Université Batna 2

Faculté des sciences de la nature et de la vieDépartement de biologie des organismes

Cours de Génomique et protéomique fonctionnelle M2 BMC

2022-2023

Table des matières

Introduction :

La notion de marqueur génétique a été introduite en 1980. Le marqueur génétique est un gène ou une séquence polymorphe d'ADN **aisément** détectable grâce à un emplacement **connu** sur un chromosome. On peut l'utiliser en cartographie génétique pour « **baliser** » le génome et identifier des individus ou des espèces. Les marqueurs peuvent être de différentes natures : RFLP, microsatellites, **SNP**, EST, etc….

Par convention, **une variation rare** (affectant par définition **moins de 1 %** des chromosomesd'un individu) est appelée **une mutation** et **une variation fréquente** (**à partir de 1%**), un **polymorphisme** ou **variant** génétique. Sur les trois milliards de bases du génome humain, plusde 99 % sont identiques d'un individu à l'autre. Les autres représentent les variants génétiquesou polymorphismes. Ce pourcentage peut paraître négligeable, mais représente pourtantenviron trois millions de nucléotides qui sont à la base des différences entre les individus. Ces polymorphismes résultent de trois types de modifications de la séquence d'ADN :

_ des variations de nucléotides uniques (**SNP** pour single nucleotide polymorphism) ou polymorphismes touchant un seul nucléotide, qui correspondent à la substitution d'un nucléotide par un autre ;

_ des insertions ou des délétions de bases, qui peuvent impliquer une seule base ou des centaines de milliers de nucléotides ;

_ des délétions d'ADN répétitifs (tandem repeats, microsatellites).

Polymorphisme mono-nucléotidique (single-nucleotide polymorphism en anglais SNP) :

Le polymorphisme mono-nucléotidique est, en génétique, la **variation** (polymorphisme) d'une **seule** paire de bases du génome, entre individus d'une même espèce. La variation doit être située à un endroit spécifique du génome et apparaître sur une proportion supérieure à **1%** de la population pour être caractérisée comme SNP. Ces variations sont très fréquentes. Ils constituent la plus **petite forme** de polymorphisme car ils n'affectent qu'une seule paire de bases et le type le plus commun de variation génétique intervenant dans la variabilité interindividuelle. Ils représentent plus de **90%** de l'ensemble des variations génétiques humaines. Un SNP se caractérise par sa position chromosomique, ses allèles et sa fréquence allélique mineure appelée (Minor Allele Frequency en anglais ou MAF).

Dans deux génomes humains tirés au hasard, 99,9% de la séquence d'ADN est identique. Les 0,1% restants contiennent des variations de séquence dont le type le plus commun est le polymorphisme pour un nucléotide (SNP). Les SNP sont stables, très abondants et distribués uniformément dans tout le génome. Ces variations sont associées à la **diversité** entre

populations ou individus, une différence de **sensibilité** à des maladies et la **réponse** individuelle aux médicaments.

Les marqueurs SNP, qui sont des polymorphismes dus à la substitution d'un nucléotide ont remplacé les marqueurs microsatellites, qui sont des répétitions de séquences très courtes. Les marqueurs SNP sont bi-alléliques et présentent les avantages d'être **très nombreux** sur le génome et facilement identifiables par les outils technologiques actuels à un coût raisonnable. A contraire, les marqueurs microsatellites sont beaucoup **moins nombreux**, ont un coût plus élevé, mais présentent l'avantage d'être très polymorphes.

Le séquençage du génome humain et les progrès des techniques de biologie moléculaire ont permis de répertorier plus de 10 millions de SNP dans des bases de données publiques, notamment la banque de données **dbSNP** (http://www.ncbi.nlm.nih.gov/projects/SNP/) géré par NCBI « *National Center of Biotechnology Information* » et celle du projet « **HapMap** » (http://hapmap.ncbi.nlm.nih.gov/). Le projet HapMap (2003) a pour objectif de créer une base de données publique détaillée des variations génétiques les plus fréquentes chez l'humain. Il apermis d'établir un catalogue des SNP chez les individus provenant d'Afrique, d'Europe et d'Asie. Plus récemment, le projet «1000 génomes » (http://www.1000genomes.org/), basé sur un séquençage de nouvelle génération, ont apporté une contribution énorme à l'identification et la caractérisation des SNP pour les populations du monde entier.

1.Localisation :

Les SNP peuvent se retrouver au sein de régions codantes de gènes (exon), de régions non codantes de gènes (intron), ou de régions intergéniques, entre les gènes. Dans le cas où les SNP se retrouvent au sein des régions codantes, celles-ci ne vont pas obligatoirement modifier la séquence d'acide aminé de la protéine produite, et ce, grâce à la redondance du code génétique. Les SNP qui se retrouvent dans des régions non codantes peuvent avoir des conséquences sur l'épissage, les facteurs de transcription, ou sur les séquences d'ARN non codant. Il existe des variations de la répartition des SNP au sein du génome car ils sont plus fréquents dans les introns et les régions intergéniques que dans les exons.

2.Types :

On parle de formes alléliques synonymes dans le cas où plusieurs formes d'un SNP mènent à la même séquence polypeptidique, et de formes non synonymes dans le cas où les séquences produites diffèrent. Bien entendu ceci ne s'applique qu'aux régions codantes du génome. Les SNP synonymes sont plus fréquents que les SNP non synonymes.

3.Fréquence :

En moyenne un SNP est rencontré tous les 100 à 1000 nucléotides. Il y en a de l'ordre

de 5 x 10^6 dans le génome humain. Certaines associations de SNP sont caractéristiques de certaines populations. La distribution des SNP est au hasard. Dans n'importe quel gène on peut attendre une moyenne de 10 SNP, mais certains peuvent n'en présenter aucun. En 2001 on avait recensé 800 000 SNP dans le génome humain (Tableau1).

Tableau1 : Répartition des SNP connus sur les chromosomes humains

Chromosome	nombre de SNP	Chromosome	nombre de SNP	Chromosome	nombre de SNP
1	16759	9	5790	17	6392
2	12748	10	6014	18	2682
3	10112	11	6931	19	7664
4	6995	12	7375	20	5381
5	9146	13	2847	21	3478
6	13888	14	5827	22	5400
7	12389	15	4343	X	3253
8	4962	16	5771	Y	63
				Total	**177594**

4. Application :

- Les SNP constituent une trace historique pour l'étude de la **phylogénie** humaine.
- Les SNP sont à l'origine de **susceptibilité** ou de **résistance** à de nombreuses maladies. Elles peuvent servir à l'identification des groupes à risques pour une pathologie ou à recruter des patients pour les essais cliniques en fonction, non plus de symptômes mais de **prédispositions génétiques**. Les personnes présentent des niveaux de sensibilité différents pour un large éventail de maladies humaines. Cela est principalement dû aux SNP dans le génome humain. La gravité de la maladie et la façon dont le corps réagit aux traitements sont également déterminées par les SNP présents dans le génome humain. Par exemple, les individus porteurs d'une mutation d'une base du gène APOE (gène de l'apolipoprotéine) présentent un risque plus élevé de contracter la maladie d'Alzheimer.
- Elle est très précieuse en pharmacogénomique, méthode visant à distinguer au sein de la population générale, sur la base de leur profil SNP, les individus qui répondent positivement à l'administration d'un médicament de ceux qui n'y répondent pas ou qui développent des effets secondaires.
- Les SNP permettent de cartographier le génome dans le cas de maladies multifactorielles complexes telles le cancer et le diabète en vue d'identifier quels sont les loci impliqués dans ces maladies. On parle des études de **liaison** et **d'association**

4.1. Cartes génétiques et carte physique :

Il existe deux types de cartes, ou plus exactement **deux distances** : distance physique et distance génétique. La première se définit comme étant la longueur de la séquence d'ADN séparant deux points du génome, mesurée en nombre de paires de bases azotés (pb), kilobases (1 kb =1000pb) ou mégabase (1 Mb = 10^6 pb).

La seconde est mesurée par **le taux de de recombinaison méiotique**. Elle est mesurée en **centimorgans** (cM). Lors de la méiose, des marqueurs situés sur un même chromosome peuvent être séparés s'il se produit un crossing-over. La probabilité qu'un tel évènement se produise est proportionnelle à la distance qui sépare ces marqueurs. La fréquence de recombinaison reflète donc des distances entre marqueurs. Une distance de 1 cM entre deux locus correspond à une probabilité de **1%** pour eux d'être séparés par un crossing-over. Notons que l'ordre est toujours le **même** sur une carte génétique et une carte physique. Seules les distances relatives **changent**.

La recombinaison survenant de façon hétérogène le long du génome, la correspondance entre distance génétique et physique varie donc selon les régions considérées. En moyenne, 1 cM équivaut à 0,88 Mb (il est considéré grossièrement que 1 cM équivaut à 1 Mb) avec une variabilité selon le sexe. En effet, chez les femmes les recombinaisons sont deux fois plus fréquentes que chez les hommes, ce qui entraîne pour un même chromosome des tailles différentes en cM, mais pas en Mb.

4.2. Etude de liaison et d'association :

Un des grands objectifs de la recherche en génétique est de trouver les gènes causant les maladies de type génétique. 1l y a principalement deux types d'études permettant d'identifier le, ou **les gènes causant une maladie** : les études de liaison et les études d'association.

A.Les études de liaison :

La notion d'haplotype :

Un haplotype est une combinaison allélique de plusieurs marqueurs polymorphes sur un même chromosome à différents loci. L'ensemble de ces allèles sont localisés à une distance **très proche** et ségrégent en **bloc** comme un seul marqueur. Au sein d'un même bloc, la probabilité de recombinaison est **très faible**. Le concept d'haplotype revêt une importance capitale dans le domaine de la génétique. Il permet de mieux étudier le comportement des SNP entre eux. De plus, ils permettent de renseigner les variabilités d'origine génétique des populations humaines.

La liaison génétique :

La liaison génétique est définie comme la **co-ségrégation** de deux, ou plusieurs gènes

(ou loci), au cours des générations en raison de leur **proximité** physique sur le génome. En d'autres termes, la liaison génétique est le phénomène par lequel les allèles d'un **haplotype** transmis par un grand parent au parent, auront tendance à être transmis encore du parent au fils. On dit alors, que les allèles des différents gènes sont **génétiquement liés**.

Comme on a déjà vu, deux loci sont liés s'ils sont situés proches l'un de l'autre, ce qui **diminue** donc les chances de recombinaison et, par le fait même, augmente les chances que leurs allèles soient transmis ensemble. Le locus de la maladie **n'étant pas connu**, on cherche alors des marqueurs situés dans son **voisinage**. **Si l'allèle d'un marqueur est fréquemment transmis aux personnes malades dans les familles**, on dira que ce marqueur est **lié** à la maladie. On peut alors concentrer la recherche du gène responsable **autour** de ce marqueur. Les analyses de liaison sont considérées comme un outil très puissant pour détecter la position d'un gène causant une maladie dans une région chromosomique.

Les analyses de liaison génétique testent **la co-transmission** d'une région chromosomique avec un phénotype. A partir de données **familiales**, elles permettent de tester l'indépendance de transmission entre des marqueurs polymorphes et la maladie dans les familles, et si ce n'est pas le cas, de situer le locus de maladie par rapport aux marqueurs. En effet, plus un marqueur estproche du locus de la maladie, plus la probabilité qu'il en soit séparé par un crossing-over est faible et celle qu'il soit **co-hérité** avec lui élevée. Montrer que la transmission d'une pathologie et d'un marqueur, dont la localisation est connue, se fait de façon non indépendante permet donc de positionner approximativement le locus de la maladie sur le génome.

B.Études d'association :

Les études d'association ont pour but d'identifier **un marqueur pour lequel un allèle est plus fréquent chez les sujets malades que chez les sujets sains**. **On dit alors que cet allèle est associé à la maladie.** La recherche d'associations, entre des marqueurs génétiques et des maladies, est une des voies possibles pour identifier des gènes de susceptibilité aux maladies. La technique d'étude est relativement simple : on compare la fréquence d'un marqueur génétique chez les sujets atteints et chez les témoins (étude cas-témoins). Un résultat positif (différence de distribution allélique entre les cas et les témoins) suggère que le marqueur est soit **directement impliqué**, soit **en déséquilibre de liaison** avec une ou plusieurs variation(s) génétique(s) causale(s).

Il est possible d'effectuer ces études à partir d'échantillons de cas et de témoins, mais également à partir d'échantillons de familles. Une association peut être due, soit à l'existence d'une **liaison**,soit à d'autres facteurs.

Une distinction importante s'impose entre liaison et association. Il y a liaison quand on

estimeque le marqueur se trouve à une **distance proche** de l'allèle morbide, de simple faite de se trouver sur le même chromosome, alors qu'il y a association quand le SNP a une **association significative** avec la maladie ou le caractère étudié. On peut avoir une association sans qu'il yait liaison.

Les études d'association sont utilisées dans le but de confirmer l'implication d'un allèle qu'on pense être important dans l'étude d'une maladie ou pour découvrir de nouveaux gènes pouvant jouer un rôle dans la maladie. Les études d'association comparent la répartition des allèles en fonction du trait étudié entre les individus qui portent le trait et ceux qui ne le portent pas. Plus la répartition est **différente**, plus le SNP est susceptible d'être impliqué avec le trait étudié.

Les études d'association ont été mises en œuvre depuis 2006, comme l'*International HapMap project* ou les *Genome-wide association studies* (GWAS). Ces études permettent, en analysant le génome d'un grand nombre d'individus **sains et de patients**, de détecter l'effet, même modeste, des variants génétiques dans une pathologie donnée. Ainsi, un nombre important de **polymorphismes synonymes** ont été associes à diverses pathologies.

Par exemple, quelques polymorphismes synonymes sont associés à un risque accru de développer un cancer. En effet, si le risque relatif d'apparition de cancer est généralement faible, la fréquence relativement importante de ces polymorphismes dans la population leur confère un poids considérable.

La découverte du gène *TERT* (*telomerase reverse transcriptase*), qui joue un rôle majeur dans l'immortalité et/ou le vieillissement cellulaire, a valu a Elizabeth Blackburn, Jack Szostak et Carol Greider l'attribution du prix Nobel de médecine en 2009. Ce gène code pour la sous-unité catalytique de la **télomérase** et possède de très nombreuses variations génétiques. Par exemple, le polymorphisme synonyme rs2736098 est associe à une augmentation du risque de développer un cancer. Il est **fréquent** dans la population puisque trouve chez 10 % de la population européenne a l'état homozygote, et 53 % à l'état hétérozygote.

C.Déséquilibre de liaison (linkage disequilibrium) :

Les études d'association sont fondées sur le principe du déséquilibre de liaison. On définit le déséquilibre de liaison comme une association **non aléatoire** entre les allèles de deux loci.

S'il existe un déséquilibre de liaison, cela signifie que les gènes appartenant à des loci différents ne sont **pas associés au hasard** dans la population : **certaines combinaisons sont moins fréquentes, d'autres plus fréquentes** que ne le voudrait le hasard si on avait une association aléatoire de ces différents gènes, c'est-à-dire il y **a une association préférentielle**

entre les deux allèles. Un **Déséquilibre de liaison entre deux loci est défini par l'existence d'une combinaison d'allèles à ces loci plus fréquente que celle attendue sous l'hypothèse d'indépendance.**

Principe d'équilibre de Hardy-Weinberg :

Le modèle d'équilibre de Hardy -Weinberg est l'un des principes fondamentaux de la génétique des populations. Cette loi a été établie indépendamment en 1908 par le médecin W. Weinberg (1862-1937) et le mathématicien G.H. Hardy (1877-1947). Dans leur modèle, Hardy et Weinberg ont supposé les hypothèses suivantes :

- La population est de taille infinie,
- Les gamètes (ovules et spennatozoïdes) s'associent au hasard c'est-à-dire que les accouplements sont aléatoires (hypothèse de panmixie),
- Il n'y a pas de sélection dans la population,
- Il n'y a ni mutation, ni migration dans la population.

Alors, les fréquences des allèles et des génotypes au cours des générations suivent une loi simple appelée loi de Hardy-Weinberg. Cette loi stipule que les fréquences alléliques et les fréquences génotypiques restent **constantes** de génération en génération. On dit alors que la population est **à l'équilibre**, et il existe une relation simple entre les fréquences alléliques et les fréquences génotypiques.

Pour conclure, en supposant les hypothèses de Hardy-Weinberg, dans une génération, on peut à partir des fréquences alléliques déduire les fréquences génotypiques, et celles-ci restent **stables** à partir d'une première génération.

La loi de Hardy-Weinberg permet, sous certaines conditions, le calcul des fréquences génotypiques à partir des fréquences alléliques.

Ainsi :

A1 et A2 deux allèles d'un même locus

- p est la fréquence de l'allèle A1 : $0 < p < 1$
- q est la fréquence de l'allèle A2 : $0 < q < 1$

avec $p+q=1$, avec une répartition identique des fréquences allèliques chez hommes et femmes, soit :

- hommes (p, q), femmes (p, q)
- si ils procréent : $(p+q)^2=p^2+2pq+q^2=1$

où :

- p^2 = fréquence du génotype A1 A1 ← HOMOZYGOTE
- 2pq = fréquence du génotype A1 A2 ← HETEROZYGOTE
- q^2 = fréquence du génotype A2 A2 ← HOMOZYGOTE

Fréquences constantes au fil des générations.

	Gamète femelle **A** (p)	Gamète femelle **a** (q)
Gamète mâle **A** (p)	**AA** (p^2)	**Aa** (pq)
Gamète mâle **a** (q)	**Aa** (pq)	**aa** (q^2)

Fréquence de AA = p^2
(Homozygote)
1/4

Fréquence de Aa = 2pq
(Hétérozygotes)
1/2

Fréquence de aa = q^2
(Homozygote)
1/4