UE 11 - Biomédecine Quantitative

06/03/14 de 13h30 à 15h30

Dr Laouénan

Ronéotypeur : Pierre Vandevelde

Ronéolectrice : Elodie Meiranesio

**Cours 5 :**

**Techniques d’analyse de la biologie moderne Epidémiologie génétique**

**Plan**

**I - Techniques d’analyse de la biologie moderne**

1. Génome - Transcriptome - Protéome
2. Biopuces (microarrays)

**II - Épidémiologie Génétique**

1. Maladies monogéniques / multifactorielles
2. Montrer l'existence d'une agrégation familiale

Études épidémiologiques

1. Mise en évidence de l’existence de la composante génétique

Études de jumeaux et d'adoption

Analyses de ségrégation

1. Analyser la composante génétique

Analyses de liaison

Études d’association

*Le professeur n’a fait que lire ses diapos. A la fin du cours il a insisté sur l’importance des définitions dont la majorité est déjà vue (je vous rassure). Pour le partiels, je cite : l’utilisation de ces notions est également possible dans un arbre généalogique à interpréter.*

**I - Techniques d’analyse de la biologie moderne**

L’étude en biologie moderne ce fait en premier lieu grâce à l’utilisation de biomarqueur. Ces biomarqueurs seront de différents types en fonction de l’échelle étudiée.

Biomarqueur : Elément quantifiable qui va permettre de renseigner sur un effet biologique, physiologique ou une susceptibilité à une pathologie. Ils sont de deux types : les biomarqueur prédictif (prévoir l’efficacité d’une thérapeutique) et les biomarqueur pronostic (caractériser l’évolution de la maladie, la survie des patients).

Un biomarqueur peut être prédictif et pronostic mais le plus souvent il n’est pas de bonne qualité pour chaque fonction.

1. Génome - Transcriptome - Protéome

Dans le passé l’étude de ces notions était basée sur une relation fausse : 1 gène = 1 ARN = 1 protéine Le génome code pour 20 000 gènes, il y a plus d’un million d’ARNm et plus de 10 millions de protéines

Les biomarqueurs à différentes échelles :

Génome : l’ADN

Transcriptome : l'ARN

Protéome : les protéines

Le Génome :

C’est l’ensemble du matériel génétique d’un individu, le support de l'information génétique, composé de près de trois milliards de paires de bases, réparties entre les 23 paires de chromosomes et le même pour toutes les cellules d'un même organisme (à de rare exception du type Lymphocyte B avec les réarrangements VDJ)

Chez l’Homme il y a 20 000 gènes codant chacun pour une ou plusieurs protéines (+/- 2 % de l’ADN). Avant l’ADN non codant était vu comme de l’ADN poubelle.

Avec le Projet ENCODE lancé en 2003 (ENCyclopaedia Of DNA Elements) :

–Plusieurs millions d’interrupteurs régulant l’activité des gènes et des mutations dans ces régions pourraient induire des maladies.

– Certains sont transcrits en ARN (non codant) d’autres sont des pseudogènes (non transcrits).

– 80 % du génome humain est fonctionnel (rôle de régulation, stabilité...)

Diversité génétique :

Polymorphisme : variations de la séquence d'ADN présente dans plus de 1% de la population.

Les erreurs de copie se produisant lors de la réplication de l'ADN ensuite transmises de génération en génération et se fixent dans la population pour former un polymorphisme.

Quand une mutation dépasse les 1% de la population, on ne parle plus de mutation, mais on parle de polymorphisme.

Locus : une position du génome.

Allèle : chacune des variations possibles de l'ADN en un locus du génome.

Pour chaque polymorphisme, le génotype d'un individu est défini comme la combinaison des deux allèles présents sur chacun des brins d'ADN hérités de ses parents

–Homozygote : individu qui possède 2 allèles identiques pour un même gène.

–Hétérozygote : individu qui possède 2 allèles différents pour un même gène.

Haplotype : combinaison de plusieurs allèles situés sur des locus différents d'un même chromosome.

Les mutations de l'ADN amenant à la formation de polymorphismes sont :

1. Substitutions d'une partie de la séquence par une séquence alternative de la même longueur

= remplacement d'une base par une autre

= polymorphisme à un seul nucléotide (SNP pour Single Nucleotide Polymorphism)

Les SNPs apparaissent en moyenne une fois tous les 2 000 bases.

3 millions de SNP ont aujourd'hui été recensés dans le projet HapMap visant à cartographier la variabilité du génome humain.

2. Insertions/délétions de séquences pouvant aboutir à des répétitions de certains éléments de séquence et à la formation de polymorphismes multi-alléliques (plus de 2 allèles possibles)

Variations du nombre de copies (CNV : Copy Number Variation) : duplications ou délétions au niveau de certains segments de chromosomes

Microsatellites de répétitions : motif de plusieurs bases répétées n fois.

Impact des variations du génome :

Polymorphismes non fonctionnels : aucune conséquence

Polymorphismes fonctionnels :

–modification de l’expression du gène

–modification de l’épissage de l’ARN

–stabilité de l’ARNm

–changement d’AA

Transcriptome :

C’est l’ensemble des ARN issus de la transcription. La variabilité du transcriptome dépend de la différenciation cellulaire et des stimuli environnementaux

•ARN codant (ARNm)

•ARN non codant (microARN) qui régulent l'expression des ARNm

–Les régulations épigéniques

–Implication dans l'apparition de nombreuses pathologies (cancer…)

–constitués d'une vingtaine de nucléotides

–1000 microARNs

Protéome :

C’est la caractérisation et étude de l’ensemble des protéines exprimées par un génome et plus particulièrement celui d’une cellule, d’un tissu ou d’un organisme. C’est une entité dynamique et complexe.

Human Proteome Organisation (HUPO) : base de données unique permettant de décrire les protéines correspondant aux 20 300 gènes codants chez l’homme

On l’analyse par spectrométrie de masse. Avec un objectif majeur : la découverte de biomarqueurs (dosage sérique de 5 biomarqueurs de l'infarctus du myocarde par ex.)

Les étapes pour avoir une protéine fonctionnelle sont : transcription, traduction, modifications post –traductionnelles.

Métabolome :

C’est l’étude de l'ensemble des métabolites (sucres, acides aminés, acides gras...) présents dans une cellule, un organe ou un organisme

Métagénome :

C’est l’étude des microorganismes, jusqu'à ce jour, incultivables (bactéries, champignons, levures…) Il permet des applications au corps humain mais aussi à des écosystèmes complexes.

Extraction de l'ADN et analyse des séquences génomiques des organismes présents.

Exemple : microbiote intestinal (100 000 milliards de bactéries).

–Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach (Manichanh C et al., Gut 2006).

Autre exemple : signatures bactériennes associées à l’obésité.

On définie des “paysages écologiques” différents = entérotypes.

Chaque entérotype est spécifique a un individu.

Les « omes » permettent d’aborder la complexité du vivant dans son ensemble. Ces approches sont particulièrement utiles pour mieux connaître les maladies héréditaires et adapter les traitements au profil génétique.

1. Biopuces (microarrays)

Elles permettent d’analyser des séquences d’ARN ou d’ADN a très haut débit. Elles détectent et quantifient simultanément plusieurs centaines de milliers de séquences et exploitent la propriété d'hybridation des séquences nucléiques.

La présence de légères différences (SNP) dans la séquence des deux brins diminue l’efficacité de l’appariement et fragilise les liaisons.



Seules deux sont à retenir :

Les puces d’expression :

Mode de fonctionnement :

1. Les ARNm issus de la transcription des gènes sont extraits de la cellule ou du tissu. Des ADN complémentaires (ADNc) sont synthétisés par rétro-transcription puis purifiés et amplifiés.

2. Le brin complémentaire des ADNc purifiés est re-synthétisé à partir de nucléotides marqués à la biotine. Les ADN ainsi obtenus fournissent alors une image fidèle de la séquence du transcriptome, où chaque morceau de séquence est présent en quantité proportionnelle à la quantité de transcrit correspondant.

3. Les ADN sont alors hybridés sur la puce et la puce est lavée.

4. Un fluorochrome Cy3 est appliqué sur la puce.

5. La puce est scannée. L'intensité lumineuse qui est mesurée sur un point de la puce évalue alors la quantité de l'ARNm correspondant à ce point, dans la cellule ou le tissu, pour l'échantillon étudié.

Les puces de génotypage :

Le génotypage par biopuce a pour objectif d'identifier simultanément le génotype de plusieurs centaines de milliers de SNPs.

La capacité des puces de génotypage n'a cessé de croître, passant de 10 000 SNPs en 2004 à plus de 2.5 millions sur la dernière génération de puces.

Les stratégies de choix des SNPs cherchent à établir une couverture fine de tout le génome.

**II - Épidémiologie Génétique**

La plupart des maladies humaines présentent une composante génétique plus ou moins forte.

L’épidémiologie génétique est la description et la compréhension de ces facteurs génétiques

Les maladies monogéniques ou mendéliennes :

–Présence d'une version déficiente du gène (majeur) responsable de la maladie (mutation transmise)

–Maladies rares et le plus souvent graves (mucoviscidose, myopathies…)

Les maladies multifactorielles ou complexes :

–Composante génétique et environnementale et résultent généralement d'interactions complexes entre ces deux facteurs

–Maladies fréquentes (obésité, maladies cardiovasculaires…)

–Présence simultanée de nombreux allèles de prédisposition (ou de susceptibilité) ayant un impact individuel modéré

1. Maladies monogéniques / multifactorielles

Les maladies monogéniques :

**Fréquence en population** : rare

**Agrégation familiale** : élevée

**Phénotype** : souvent sévère, invalidant voire fatal

**Facteurs environnementaux :** peu nombreux voire absents

**Facteurs génétiques :** une mutation au niveau d’un seul gène est nécessaire et suffisante pour faire apparaître la maladie

**Transmission** :

–Autosomique dominante (50% des enfants touchés)

–Autosomique récessive (25% des enfants touchés)

–Dominante liée à l’X

–Récessive liée à l’X

**Pénétrance** : forte (pénétrance = probabilité qu’un individu soit atteint sachant son génotype)

Les maladies multifactorielles :

Maladies à hérédité complexe ou polygéniques

**Fréquence en population** : modérée à élevée

**Agrégation familiale** : faible

**Facteurs impliqués** :

–Maladies polygéniques avec gène majeur (nécessaire mais pas suffisant)

–Maladies polygéniques sans gène majeur

**Pénétrance** : faible

**Effet de l’environnement** : important

Stratégies de recherche : de l'approche « gène candidat » aux « études génome entier »

Gène candidat: gènes connus et potentiellement impliqués dans l'étiologie de la maladie étudiée

Etudes génome entier: recherche de polymorphismes génétiques prédisposant au développement d'une maladie en balayant l'ensemble du génome à la recherche de signaux d'association indiquant la présence d'un locus de prédisposition (études d'association génome entier ou GWAS).

Ce changement de stratégie c’est fait grâce au développement du génotypage à haut débit. L’approche gène candidat est toujours utilisée actuellement mais a vu grandement réduire son importance.

Dans tous les cas il faut :

1. Montrer l'existence d'une concentration familiale (agrégation familiale) de la maladie : **études épidémiologiques**

2. Montrer que cette agrégation est due à une composante génétique et la caractériser : **études de jumeaux et d'adoption** et **analyses de ségrégation**

3. Localiser, identifier et préciser les effets de cette composante (gènes impliqués, polymorphismes fonctionnels, interactions entre les gènes et avec les facteurs d'environnement) : **études de liaison génétique** et **études d'association**

•utilisent les polymorphismes de l’ADN (microsatellites ou SNP).

•permettent de définir les régions du génome pouvant contenir des gènes de susceptibilité (criblage systématique du génome).

•les études fines des régions de susceptibilité permettent ensuite de définir les variants génétiques responsables de la susceptibilité génétique à la maladie, et la façon dont l'environnement module l'effet de ces gènes.

1. Montrer l'existence d'une agrégation familiale

Excès de cas familiaux = agrégation familiale

Pour cela il faut montrer que chez les apparentés du premier degré (parents, frères, soeurs ou enfants) des malades, la maladie est plus fréquente que dans la population générale et chez les apparentés du premier degré de sujets sains (témoins)

Concentration familiale peut être due à :

–une transmission socioculturelle

–une exposition environnementale commune aux différents membres de la famille

–une susceptibilité génétique

1. Mise en évidence de l’existence de la composante génétique

Les études de jumeaux :

Monozygotes (MZ) : partagent le même patrimoine génétique

Dizygotes (DZ) : partagent 50% de leur patrimoine génétique

Hypothèse : les différences d’environnement entre 2 jumeaux (au sein d'une même paire) sont les mêmes pour les MZ et les DZ (ce n’est pas totalement vrai, les MZ partagent plus de caractéristiques environnementales que les DZ)

Les cas possibles sont :



On peut avec calculer le taux de concordance.

**Principe :**

Pour les MZ : une discordance (1 jumeau malade, 1 jumeau sain) ne peut être que d’origine environnementale.

Pour les DZ : une discordance peut être d’origine génétique et/ou environnementale.

Il faut comparer le taux de concordance pour la maladie entre les paires MZ et les paires DZ.

**Interprétation :**

Si une maladie est totalement génétiquement déterminée alors :

–100% des jumeaux MZ devraient être concordants pour la maladie

–50% des jumeaux DZ devraient être concordants pour une maladie autosomique dominante, 25% pour une maladie autosomique récessive

Si une maladie n’a rien de génétique alors :

–La concordance au sein des paires MZ devrait être égale à la concordance des paires de DZ

Si une maladie est multifactorielle avec une composante génétique

–Les jumeaux MZ ne devraient pas toujours être concordants mais ils le seront plus souvent que les DZ

**Limites :**

Souvent lent et difficile

Environnement plus souvent partagé pour les MZ que pour le DZ

Les études d’enfants adoptés :

**Principe :**

L’adoption sépare les enfants de leurs parents biologiques

Un enfant adopté :

Par rapport à ses parents adoptifs : Il partage une partie de son environnement mais a peu de similarités génétiques.

Par rapport à ses parents biologiques : Il partage en moyenne, 50% de leur patrimoine génétique mais a peu partagé leur environnement.

**Conditions :**

Les enfants doivent avoir été abandonnés très jeunes (si possible dès la naissance)

Les enfants doivent avoir été placés très tôt dans leur famille adoptive

Les enfants doivent avoir été placés au hasard

Les données sur les parents biologiques et adoptifs doivent être complètes

**Interprétation :**

Si trait est génétiquement déterminé alors on s’attend à retrouver plus fréquemment la maladie dans la famille biologique de l’enfant adopté.

Si une maladie n’a rien de génétique, si elle est totalement sous l’influence de facteurs environnementaux, l’enfant ressemblera alors d’autant à ses parents adoptifs.

**Limites :**

Enquêtes très difficiles et quasiment impossibles en France car elles nécessitent dans l’idéal :

–de recenser des enfants adoptés

–de connaître les parents adoptifs

–de connaître les parents biologiques

Analyses de ségrégation :

Elles permettent de répondre à deux types de question :

Quel est le mode de transmission d'un phénotype donné ?

Peut-on individualiser un gène majeur parmi l’ensemble des facteurs impliqués ?

**Objectif :**

Déterminer le modèle génétique qui explique le mieux les distributions familiales d’une maladie

Cherche à mettre en évidence l'effet d'un gène transmis de façon mendélienne (classiquement appelé gène majeur ou à effet fort) parmi l'ensemble des facteurs génétiques et environnementaux impliqués.

Estimer les paramètres du modèle génétique :

–Fréquence de l’allèle délétère

–Mode de transmission

–Pénétrance

Déterminer, par des tests statistiques, le mode de transmission expliquant le mieux les distributions familiales observées du phénotype étudié

Cette analyse est compliquée dans le cas de maladies multifactorielles en raison de l'existence d'une hétérogénéité phénotypique et génétique

L'hétérogénéité phénotypique : observation qu'une même maladie peut se présenter sous différentes formes cliniques

L'hétérogénéité génétique : cas où des mutations différentes d'un même gène sont la cause de la même maladie

1. Analyser la composante génétique

1. Analyse de liaison génétique (**sur données familiales**) :

Le but est de localiser le(s) gène(s) responsable(s) de la maladie sur le génome.

On fait une analyse de régions candidates si il y a une hypothèse sur le gène en cause.

On pourra confirmer ou infirmer une liaison génétique d’une région avec le phénotype étudié.

Sinon on fait une recherche sur l’ensemble du génome car pas d’hypothèse a priori sur le gène en cause. On cherche à savoir si la maladie et des allèles de marqueurs coségrègent dans les familles

Le déséquilibre de liaison gamétique :

La transmission d'une maladie implique la transmission d'une région chromosomique particulière.

Dans cette région sont localisés :

–Le gène morbide (mutation responsable de la maladie)

–Un marqueur génétique (mutation non responsable de la maladie) (SNP)

Si le marqueur et le gène morbide sont localisés à proximité l'un de l'autre sur le même segment chromosomique, ils seront toujours transmis simultanément. On dira qu'ils sont liés (d'où le terme de liaison génétique)

Ils pourront être séparés par recombinaison lors de la méiose. La probabilité de cet événement est dépendante de la **fréquence de recombinaison** et de la **distance entre le marqueur et le gène morbide.**

Indépendance : Si deux locus sont répartis de manière indépendante alors ils sont situé sur des chromosomes différents ou sur le même mais très éloignés l’un de l’autre. On aura autant de gamètes parentaux que de gamètes recombinés.

Liaison : Si deux locus sont liés génétiquement, ils sont sur le même chromosome. On aura plus de gamètes parentaux que de recombinés. Si liaison totale alors 100% de gamètes parentaux.

**Objectif :**

Analyser la ségrégation conjointe de deux locus d’une génération à la suivante dans des familles comportant plusieurs sujets atteints

**Résultats :**

Existence d’une liaison (transmission simultanée) entre un marqueur et le gène morbide.

Estimation de la distance génétique entre le marqueur et le gène morbide.

Un des allèles du marqueur génétique est-il transmis de façon préférentielle avec la maladie ?

Si oui, la mutation responsable de la maladie se situe probablement à proximité de ce marqueur.

2. Etudes d’association (GWAS) (**sur données de population**) :

Elles concernent des populations de sujets indépendants (sans lien familial) et font appel aux techniques classiques de l’épidémiologie.

Le plus souvent on réalise des **études cas-témoins.**

Comparaison de la distribution d’un marqueur génétique entre :

–un groupe de sujets atteints de la maladie : cas

–et un groupe de sujets indemnes de la maladie : témoins

Un allèle ou génotype est dit associé à une maladie s’il est plus fréquent (facteur de risque), ou au contraire s’il est moins fréquent (facteur protecteur), chez les cas que chez les témoins.

Elles permettent de comparer la distribution génotypique et allélique d’un polymorphisme entre cas et témoins. Ainsi que d’estimer le risque de maladie associé à la présence d’un allèle ou d’un génotype (même si les effets sont restreints) et de rechercher des interactions gène-gène ou gène-environnement.

Elles reposent sur :

–Des échantillons de population

–Des polymorphismes

–Le déséquilibre de liaison

–Des stratégies de sélection du ou des gène(s) à étudier

**Sélection des échantillons :**

Cas :

Définition de la maladie (critères d’inclusion)

Méthode de recrutement :

–cas prévalents, incidents ?

–cas hospitaliers, en ville, registre ?

–cas consécutifs, tirage au sort ?

Témoins

Comment s’est-on assuré de l’absence de maladie ?

•Méthode de recrutement :

–population générale ?

–voisins, époux ?

–témoins hospitaliers ?

•Représentatifs de la population dont sont issus les cas

Il faut réduire au maximum tous les risques de biais !!!

**RONEOFICHE Cours 5 Biomédecine quantitative UE11**

J’ai repris tous les passages que le prof a déclarés important.

**Biomarqueur**: Elément quantifiable qui va permettre de renseigner sur un effet biologique, physiologique ou une susceptibilité à une pathologie. Ils sont de deux types : les biomarqueur prédictif et les biomarqueur pronostic.

**Polymorphisme** : variations de la séquence d'ADN présente dans plus de 1% de la population.

**Type de Mutation :** insertion/délétion (CNV), substitution (SNP)

Impact des variations du génome :

Polymorphismes non fonctionnels : aucune conséquence

Polymorphismes fonctionnels :

–modification de l’expression du gène

–modification de l’épissage de l’ARN

–stabilité de l’ARNm

–changement d’AA

**Génome :** matériel génétique d’un individu, le support de l'information génétique, 3 milliards de paires de bases, 23 paires de chromosomes et le même pour toutes les cellules d'un organisme

20 000 gènes codant chacun pour une ou plusieurs protéines (+/- 2 % de l’ADN). Mais surement 80% de l’ADN est fonctionnel.

Biomarqueur : ADN

**Transcriptome :** C’est l’ensemble des ARNs issus de la transcription. La variabilité du transcriptome dépend de la différenciation cellulaire et des stimuli environnementaux

Biomarqueur : ARN

**Protéome :** C’est la caractérisation et étude de l’ensemble des protéines exprimées par un génome et plus particulièrement celui d’une cellule, d’un tissu ou d’un organisme. C’est une entité dynamique et complexe.

Biomarqueur : Protéine

**Biopuces :**



**Les maladies monogéniques** :

Fréquence en population : rare

Agrégation familiale : élevée

Phénotype : souvent sévère, invalidant voire fatal

Facteurs environnementaux : peu nombreux voire absents

Facteurs génétiques : une mutation au niveau d’un seul gène est nécessaire et suffisante pour faire apparaître la maladie

Transmission :

–Autosomique dominante (50% des enfants touchés)

–Autosomique récessive (25% des enfants touchés)

–Dominante liée à l’X

–Récessive liée à l’X

Pénétrance : forte (pénétrance = probabilité qu’un individu soit atteint sachant son génotype)

**Les maladies multifactorielles :**

Maladies à hérédité complexe ou polygéniques

Fréquence en population : modérée à élevée

Agrégation familiale : faible

Facteurs impliqués :

–Maladies polygéniques avec gène majeur (nécessaire mais pas suffisant)

–Maladies polygéniques sans gène majeur

Pénétrance : faible

Effet de l’environnement : important

**Gène candidat :** gènes connus et potentiellement impliqués dans l'étiologie de la maladie étudiée

**Etudes génome entier**: recherche de polymorphismes génétiques prédisposant au développement d'une maladie en balayant l'ensemble du génome à la recherche de signaux d'association indiquant la présence d'un locus de prédisposition (études d'association génome entier ou GWAS).

**Agrégation familiale :** accumulation de cas dans une famille à cause de facteurs génétiques, environnementaux ou sociaux culturels.

**Les études de jumeaux :** l’objectif est de mettre en évidence le rôle de l’environnement dans une pathologie donnée. En prenant en compte que chez les monozygotes, c’est le seul facteur de variation alors que chez les dizygotes le patrimoine génétique n’est pas partagé a 100%. On calcule le taux de concordance (jumeau 1 malade/sain vs jumeau 2 malade/sain). Pour une maladie à dominante génétique la concordance sera supérieure chez les MZ alors que pour une composante environnementale on retrouvera moins de différence entre MZ et DZ

**Les études d’enfants adoptés :** Même principe sauf qu’ici les enfants ne partage pas un patrimoine génétique comme leurs parents biologiques sont différents. Mais ils partagent, pour une bonne partie les mêmes caractéristiques environnementales.

**Analyse de liaison génétique** (**sur données familiales !**) : Le but est de localiser le gène responsable de la maladie.On fait une analyse de régions candidates si il y a une hypothèse sur le gène en cause.

Sinon on fait une recherche sur l’ensemble du génome. On cherche à savoir si la maladie et des allèles de marqueurs coségrègent dans les familles.

**Etudes d’association (GWAS) (sur données de population !) :** Elles concernent des populations de sujets indépendants (sans lien familial).Le plus souvent on réalise des **études cas-témoins.**

Un allèle ou génotype est dit associé à une maladie s’il est plus fréquent (facteur de risque), ou au contraire s’il est moins fréquent (facteur protecteur), chez les cas que chez les témoins.

Elles permettent de comparer la distribution génotypique et allélique d’un polymorphisme entre cas et témoins. Ainsi que d’estimer le risque de maladie associé à la présence d’un allèle ou d’un génotype (même si les effets sont restreints) et de rechercher des interactions gène-gène ou gène-environnement.