). Ces 46 chromosomes représentant 3,2 milliards de paires de bases (répartis sous forme de bases nucléiques adénines [A], guanines [G], cytosines [C] et thymines [T]) et environ 30 000 gènes, codant pour autant de protéines, sont retrouvés tout au long de cette séquence nucléotidique (Human Genome Project, 2018). D'autres éléments génomiques sont retrouvés avec un nombre de copies plus ou moins important tout au long du génome. De ces éléments, diverses séquences régulatrices (éléments régulateurs) sont réparties à travers le génome humain et certaines sont regroupées dans la grande classe des acides ribonucléiques non codants [**ncARN**]. Les sous-classes retrouvées chez la famille des ncARN sont présentées dans le tableau 1.

**Tableau 1** Classification des ARN non codants du génome humain [tiré de 5]

| Sous-classes d'ARN non codants (ncARN) | Fonctions et caractéristiques génomiques | Taille (Nucléotides [nt]) |
|---|---|---|
| Longs ARN non codants (lncARN) | Régulation épigénétique et génétique | > 200 nt |
| Petits ARN interférents (siARN) | Régulation génique, transposons et défense contre les pathogènes viraux | ~ 21-22 nt |
| Micro-ARN (miARN) | Associé à la protéine Argonaute et agit sur la régulation post-traductionnelle des gènes | ~ 22 nt |
| ARN interagissant avec Piwi (piARN) | Régulation des transposons et de la chromatine dans les cellules germinales | ~ 26-30 nt |
| ARN associé aux promoteurs | Rôle potentiel dans la régulation de l'expression génique | ~ 200-500 nt |
| Petits ARN nucléaires (snARN) | Guides pour la méthylation à l'ARNr et la pseudouridylation | ~ 60-45 nt |
| Petits ARN télomériques (tel-sARN) | Maintenance et stabilisation des télomères | ~ 24 nt |
| ARN associé au centrosome (crasiARN) | Guide dans les processus locaux de modification de la chromatine | ~ 34-42 nt |

Les éléments mobiles dispersés (transposons, rétrotransposons à séquence terminale longue répétée, longs éléments nucléaires intercalés et les petits éléments nucléaires intercalés) comptent pour environ 45% de l'acide désoxyribonucléique [**ADN**] humain. Finalement, des éléments mobiles répétés en tandem (minisatellites et microsatellites) (~6%) et des régions intergènes (25%) sont les derniers éléments qui composent le génome humain (Tableau 1). En plus de cette séquence de 3,2 milliards de paires de bases, une séquence circulaire additionnelle d'ADN mitochondrial de 16 569 paires de bases est retrouvée à l'intérieur des mitochondries [6].

**Tableau 2** Architecture du génome humain [tiré de 7]

| Éléments génomiques | Longueur (pb) | Nombre de copies | Fraction du génome |
|---|---|---|---|
| Gènes codants pour des protéines | 0,5 – 2 200 kpb | ~ 21 000 | ~ 40%† (2%††) |
| Gènes codants pour de longs ARN non-codants (lncARN) | 0,2 – 50 kpb | ~ 10 000 | ~ 15% |
| ARN ribosomal (ARNr) | 43 kpb | ~ 300 | 0,4% |
| Petits ARN nucléaire U2 (ARNsn U2) | 6,1 kpb | ~ 20 | < 0,0001% |
| Séquences simples (microsatellites, minisatellites) | 1 – 500 pb | Variable | ~ 6% |
| Transposons | 2 – 3 kpb | 300 000 | 3% |
| Rétrotransposons à LTR | 6 – 11 kpb | 440 000 | 8% |
| Longs éléments nucléaires intercalés (LINEs) | 6 – 8 kpb | 860 000 | 21% |
| Petits éléments nucléaires intercalés (SINEs) | 100 – 400 pb | 1 600 000 | 13% |
| Pseudogènes processés | Variable | ~ 12 500 | ~ 0,4% |
| Régions intergéniques | Variable | Inconnu | ~ 25% |

†En incluant les introns et les exons

††En incluant seulement les exons

La structure typique d'un gène eucaryote comprend une région amplificatrice, une région promotrice ainsi qu'une région transcriptionnelle. La région amplificatrice d'un gène comprend plusieurs amplificateurs (de tailles variables pouvant aller de 50 pb à 1 500 pb) qui ont pour fonction d'activer (ou bien de réprimer) le promoteur et la transcription du gène dans un type cellulaire particulier ou encore, à un moment précis, lors du développement où d'autres produits protéiques sont requis par les cellules par la liaison de protéines nommées facteurs de transcription [8].

En amont du site d'initiation de la transcription [**TSS**], à environ 35 paires de bases, se trouve le promoteur du gène, élément critique à la régulation de l'expression génique chez les eucaryotes [9]. Les promoteurs sont reconnus par les facteurs de transcription grâce à des motifs nucléotidiques distinctifs tels les boîtes TATA (ou également boîtes de Goldberg-Hogness) et les régions riches en CpG. D'autres motifs nucléotidiques existent également dans certains promoteurs eucaryotes [**Inr**, **BRE**] [10]. En aval du promoteur, une courte séquence d'acides nucléiques nommée région 5' non traduite) est retrouvée et y joue un rôle dans la régulation de l'expression génique en ajoutant un second contrôle dans la transcription d'ARN messager [**ARNm**] [11]. La région 5' non traduite peut s'avérer être un handicap pour le profil mutationnel dû à leur séquence relativement longue dans les gènes eucaryotes augmentant considérablement, dans le un tiers des cas, un allongement du produit protéique sur la séquence N-terminale [12]. La section codante (ou unité transcriptionnelle) est composée de séquences introniques et exoniques en alternance. Après épissage de l'unité transcriptionnelle, les introns sont retirés et les exons sont regroupés pour former les ARNm retrouvés dans le cytoplasme. Les gènes humains sont ceux qui possèdent le plus d'introns ainsi que les introns contenant le plus de nucléotides avec une moyenne de 3 413 nucléotides par intron (quoique la longueur la plus commune soit de 75 à 150 nucléotides dans la plupart des gènes) [13]. En regroupant les 23 paires de chromosomes, il est estimé que la longueur intronique totale est de

1 123 657 235 paires de bases, avec le chromosome 7 contenant à lui seul 20 537 introns, pour seulement 39 841 315 paires de bases d'exons dans le génome diploïde complet [14]. Ce déséquilibre introns-exons semble être bénéfique pour la cellule malgré le coût bioénergétique élevé à leur stabilité. En effet, les introns sont essentiels à la régulation positive de l'expression des gènes, dans le contrôle de la qualité de la dégradation des ARNm non-sens et d'augmenter l'efficacité de la sélection naturelle en augmentant les possibilités de recombinaisons [15].

### 1.2 Motifs génomiques d'intérêts en analyse génétique

De nombreux marqueurs génomiques sont utilisés en génétique des populations pour mettre en évidence certaines distinctions génétiques interindividuelles et des structures populationnelles. En estimant qu'un génome individuel comporte environ 4,1-5,0 millions de paires de bases qui diffèrent de celui de référence établie par le Projet Génome Humain (Human Genome Project, 2018), il est primordial d'être en mesure d'identifier ces différences plus en détail pour établir des associations et des liens avec des pathologies et des structures de populations [16]. Les microsatellites (séquences simples répétées) sont de courtes séquences nucléotidiques de 1 à 6 paires de bases répétées à plusieurs reprises et dispersées dans tout le génome et représentant environ 3% de ce dernier et hérités selon une transmission mendélienne [17, 18]. Plusieurs motifs de microsatellites semblent être plus récurrents dans le génome humain. Les unités monomériques répétées d'adénine sont les plus fréquentes tandis que les unités dimériques AC et AT sont les plus communes. Pour les répétitions de plus longue séquence, les trimères AAT et AAG sont les plus fréquents, les tétramères AAAT, AAAG, AAAC et AAGG sont les plus communs à l'échelle du génome tandis que les pentamères AAAAT et AAAAC sont retrouvés en majorité [17]. L'analyse de ces séquences simples répétées permet donc de faire ressortir certaines régions génomiques. En allongeant la séquence nucléotidique des microsatellites, on retrouve les répétitions en tandem en nombre variable (**VNTR**)

qui sont classés comme microsatellites et minisatellites [19, 20]. Ces séquences de 6 à plus de 50 nucléotides sont réparties à travers le génome mais spécialement plus rassemblées dans les régions télomériques des chromosomes. Leur grande nature polymorphique les rend utiles pour des analyses d'association et l'étude des traits monogéniques [21]. Une autre classe de marqueur génomique qui a gagné une certaine notoriété depuis les années 1980 est celle des polymorphismes d'un seul nucléotide (**SNP** – *Single Nucleotide Polymorphism*). Avec l'avènement de l'étude des traits complexes dans les années 1990, les SNP se sont montrés plus efficaces pour l'étude des maladies multifactorielles que les minisatellites et microsatellites qui sont plus propices à subir un taux de mutation plus élevé en raison de leur taille [22]. Les éléments simples répétés étaient utilisés pour identifier de potentielles régions du génome qui pouvait avoir un lien avec l'état de la pathologie mais les informations qu'ils révélaient ne renseignaient pas sur le phénotype à proprement parlé, un changement de marqueur s'avérait donc nécessaire [22]. À l'opposé, les SNP peuvent se retrouver dans les sections codantes et donc engendrer des conséquences phénotypiques qui seront, dans quelques cas, visibles. En date de 2015, 84,7 millions de SNP ont été répertoriés dans le génome humain et représentent 99,9% des différences entre les génomes soulignant leur importance ainsi que leur contribution pour de futures études génétiques et génomiques au côté d'autres variations structurales touchant plusieurs bases à la fois [16]. Un autre élément utilisé en génomique est le génome mitochondrial (mitogénome), une séquence d'ADN exclusivement transmise par la mère [23]. Le mitogénome ne subit pas de recombinaisons contrairement au génome et rend donc la tâche plus facile pour retracer des liens entre des individus [24]. Plus de 300 variants ont été recensés à ce jour dans le mitogénome soulevant donc la contribution de cette courte séquence de 16,5 kpb [25].

### 1.3 Études d'association pangénomiques

Historiquement, la première étude d'association pangénomique (**GWAS** – *Genome-wide Association Study*) remonte au début des années 2000 avec la publication d'un article scientifique par une équipe japonaise, dirigée par le généticien Yusuke Nakamura, portant sur l'infarctus du myocarde [26]. Cette première tentative d'étude à l'échelle globale du génome a marqué un grand pas en génétique moderne. La structure d'une GWAS est basée sur une approche expérimentale visant à identifier des associations entre un trait et des variants génétiques. Pour pouvoir effectuer une étude GWAS typique une population d'intérêt doit être recrutée. Sous des bases phénotypiques, un sous-groupe d'individus étant atteint d'une maladie sera classé comme patients (cas) et un second sous-groupe, composé d'individus sains, sera classé comme témoins. Ces deux sous-groupes d'individus seront alors échantillonnés pour obtenir une fraction de leur ADN dans l'optique d'un génotypage, c'est-à-dire, l'identification moléculaire des variations interindividuelles, appelées SNP devant posséder une fréquence allélique se situant aux alentours de 5% dans la population, entre les individus atteints et non atteints de la pathologie étudiée. À partir des informations récoltées lors du génotypage des individus, il est alors possible de cartographier les variations et, à l'aide d'outils statistiques, déterminer quel(s) SNP se démarquent plus particulièrement dans la maladie étudiée [27]. Pour qu'un SNP soit associé à une maladie, à l'échelle populationnelle, il doit être en déséquilibre de liaison avec le locus de susceptibilité, c'est-à-dire une forte association non-aléatoire entre les allèles du locus d'intérêt et le locus de susceptibilité (dans le cas d'une étude GWAS, on parle de tagSNP) [28]. Les études GWAS ont permis de mieux documenter plus d'une centaine de pathologies allant des maladies cardiovasculaires aux différents types de cancer et également de parfaire les connaissances aux sujets de gènes suspectés d'être impliqués dans certaines pathologies. En effet, seulement 9,5% des 348 premières études GWAS ont identifié des SNPs qui étaient localisés dans

les exons ou bien dans les régions non traduites 5' et 3' tandis qu'environ 45% des SNPs identifiés étaient retrouvés dans les introns et un autre 45% dans des régions associées à aucun gène [28]. Toutefois, malgré le fait que de nombreuses variations génétiques soient identifiées, les GWAS sont toujours peu prédictifs de l'héritabilité des traits complexes et l'expliquent seulement en partie (Tableau 3). Cela peut être expliqué par le fait que la sélection naturelle tend à faire disparaître les variants génétiques qui ont une trop grande influence sur la maladie [28]. D'autres auteurs mentionnent également le fait que les GWAS sont peu puissants face à l'identification de variants rares et ayant un effet moins prononcé sur le phénotype [29]. La contribution des études GWAS en génétique humaine a donc permis de développer des outils forts utiles comme les scores de risque polygénique (*Polygenic Risk Score* [**PRS**]), l'évaluation du déséquilibre de liaison, la randomisation mendélienne ou encore l'estimation de l'héritabilité pour des traits simples comme complexes [30].

**Tableau 3** Proportion de l'héritabilité expliquée par des loci associés par GWAS [adapté de 31]

| Pathologies | Nombre de loci associés | %Héritabilité expliqué |
|---|---|---|
| Diabète de type 2 | 76 | ~ 10% |
| Cancer du sein | 67 | ~ 14% |
| Arthrite rhumatoïde | 48 | ~ 51% |
| Troubles bipolaires | 56 | ~ 2% |
| Schizophrénie | 108 | ~ 3-7% |