UE11 BIOMEDECINE QUANTITATIVE
cours du 14/03/13
par Dr Cédric Laouénan
RT : ZELMAT Lillia
RL : MOUROUGANE Brigitte

COUR n° 6 :
ANALYSE GENOMIQUE

Le prof a essentiellement lu ses diapos qui sont d'ailleurs très bien foutues. Donc j'ai seulement rajouté des petits commentaires en italique afin d'un peu plus vulgariser le contenu.
Le cours se divise en deux parties complémentaires: la première qui introduit quelques notions et la deuxième qui porte sur l'épidémiologie génétique
Je vous souhaite donc une très bonne lecture...

SOMMAIRE :

I) TECHNIQUES D'ANALYSE DE LA BIOLOGIE MODERNE

Grâce aux nouvelles techniques de biologie moderne, de plus en plus de pathologies peuvent être suivies avec la mesure de biomarqueurs

Biomarqueur : **élément quantifiable qui va permettre de renseigner sur un effet biologique, physiologique ou une susceptibilité à une pathologie.**

2 types de biomarqueurs
–**biomarqueur prédictif : prévoir l'efficacité d'une thérapeutique** (*un polymorphisme au niveau de l'ADN peut-être associé à une meilleure réponse à un traitement ou pas; d'où l'intérêt de génotyper le patient avant traitement afin de savoir s'il est porteur d'une mutation qui le rendra plus ou moins sensible au traitement)*

–**biomarqueur pronostic : caractériser l'évolution de la maladie (survie des patients)** *certains polymorphismes au niveau de l'ADN sont associés ou non à une évolution plus ou moins rapide de la maladie.*

Pourquoi parle-t-on de Génome, Transcriptome ou encore de Protéome ?

Auparavant on ne s'intéressait qu'à l'analyse du génome parce qu'on partait du postulat :
1 gène = 1 ARN = 1 protéine → *donc en étudiant le génome, on comprends tout ce qui se passe par la suite.* Or cette relation est fausse :
Le génome code pour 20.000 gènes, il y a plus d'un million d'ARNm et plus de 10 millions de protéines → *il y a dès lors nécessité d'analyser en plus du génome, le transcriptome et le protéome car la diversité de ARN et des protéines peut expliquer certaines maladies même si au niveau du génome il n'y a pas de modification de la séquence d'ADN*

Il y a donc des **Biomarqueurs à différentes échelles :**
ADN → étude génomique
ARN → étude transcriptomique
Protéine → étude Protéomique

A) Le génome

=Ensemble du matériel génétique d'un individu

-Support de l'information génétique, composé de près de trois milliards de paires de bases, réparties entre les 23 paires de chromosomes

-Le même pour toutes les cellules d'un même organisme

-20 000 gènes codant chacun pour une ou plusieurs protéines (correspond à +/- 2 % de l'ADN)

-2 « sortes » d'ADN : **ADN codant et non codant** (= ADN poubelle car non traduit en protéines) :
-Projet ENCODE lancé en 2003 (ENCyclopaedia Of DNA Elements) → séquençage de tout l'ADN non codant : 2 résultats à ce projet

1/mise en évidence de plusieurs millions **de séquences « interrupteurs » régulant l'activité des gènes et des mutations** dans ces régions *(peuvent induire des maladies)*

2/une partie de cet ADN non codant est transcrit en **ARN (non codant)** le reste donne des **pseudogènes (non transcrits)**

Cl de ce projet → **80 % du génome humain est fonctionnel** au lieu de 2 % car au sein de l' ADN non codant on retrouve ces séquences de gènes (interrupteurs) qui vont régulé l'expression d'autres gènes.
Cmt ? Des zone d'ADN qui sont éloignées au niveau de la séquence se retrouvent très proches avec la structure 3D de l'ADN (histones,...) → une région d'ADN non codant peut alors réguler un gène situé juste à coté de lui.

On a **une diversité génétique = le polymorphisme ie des variations de la séquence d'ADN**
-Cette diversité naît d'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 . *La mutation a un avantage sélectif et est transmise, se retrouve fixée ie présente dans les générations suivantes.*
-Quand une mutation dépasse **les 1% de la population**, on ne parle plus de mutation, mais on parle de polymorphisme

Rappel de quelques définitions pour appréhender cette notion de polymorphisme :

1/**Locus** : une position du génome
2/**Allèle** : chacune des variations possibles de l'ADN en un locus du génome
3/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
4/**Haplotype** : combinaison de plusieurs allèles situés sur des locus différents d'un même chromosome. *L'haplotype est transmit dans son ensemble au générations suivantes sauf si recombinaison au milieu du chromosome qui détruit l'haplotype.*
5/ **Mutations de l'ADN amenant à la formation de polymorphismes :**
-**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 séquences ne diffèrent que d'une seul base, on se retrouve avec un polymorphisme bi-allèlique .*
Les SNPs apparaissent en moyenne une fois tous les 2000 bases → **3 millions de SNPs** ont aujourd'hui été recensés dans le projet HapMap visant à cartographier la variabilité du génome humain.

-**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*, à chaque nombre de répétition, on crée un allèle différent*). Elles peuvent être des :
→ 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 *(souvent les bases C-A)*
…

Quelle est l'impact de tous ces polymorphismes sur le génome ?

-**Polymorphismes non fonctionnels** : aucune conséquence
-**Polymorphismes fonctionnels** : sont responsables :

*Modification de l'expression du gène
*Modification de l'épissage de l'ARN
*Changement de la stabilité de l'ARNm
*Changement d'AA

B) Le transcriptome

=Ensemble des ARN issus de la transcription
-Variabilité du transcriptome par :
*différenciation cellulaire
*stimuli environnementaux
-2 types d'ARNs :
***ARN codant (ARNm)**
***ARN non codant (microARN)** qui **régulent l'expression des ARNm via des régulations épigéniques**, sont impliqués dans l'apparition de nombreuses pathologies. On dénombre env 1000 microARN.

C)Le protéome :

=**Caractérisation et étude de l'ensemble des protéines exprimées par un génome et plus particulièrement celui d'une cellule ou d'un tissu** (car les protéines sont différentes en fonction du tissu et de la cellule qui les produit)
Pour avoir une protéines fonctionnelle, **il faut des étapes de: transcription, traduction et modifications post -traductionnelles**. *Donc l'étude du protéome prend en compte toutes ces étapes de « fabrication » qui contribuent à la variabilité du protéome.*

D)Le Métabolome

Déclinaison des trois concepts précédents au niv des métabolites :
=**Étude de l'ensemble des métabolites (sucres, acides aminés, acides gras...) présents dans une cellule, un organe ou un organisme.**
Par exemple, on va essayer de distinguer, grâce au métabolome, un foie malade d'un foie sain ou le foie gauche du foie droit.

En CL : **Les « omiques » (génomique, transcriptomique,...) permettent d'aborder la complexité du vivant dans son ensemble. Ces approches « omiques » sont particulièrement utiles pour mieux connaître les maladies héréditaires et adapter les traitements au profil génétique (*possibilité de faire des « cartes » et ainsi caractériser un ind malade et non-malade en fonction de son génome, transcriptome,...)***

E)Mesure du génome et du trancriptome par biopuces (microarrays)

→ **Analyse des séquences d'ADN ou d'ARN avec un très haut débit.**
Le principe de ces biopuces est l'utilisation de puces sur lesquelles sont fixé des sondes et où on applique l'ADN ou l'ARN d'un ind pour connaître son génome ou son transcriptome.

Il y a différents types de puces : *!!! Ne retenir que les 2 premières !!!*

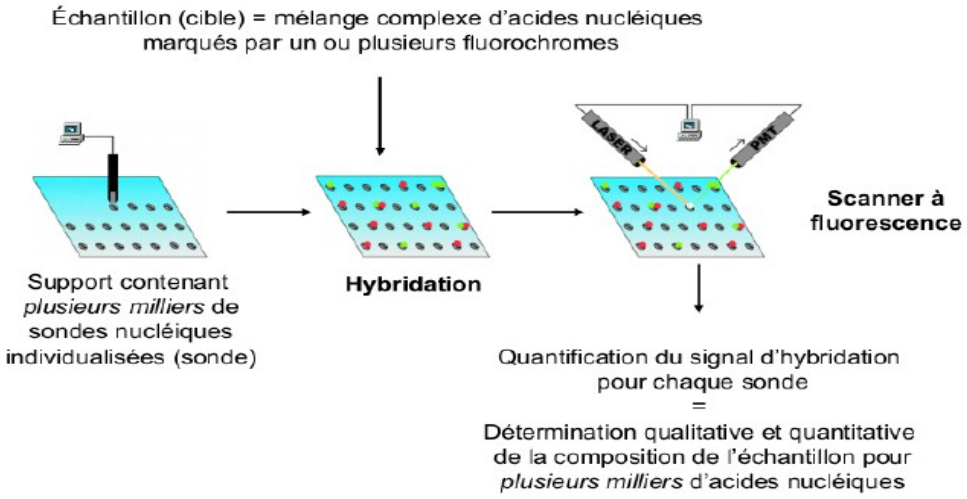
Type de puce	Cible	Objectif
Expression	transcrits	Mesure de l'expression des gènes
Génotypage	SNP	Génotypage des SNPs
Comparative Genomic Hybridation (CGH)	ADN	Repérage des duplications anormales de l'ADN
Méthylation	CpG islands	Mesure de la méthylation de l'ADN
Chromatin Immuno Précipitation (ChIP)	sites de fixation	Repérage des sites de fixation d'une protéine

Elles **détectent et quantifient simultanément** plusieurs centaines de milliers de séquences.
Elles 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 d'hybridation : *on a le mono-brin d'ADN du sujet d'un coté et la sonde d'hybridation de l'autre → si il y a une différence (un SNP) la liaison entre les deux brins est diminuée.*

1/ Les puces à expression

Ciblent **l'ARN** et leur objectif est **de mesurer l'expression des gènes (ARN sur ou sous-exprimé?)**
→ *on cherche à savoir si le mécanisme physio-pathologique de la maladie passe par une modification de la quantité d'ARN*
Les étapes de fonctionnement :



a. 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
b. 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
c. Les ADN sont alors hybridés sur la puce et la puce est lavée
d. Un fluorochrome Cy3 est appliqué sur la puce
e. 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

2/ Les puces à génotypage (+++ épidémio-génétique)

Cibles **les SNPs (ADN)** et servent au **repérage et au génotypage des SNPs** → *Quel SNP(s) est associé(s) à la maladie ? Et on regarde ce qui se passe à coté :*

-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 *ie on peut détecter la présence ou non chez un ind de 2,5 millions de SNPs sachant après que ces SNPs sont associés à des maladies ou pas*

-Les stratégies de choix des SNPs cherchent à établir **une couverture globale et fine du génome entier.**

II) ÉPIDEMIOLOGIE GENETIQUE :

Toutes les notions abordées ci-dessus servent à la réalisation d'études épidémiologiques.
L'épidémiologie génétique part **du principe que la plupart des maladies humaines présentent une composante génétique plus ou moins forte.**
Donc Épidémiologie génétique = **description et la compréhension de ces facteurs génétiques**→*savoir si une maladie a un facteur génétique et dans quelle mesure celui-ci est impliqué.*

On considère deux types de maladie :

1/Les maladies monogéniques ou mendéliennes :

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

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

une mutation ponctuelle entraîne une protéine anormale voire absente et le tout entraîne une maladie ; maladie héréditaire

***2/*Les maladies multifactorielles ou complexes :**

–**Composante génétique et environnementale** sont intriquées et la maladie résulte 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 → *ces allèles déclencheront la maladie que si il y a des interactions avec des facteurs de l'environnement. Donc la mutation n'est plus nécessaire et suffisante pour provoquer la pathologie.*

A) Les maladies monogéniques

Fréquence en population : rare
Agrégation familiale : élevée *(de nombreux cas dans la famille)*
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)
Effet de l'environnement : faible

B) Les maladies factorielles

=Maladies à hérédité complexe ou polygéniques (combinaison de plusieurs gènes mutés)
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

C)Principes généraux de l'épidémiologie génétique

Il existe 2 stratégies de recherche de gènes responsables de maladie. Ces deux méthodes peuvent être utilisée aujourd'hui. Les stratégies de recherche utilisées dépendent à la fois de la pathologie étudiée et des nouvelles technologies disponibles

1/**gène candidat** = gènes **connus et potentiellement impliqués** dans l'étiologie de la maladie étudiée → *Il n'y a pas d'analyse de l'ensemble du génome mais d'une fraction qu'on suppose responsable de la maladie et on regarde si ces gènes sont retrouvés mutés chez les malades*

2/**études 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 *expliqué par la suite***) → *on recherche un polymorphisme dans l'ensemble du génome (pas d'hypothèse d'implication au départ) et on regarde ensuite s'il est associé à la pathologie, ie s'il est plus fréquent chez les ind atteints par / au ind sains.*

3 étapes au raisonnement de l'épidémio génétique : *en gros comment recherche-t-on la composante génétiques de la maladie ?*

1/Montrer l'existence d'une concentration familiale de cas (agrégation familiale) de la maladie
–> **études épidémiologiques classiques**
2/Montrer que **cette agrégation est due à une composante génétique et la caractériser** : *cette concentration familiale n'est pas due à l'environnement ou à des habitudes de vie*
–> **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**

Détaillons ces étapes :

1/Montrer l'existence d'une agrégation familiale :

-Excès de cas familiaux = **agrégation familiale**
-Montrer que **chez les apparentés du premier degré** (parents, frères/sœurs ou enfants) des malades, la maladie est plus fréquente que :

*dans la population générale

*chez les apparentés du premier degré de sujets sains (témoins)

Or la concentration familiale peut être due à :

-une **transmission socioculturelle**

-une **exposition environnementale commune aux différents membres de la famille**

-une **susceptibilité génétique** *(ce qu'on chercher à démontrer)*

Ex : la schizophrénie :

	Incidence of Schizophrenia %
General population	0.8
Parents of affected	4.4*
Siblings of affected	8.5
Children of affected	12.3

L'incidence de la S . dans la population générale est de 0,8 % alors que :
-chez les parents d'un patient atteint, l'incidence passe à 4,4 %
-chez les frères/ sœurs, l'incidence est de 8,5 %
-chez les enfants d'un schizophrène, l'incidence est de 12,3%
-------> mise en évidence d'une agrégation familiale

Attention pour rechercher l'existence d'une agrégation familiale ie concentration de cas au sein des familles il faut prendre en compte les FACTEURS DE CONFUSIONS.

2/Mise en évidence de l'existence de la composante génétique

a) Les études de jumeaux

On regarde la maladie chez des jumeaux :

-**MONOZYGOTE** (MZ) :partagent le même patrimoine génétique
-**DIZYGOTE** (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 → *s'ils vivent ensemble qu'ils soient MZ ou DZ, on considère qu'ils partagent le même environnement*

Donc au sein d'une même paire de jumeaux on peut avoir :

	Jumeau 1	Jumeau 2
Concordant	Sain	Sain
Discordant	Malade	Sain
Discordant	Sain	Malade
Concordant	Malade	Malade

Les paires sont concordantes par rapport à la maladie si les deux jumeaux sont atteints. Elles sont discordante, si seul un des deux jumeaux est atteint → il a donc 4 cas de figures possibles

On calcul **le taux de concordance** ie le taux, le nombre de paires concordantes divisé par le nombre total de paires de jumeaux étudiées

Le principe de l'étude:

-MZ : la discordance (1 jumeau malade, 1 jumeau sain) ne peut être **que d'origine environnementale** *puisque les jumeaux possèdent le même patrimoine*

-DZ : la discordance **peut être d'origine génétique et/ou environnementale**, *car seul 50% de leur patrimoine est en commun*
→ **on compare donc le taux de concordance pour la maladie entre les paires MZ et DZ**

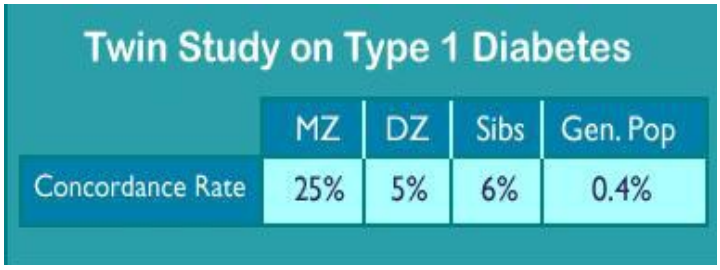
Interprétation :

1/Si une maladie est **totalement génétiquement déterminée :**
–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
2/Si une maladie **n'a rien de génétique :**
–La concordance au sein des paires MZ devrait être égale à la concordance des paires de DZ *(même environnement pour les MZ et DZ)*
3/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

Ex : *Diabète de type 1*



→ *On voit bien que la maladie a une composante génétique.*

b) Les études d'enfants adoptés (+++année 50-70)

Principe :

L'adoption sépare les enfants de leurs parents biologiques
Un enfant adopté :
–Par **rapport à ses parents adoptifs :**
- •Partage une partie de son environnement
- •Peu de similarités génétiques

–Par **rapport à ses parents biologiques :**
- •Partage en moyenne, 50% de leur patrimoine génétique
- •A peu partagé son environnement

Condition :

-Les enfants doivent **avoir été abandonnés très jeunes** (si possible dès la naissance) → *pas d'interactions d'environnement entre enfants adoptés et parents biologiques*
-Les enfants doivent **avoir été placés très tôt dans leur famille adoptive** → *pour partager un maximum d'environnement avec les parents adoptifs*
-Les enfants doivent **avoir été placés au hasard**
-**Les données sur les parents biologiques et adoptifs doivent être complètes**

Interprétation :

1/**Si trait est génétiquement déterminé** : On s'attend à retrouver **plus fréquemment la maladie dans la famille biologique de l'enfant adopté**, *l'enfant ressemblera d'avantage à ses parents biologiques qu'adoptifs*

*2/***Si une maladie n'a rien de génétique** : Si elle est totalement **sous l'influence de facteurs**

environnementaux, *l'enfant ressemblera autant à ses parents adoptifs qu'à ses parents biologiques*

Limites :

Enquêtes très difficiles et quasiment impossibles en France car elles nécessitent dans l'idéal :
-de connaître les parents adoptifs
-de connaître les parents biologiques
-de recenser des enfants adoptés

c)Analyse de ségrégation :

vise à répondre à 2 questions :

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

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

Objectifs :

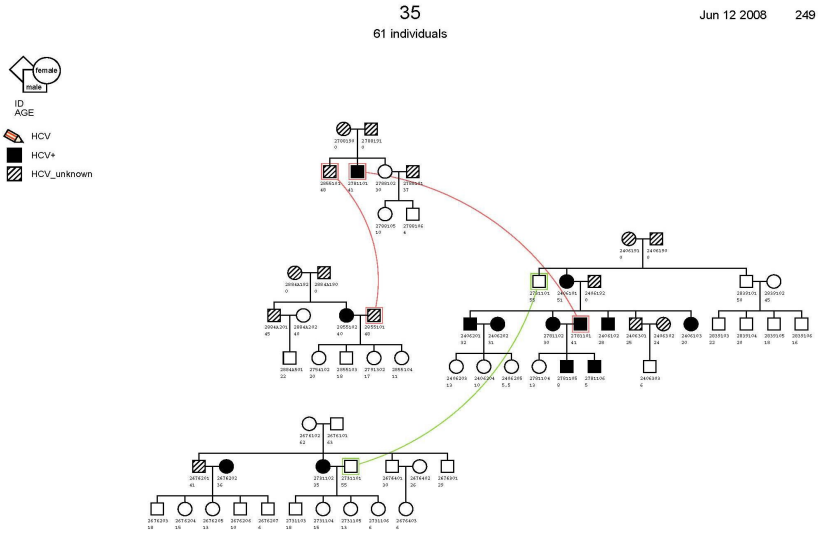
-**Déterminer le modèle génétique** qui explique le mieux les distributions familiales d'une maladie (maladie monogéniques+++)
-Cherche à mettre en évidence **l'effet d'un gène transmis de façon mendélienne** (classiquement appelé gène majeur ou à effet fort) parmi l'ensemble des facteurs génétiques 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 → ***utilisée dans le cadre de maladie mendéliennes***

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

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

Ex :



Sélection d'un pedigree (famille) de 61 ind où il y a des cas d'hépatite C. On reconstitue les liens de famille entre les différents ind et on voit bien qu'il y a une agrégation familiale et après une étude de ségrégation (étude statistique grâce à des logiciels) → mise en évidence d'un gène de susceptibilité à l'hépatite C à transmission plutôt autosomique dominant .

3/ Analyse de la composante génétique : Analyse de liaison / Étude 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**

a)Analyse de liaison génétique (sur données familiales)

But :

Localiser le(s) gène(s) responsable(s) de la maladie sur le génome soit par :

–**Analyse de régions candidates** :

*__Hypothèse sur le gène en cause :__ on regarde les zones polymorphiques considéré comme potentiellement impliquées puis

*__Confirmer ou infirmer une liaison génétique__ d'une région avec le phénotype étudié

–**Recherche sur l'ensemble du génome** = **pas d'hypothèse a priori** sur le gène en cause

Objectif :

-On cherche à savoir si **la maladie et des allèles de marqueurs coségrègent dans les familles** : *est-ce que un SNP est associé à la présence de la maladie ?*

-Pour cela on utilise des **familles recensées par un sujet malade avec au moins 2 germains**

-L'analyse de liaison génétique repose sur la notion **de Déséquilibre de liaison gamétique/ génétique :** 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) .** Cependant, 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 (exprimé en centimorgans)**

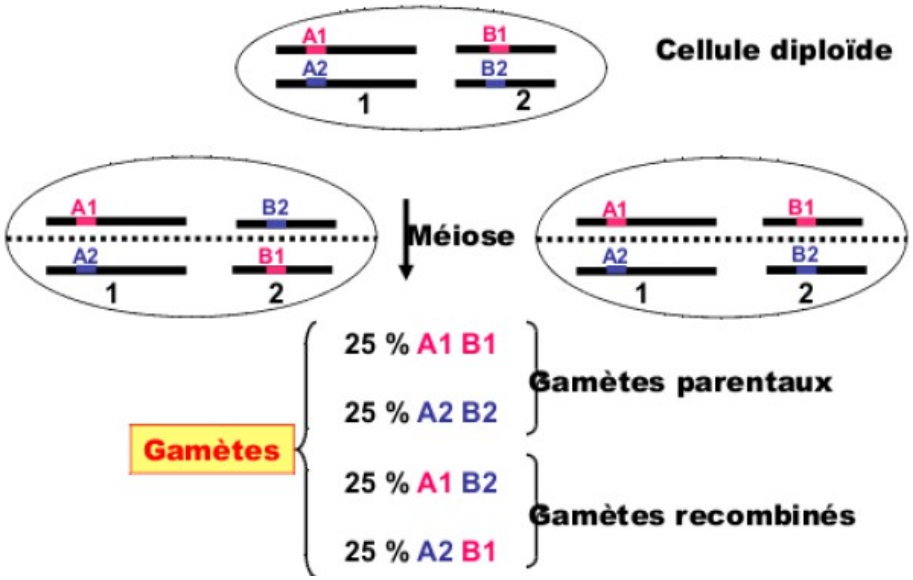
→ CALCUL DU TAUX DE RECOMBINAISON :

-si faible alors le SNP et le gène morbide sont proches

-si très très faible alors le SNP ne se trouve qu'à quelques paires de bases du gène morbide

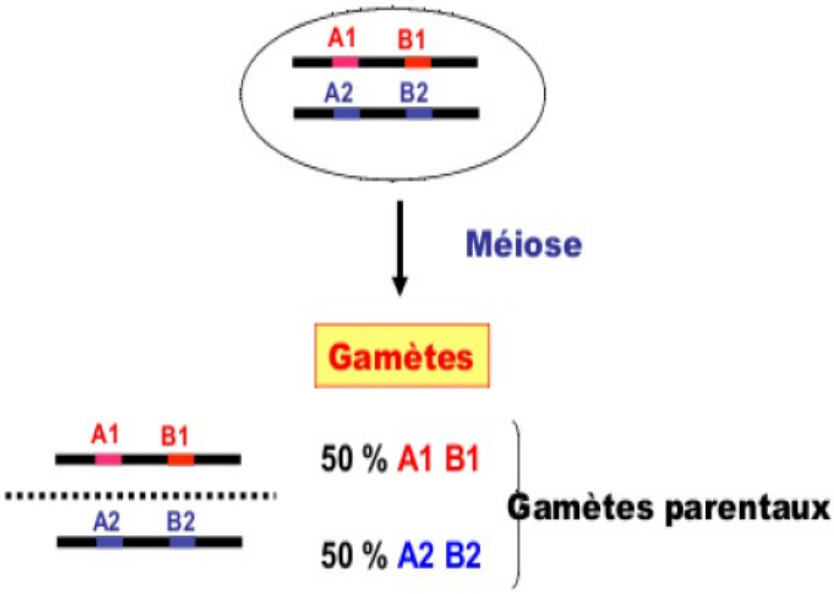
Deux cas de figures se présentent :

1/ **les deux locus sont indépendants** (situés sur deux chromosomes différents)



les gènes A et B sont répartis de façon **indépendante** donc on aura lors de la méiose des **gamètes parentaux** (25% X 2) mais aussi **recombinés** (25% X2)

2/ **En cas de déséquilibre de liaison gamétique** (locus séparé par une petite distance génétique)

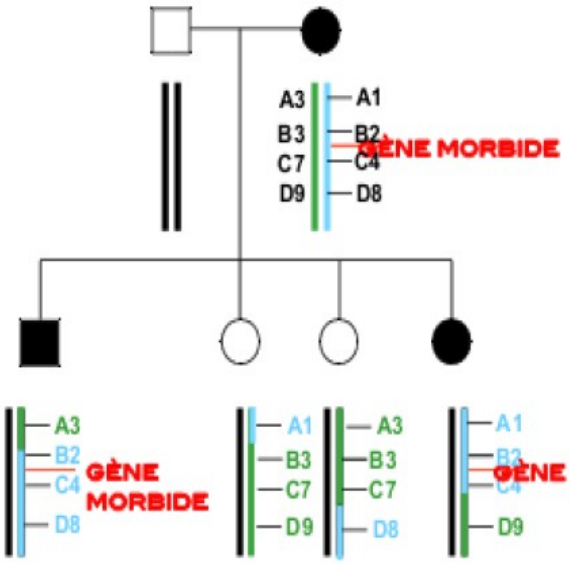


Si les deux locus sont **totalement liés génétiquement** alors on obtient **uniquement des gamètes parentaux et aucunes gamètes recombinées** (sauf si cross-over).

Ainsi, en fonction du nombre de gamètes recombinées obtenues, on évalue la distance entre deux locus : **+ il y a de gamètes recombinées et + la distance génétique est importante .** *Avec cette analyse, on se connaît pas le gène morbide mais on peut estimer les marqueurs ie les SNPs proches de celui-ci : la présence de ces SNPs signent dès lors la présence du gène morbide, ils deviennent associés, marqueurs de la pathologie.*

<u>Objectifs de l'étude de liaison :</u>

Analyser **la ségrégation conjointe de deux locus d'une génération à la suivante comportant plusieurs sujets atteints →** *puisque si les deux locus sont situés à proximité l'un de l'autre alors ils seront transmis simultanément aux générations suivantes*



Par ex : ici on peut faire l'hypothèse que les allèles B2 et C4 sont liés génétiquement au gène morbide

*

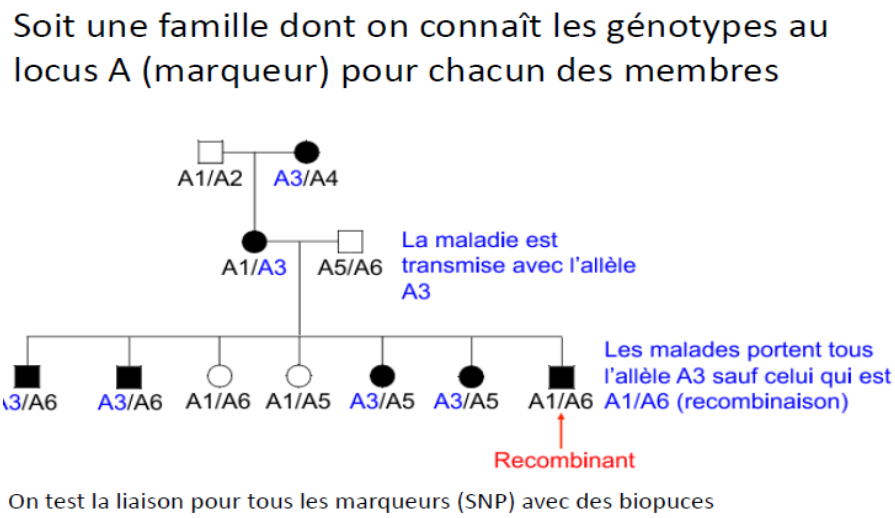
Les résultats de cette études permettent :

-**Montrer l'existence d'une liaison** (transmission simultanée) entre un marqueur et le gène morbide

-**L'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**

Exemple :



A noter que pour ce genre d'étude sachant qu'on teste les 2,5 millions de SNPs on ne peut pas se contenter d'un degré de signification de 5% → des techniques statistiques corrigent ce seuil qui est abaissé à 10^-27

b)Études d'association ou GWAS (sur données de population)

Principe :

-Concernent des populations **de sujets indépendants (sans lien familial)**

-Font appel **aux techniques classiques de l'épidémiologie**

-Le plus souvent on utilise **des études Cas-Témoins** → On fait un test du χ^2 et on mesure l'association en calculant **des Odd Ratio** *(sachant que le prof a dit qu'il ne demanderai jamais pour son cours de faire un test du χ^2 mais qu'on était sensé le connaître et notamment pour le cours n°1)*

-**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 (un SNP) 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**

Objectifs :

-**Comparer la distribution génotypique et allélique d'un polymorphisme entre cas et témoins**
- Permettent :

–> **D'estimer le risque de maladie associé à la présence d'un allèle ou d'un génotype** (même si les effets sont restreints)

–>De rechercher **des interactions gène-gène ou gène-environnement** : *l'avantage des GWAS est qu'ils sont moins puissants que les études familiales pour les maladies monogéniques mais très utiles pour l'analyse des maladies multifactorielles*

-Reposent sur :

–>Des échantillons de population

–>Des polymorphismes (les SNPs)

–>Le déséquilibre de liaison

–>Des stratégies de sélection du ou des gène(s) à étudier (soit stratégie gène-candidat ou étude génome entier)

La principale limite

la sélection des échantillons :

-Les Cas :

-Comment définir la maladie ?

-Méthodes de recrutement ? → Cas incidents ou prévalents ? Cas hospitaliers, en ville ou sélectionnés à partir de registres ? Cas consécutifs ou tirage au sort ?

-Les Témoins :

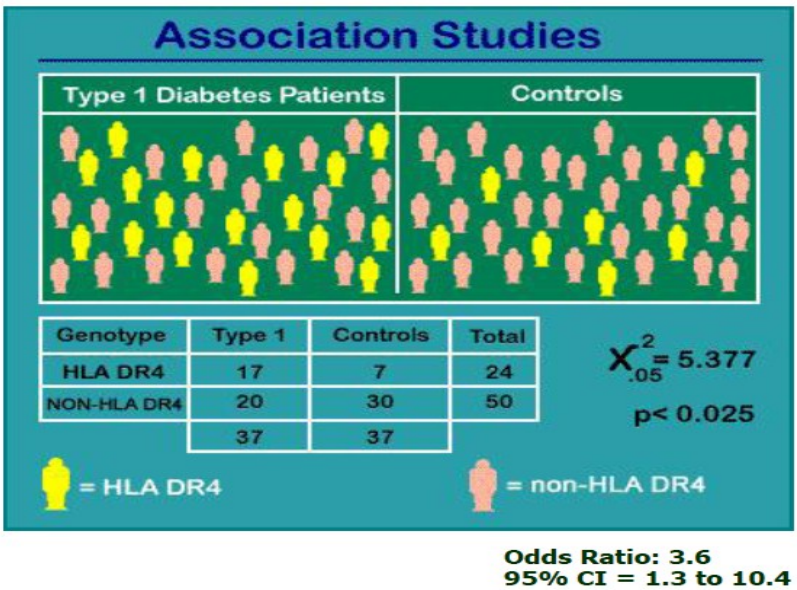
-Comment s'est-on assurer de l'absence de maladie ?

-Méthodes de recrutement ? → population générale ? Voisins, époux ? Témoins hospitaliers ?

-Sont-ils représentatifs de la population dont sont issus les cas ?

Les études d'association ne sont ni plus ni moins que des test de χ^2 que l'on reproduit pour chacun des 2,5 millions de SPNs

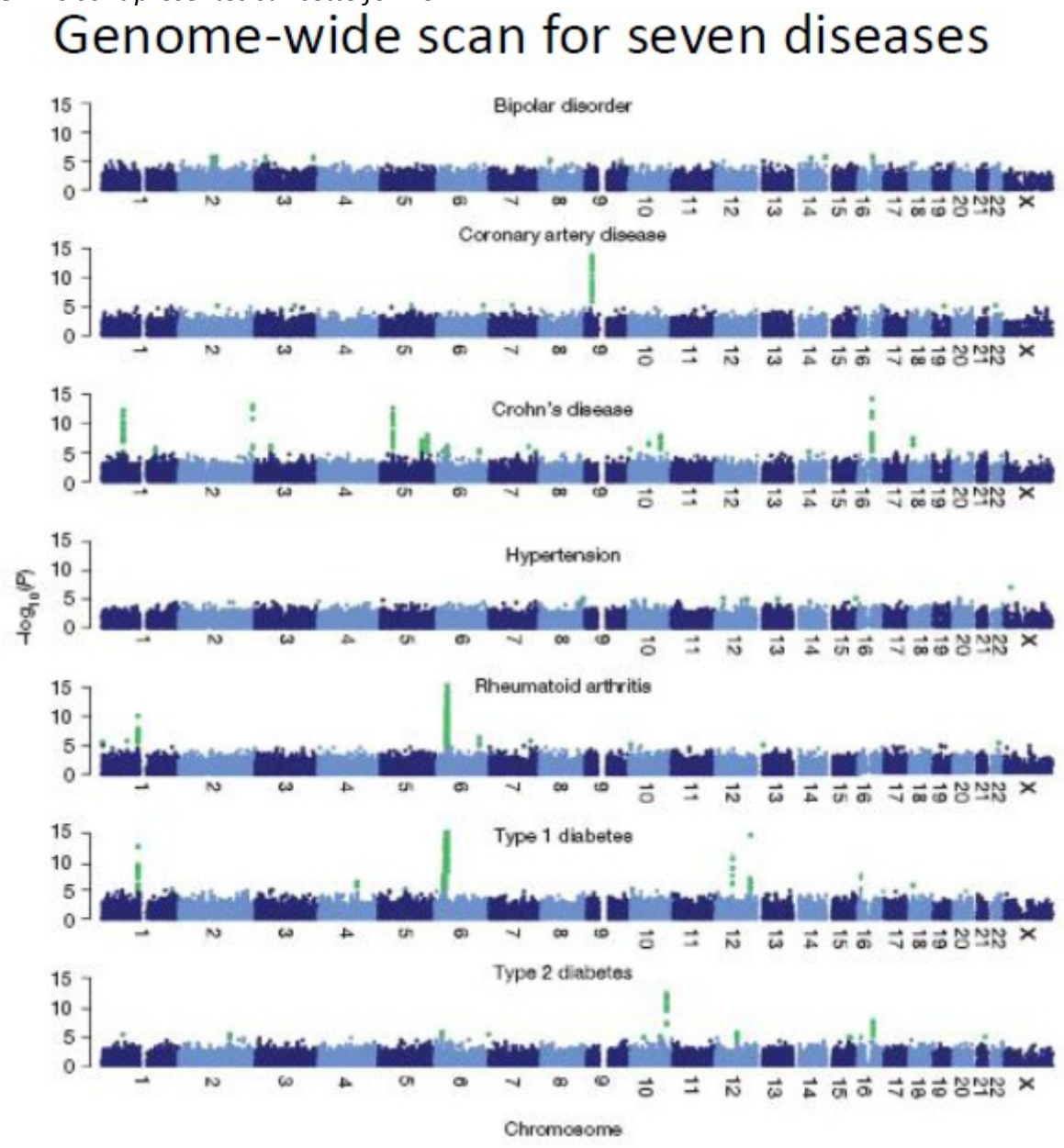
Ex :



On s'intéresse ici à l'association entre être atteint du diabète de type 1 et le fait de porter dans son génome l'allèle HLA DRA : On génotype les cas et le témoins et on constate que 17 cas sur les 37 sélectionnés porte l'allèle HLA DR4 alors que seulement 7 témoins sur les 37 sélectionnés porte l'allèle → peut-être que le D de type 1 et le génotype HLA DR4 sont associé → on fait un χ^2 qui montre qu' il y a bien

association et on calcule OR = 3,6 donc il y a ***3,6 fois plus de risque d'être atteint du D type 1 si on est porteur du génotype HLA DR4.*** *Et on refait ça pour tous les SNPs possibles et inimaginable...*

Les GWAS sont présentés sur cette forme :



On étudie 7 pathologies différentes. En abscisse sont représentés les chromosomes : ce qui faut retenir c'est qu'on regarde s'il y a chez les cas et le témoins des zones qui sont associé à la pathologie. Chaque point représente la puce d'un patient pour un SNPs donné → on regarde s'il existe des signaux ie si chez les malades il a beaucoup plus de SNPs présents par / aux témoins (les points verts qui se démarquent des autres)
donc par ex pour les coronopathies on remarque un pic au niveau du chromosome 9 → il doit y avoir un SNP associé à la maladie et ainsi de suite...