Introduction

La connaissance du génome dans son integralité passe par son séquençage et sa cartographie. Cependant, la taille des génomes étant de plusieurs millions de bases, il est necessaire de coupler les approches de biologie moléculaire et d'informatique, pour pouvoir gérer les quantités immenses de données.

La génomique a ainsi vu le jour, et a permis de caractériser l'organisation du genome et de mettre en evidence l'existence de séquences codantes et de séquences non codantes. Ces dernieres sont présentes dans tous les génomes mais leur abondance est corrélée a la complexité des especes, et donc les mecanismes de régulation des genomes.

Le génome humain serait constitué de 5% seulement de séquences codantes qu'on nomme «gènes» . Des programmes bioinformatiques permettent l'analyse des sequences d'ADN et la prediction des genes. La traduction *in silico* de la séquence nucléotidique en séquence d'acides aminés permet éventuellement de confirmer ou proposer une fonction de la protéine codée par le gène.

1- Sequences codantes

Une séquence codante est une séquence ADN double brin transcrite en ARN et encadrée de séquences régulatrices. Dans un génome eucaryote, la séquence séquence transcrite est constituée d'exons et d'introns.

Les séquences régulatrices servent de signaux de début (promoteur, codon iniateur) et de fin de la transcription (codon stop)

Les gènes peuvent donc etre prédits par des algorithmes grace a ces caractéristiques communes aux séquences codantes.

Il faut distinguer deux types de gènes:

* Ceux qui codent pour des protéines et donc sont transcrits en ARN codant qui est l'ARNm, traduit ensuite en protéines

* Ceux qui codent des ARN non codants c'est a dire qui ne seront pas traduits en protéines.

Les ARN non codants sont impliqués dans l'expression du génome et sa régulation. Leur fonction repose sur leur structure et leur séquence:

* L'ARN ribosomique constitue 80% de l'ARN total dans les cellules eycaryotes.
* L' ARN de transfert constitue 15% des ARN dans une cellule eucaryote.

* Les petits ARN nucléaires et nucléolaires, présents dans le noyau des cellules eucaryotes, interviennent dans l'epissage et la maturation des ARNr
* Les ARN régulateurs sont impliqués dans les mécanismes d'interférences et régulent l'expression des gènes (transcription et traduction)

2- Sequences regulatrices

La transcription d'un géne est finement régulée dans le temps et l'espace. Un transcrit donné ne sera synthétisé que dans certains types cellulaires à une certaine étape du developpement de l'individu et/ou différents facteurs de l'environnement.

Les informations necesaires a cette régulation sont portées à la fois par des protéines régulatrices (éléments trans), et par des séquences nucléotidiques non transcrites (éléments cis) placées en amont de la séquence codante.
Un complexe de transcription initie la transcription en interagissant avec le promoteur. Le promoteur est situé en amont du +1 de la transcription. Des séquences particulières permettent de reconnaitre une séquence promotrice comme les boites TATA et les ilots CpG; une analyse bioinformatique peut révéler dans ces régions, des séquences nucléotidiques conservées et reconnues par les facteurs de transcription.
L'inhibition ou l'activation de la transcription dépend d'autres facteurs et séquences (enhancers et silencers)

Génome- Transcriptome - Protéome : Chez les eucaryotes contrairement aux procaryotes, le génome n'est pas corrélé au transcriptome ni le transcriptome au protéome. Un décalage important est noté entre le nombre de genes et le nombre de transcrits et de protéines produits par une cellule eucaryote.
Ce décalage révèle l'importance des mécanismes de régulation post-transcriptionnelles et post-traductionnelles dans le niveau d'expression final d'une protéine dans la cellule.
Le mécanisme d'épissage alternatif explique en partie le décalage entre le nombre de gènes et le nombre de transcrits et explique ainsi la présence de séquences introniques dans les gènes eucaryotes.

3- Sequences non codantes

3.1 Sequences répétées

a- Sequences répétées dispersées

Les elements transposables sont des sequences d'ADN capables d'etre transférées, au sein du genome, d'un site donneur a un site cible sous l'effet d'une enzyme specialisée. Les deux sequences ne presentent aucune homologie et donc le mecanisme d'insertion est une recombinaison illegitime

Les enzymes qui effectuent la transposition sont soit des transposases soit des integrases (dans le cas de retrovirus par exemple); ces deux enzymes sont des recombinases.

Elles reconnaissent et se lient a des sequences specifiques aux extremités du transposon, excisent l'ADN et lient.

Le système transposon/transposase peut etre utilisé comme outil pour la transgenese.

On distingue deux classes d'elements transposables:

Classe I ou retrotransposons a ARN

Classe II ou transposon a ADN

1- Transposons: transposent via un mode couper-coller, c'est-à-dire que le transposon est excisé de son site et intégré a un autre site sur le meme chromosome ou un autre chromosome. On parle de gene sauteur, car la sequence a changé de place

2- Retrotransposons: constituent 42% du genome humain, la majorité sont cependant inactifs. Ils transposent par un mode copier coller, c'est-à-dire que la sequence donneuse reste a sa place, elle est transcrite en ARN qui est retrotranscrit en ADN, cette sequence est integré a un nouvel endroit du genome.

Les sequences L1 sont impliquées dans les cassures de l'ADN et les cancers et l'inactivation du chromosome X chez les mammiferes; chez la drosophile, ils se sont substitués aux telomerases pour la maintenance des telomeres

b- Sequences répétées en tandem

Les ADN satellites sont des répétitions de séquences nucléotidiques (de quelques paires de bases à plusieurs centaines) qui peuvent s'étendre sur plusieurs milliers de kilobases. Ces séquences, constituant majeur de l'hétérochromatine centromérique et péri-centrique, représentent une fraction importante des génomes eucaryotes. Paradoxalement, la fonction des ADN satellites demeure largement incomprise. En effet, ces séquences ont été longtemps négligées car considérées comme de l'ADN "poubelle" sans fonction cellulaire. Ces séquences jouent un rôle crucial dans la structure des chromosomes et peuvent influencer la régulation génomique. Bien que l'ADN satellite ne code pas pour des protéines, il est essentiel pour la

stabilité chromosomique et la division cellulaire. De plus, des maladies humaines telles que la dystrophie facio-scapulo-humérale ou certains types de cancers, sont associées à des perturbations de l'organisation de satellites

Les répétitions en tandem sont constituées de répétitions adjacentes plus ou moins nombreuses d'un monomère (motif) donné. Certains de ces éléments ont la propriété d'être variables en nombre de répétitions, une conséquence de phénomènes de mutation particuliers (glissement de polymérase et recombinaison inégale). Ces répétitions en tandem peuvent être séparées en trois classes distinctes : les satellites, les minisatellites et les microsatellites.

*Les satellites se dénissent comme des répétitions en tandem possédant un très grand nombre de répétitions, pouvant régulièrement atteindre plusieurs mégabases

*Les minisatellites ou VNTR (variable number tandem repeat), sont des répétitions en tandem d'un motif compris entre dix et une centaine de paires de bases.

* Les microsatellites ou STR (Short tandem repeats) ou SSR (simple sequence repeat), le motif répété est plus court (1-6 nts) long (10-60 nts), il s'agit de minisatellite ou VNTR (variable number tandem repeat)

3.2 Les Pseudogenes

Sont vestiges de genes qui ont perdu leur fonctionnalité. Les pseudogenes donc ressemblent aux genes, ils peuvent etre reconnus grace a des sequences caracteristiques des genes comme les promoteurs et les sites d'epissage

Ils proviennent soit:

* De la duplication de gene ancestral qui existe dans le genome et qui est actif, et ont accumulé beaucoup de mutations, soit:

- Au niveau de la sequence promotrice, ce qui empeche la reconnaissance par le complexe de transcription et donc l'initiation de ce mecanisme
- Ou au niveau de la sequence codante, par l'apparition de codons stop prematurés. Dans ce cas le pseudogene est transcrit, puisque le promoteur est intact, mais le cadre de lecture ne peut pas donner une proteine fonctionnelle (trop courte)

* Soit de la retrotarnscription d'ARNm, ce qui explique que certains pseudogenes sont depourvus de sequences promotrices, ce qui empeche leur transcription et donc leur expression.

Il a été demontré que certains pseudogenes jouent un role essentiel dans la regulation de

l'expression de leurs genes parentaux. Les pseudogenes transcrits peuvent aussi former des ARN interferents. Ils peuvent aussi reguler les oncogenes et les suppresseurs de tumeur.

4- Chromatine et regulation epigenetique

La régulation de l'expression des gènes eucaryotes implique plusieurs niveaux de controle. Un premier niveau dit chromatinien implique la structure de la chromatine.

*** La chromatine** : Dans les cellules eucaryotes, le genome est organisé en une structure complexe constituée d'ADN et de proteines histones, appelée « chromatine ». Cette structure permet la compaction de l'ADN mais aussi la regulation de ses fonctions et de son expression La chromatine peut adopter deux états qui peuvent etre distingués par la présence de modifications post-traductionnelles des histones et méthylation de l'ADN.

- Euchromatine apparait decondensée aucours de l'interphase, elle correspond aux zones codantes du genome
- Heterochromatine : Région hautement condenséé, garde le meme état de condensation aucours du cycle cellulaire; elle se compose de

* heterochromatine constitutive: structure permanente et figée, contient peu de genes, est formée principalement de zones répétées dont les plus grandes regions sont situées a proximité du centromere et des telomeres

* heterochromatine facultative : structure dynamique, contient les regions codantes maintenues sous une forme compacte, transcriptionnelement inactive, selon le stade de developpement et le type cellulaire

*** La régulation epigenetique** : Chez les eucaryotes, la régulation dite épigenetique implique :

- Les processus par lesquels le genotype, l'ensemble des genes, engendrent le phenotype, les caractéristiques de l'organisme.
- Les changements dans l'expression des genes, c'est-à-dire les changements des états de la transcription des genes. Ces changements sont stables et heritables au cours des divisions cellulaires et n'impliquent aucun changement de la séquence des genes

Les changements epigénetiques sont des caracteres reversibles en général et reprogrammables selon le type cellulaire. En effet, elles permettent la différenciation cellulaire en imposant une empreinte génomique caractéristique de chaque type cellulaire. Les genes sont allumés ou éteints par compaction de la chromatine qui empeche l'acces au promoteur des facteurs de transcription

Le profil de methylation, c'est a dire les positions des methylcytosines (C5) dans le génome, permet d'identifier les gènes ainsi régulés par ces modifications chimiques
Des altérations de cette régulation génétiques dont des épimutations, et sont impliquées dans de nombreuses pathologies et cancers.

* **le code histone** : La région NH2 terminale des histones est accessible à l'extrémité du nucleosome

Les modifications post traductionnelles des histones s'effectuent dans cette region et entrainent des modifications des états de condensation de la chromatine (euchromatine et heterochromatine) et donc l'accessibilité des complexes proteiques de transcription au promoteur

Il existe une grande diversité des modifications post-traductionnelles des extremités N-terminales des histones;

Les differentes combinaisons de ces modifications (code histone) agissent de manière coordonnée pour reguler l'expression des genes. Chaque etat transcriptionnel pourrait etre associé a un code histone particulier

Les principaux types de modifications des histones sont :

- Acétylation des lysines
- Methylation des lysines/arginine
- Phosphorylation des serines/tyrosines
- Ubiquitination des lysines
- Sumoylation des lysines

La combinaison des modifications post-traductionnelles des queues des histones constituent le « code histone », un code specifique reconnu par le système des complexes proteiques « writer » et « reader » . Ce systeme permet d'interpreter ces profils de modifications et conduit a des processus biologiques specifiques (regulation de la transcription, reparation de l'ADN, inactivation du chromosome X, organisation du centromere…)

Ces modifications permettent la mise en place de l'empreinte genetique, l'empreinte parentale ainsi que l'inactivation du chromosome X :

* **Empreinte genomique** : Definit de façon stable la repression ou l'expression des genes dans des lignées cellulaires spécifiques (les deux alleles sont soumis a l'empreinte) en fonction du type cellulaire (differenciation cellulaire), selon les besoins de la cellulle et/ou en reponse a des stimuli

* **Lyonisation:** Correspond a l'inactivation du chromosome X. Un chromosome

complet condensé chez la femme connu également sous le nom de corpuscule de Barr (chromsome X condensé et inactif)

L'inactivation du chromosome X se met en place de manière aleatoire aucours du developpement embryonnaire, puis elle se maintient dans les cellules somatiques

*** Empreinte parentale**: Est un mécanisme physiologique conduisant a l'inactivation de l'un des deux alleles parentaux de certains genes, selon leur origine paternelle ou maternelle

- Empreinte maternelle : Le gene paternel s'exprime le gene maternel ne s'exprime pas
- Empreinte paternelle : Le gene paternel ne s'exprime pas et le gene maternel s'exprime