UE11 – Parcours physiologie et pharmacologie des transmissions - Cours n°11 10/04/2019 alexandre.alcais@inserm.fr	RT : Aurélie Khouri Yasmine Niati RL : Priscille Heude

Epidémiologie génétique (2)

Plan :

I. Introduction

II. Etudes sans ADN

A. Variabilité inter individuelle
B. Concentration familiale
C. Etudes de jumeaux
D. Etudes d'adoptés
E. Etudes de ségrégation

III. Etudes avec ADN

A. Analyse de liaison
B. Analyse d'association

IV. Génotypage et séquençage

A. Génotypage
B. Séquençage
C. QCM

Abréviations :
Pb : paires de base
GWAS : Genome Wide Association Study

Mot du RT :
A la fin du cours, je vous ai mis 4 QCM types corrigés par le prof qui peuvent tomber aux partiels.

I. Introduction

La génétique épidémiologique démarre à partir d'une idée formalisée sous la forme d'une **hypothèse**.
L'hypothèse de départ qui nous intéresse est que le génotype de l'autre joue un rôle dans notre génotype, ce qui implique une **variabilité du phénotype** que l'on va essayer de relier au génome. Il faut donc poser des hypothèses. Les arguments en faveur de la génétique sont du type : « des cas familiaux semblent plus concentrés au sein d'une même famille » ; « si un frère est atteint, le risque de transmettre est augmenté », etc.

Ho : la génomique de l'hôte n'influence pas le trait d'intérêt
H1 : la génomique de l'hôte influence le trait d'intérêt

Il n'y a pas de bonne ni de mauvaise idée (le prof l'a répété 5 fois), toutes sont recevables et sont à **mettre en lien avec les données**, et ces données vont valider ou invalider ces idées. Elles sont toutes bonnes tant qu'elles sont émises par l'esprit humain (les machines ont un raisonnement contre intuitif qui ne peut pas remplacer la recherche humaine). Ces hypothèses peuvent ensuite être vérifiées grâce à l'abondance d'information (Big data). Elles sont à tester en utilisant le design optimal (puissance, efficacité…).

La dernière fois, on s'est intéressé aux hypothèses portant sur la **relation génétique** (ne s'arrête pas seulement aux marqueurs de l'ADN mais concerne également tout ce qui joue sur l'ADN, les différents variants, l'épigénétique… contrairement au mot « génomique » qui est plus restreint)**/phénotype**. On s'est également penché sur les hypothèses portant sur le **degré d'homogénéité.**

L'idée des études GWAS est la suivante : beaucoup d'analyses génétiques sont des études **cas-témoins** où l'on prend 1000 cas (par ex) et sur ces 1000 cas, certains vont avoir des effets indésirables au traitement testé et d'autres non. On va génotyper ces **variants alléliques** (s'oppose au séquençage) c'est-à-dire qu'on définit ce qu'on va voir. On génotype un variant après l'autre pour couvrir tout le génome, on les teste chacun. On fait quand même l'hypothèse que des gens qui, par définition de la construction de l'étude, sont indépendants, partagent exactement le même variant, à la même position de leur génome, c'est hallucinant (au sein d'une famille, c'est potentiellement acceptable, mais entre des individus qui n'ont rien à voir entre eux, c'est hallucinant !).

En faisant ces études, on part du principe que les n personnes indépendantes possèdent le variant d'intérêt. De plus, en testant les variants un par un, on sous-entend que tous les variants du génome sont indépendants du reste du génome. Ça veut dire que si on fait ça, on pourrait faire les 3 milliards de paires de bases, si on nous mettait 3 milliards de chromosomes qui contiennent chacun 1 base, on devrait avoir la même chose qui se produit chez les individus. Mais ces études ne prennent pas en compte les loci, les séquences (nucléotides…) : c'est comme si on testait 10 millions de chromosomes ; ça n'a aucun sens mais c'est ce qui est fait.
Ce ne sont pas des hypothèses techniques ce sont des **risques** qu'on est près à prendre en faisant l'étude. (Risque des 5% hors de l'intervalle en statistiques, le risque d'observer une différence aléatoire).

Tests des hypothèses :

On est dans un monde où **la causalité n'est pas accessible chez l'être humain**, on se contente donc d'interpréter les résultats comme une causalité. La seule solution à la causalité ce sont les **cohortes** d'1 million 500 000 personnes qu'il faut suivre avec des contrôles réguliers.

Maintenant, il faut trouver quels **tests** existent pour démontrer ou être en faveur de l'hypothèse qu'il y a un rôle des facteurs génétiques de l'hôte dans votre phénotype d'intérêt. Il faut donc distinguer :
-**sans ADN/avec ADN**
-sans familles/avec familles

II- Etudes sans ADN

A) Variabilité interindividuelle

Pour parler de génétique, il faut qu'il y ait de la **variabilité entre les individus**. S'il n'y a pas de variabilité, il n'y a pas d'intérêt à aller chercher un quelconque rôle du génome de l'individu car dans ce cas-là, tout le monde est égal devant ce qui se passe et il n'y a rien à expliquer.
Dans la génétique épidémiologique, il faut donc des individus atteints, et des non atteints... L'information qu'on essaie de disséquer est la variabilité entre les individus.
Elle s'étudie lors d'études **sans familles**, chacun est indépendant des autres.

B) Concentration familiale

On observe la concentration familiale de l'étude avec des cas si on est sur un phénotype 0/1 ou des valeurs élevées s'il s'agit d'un phénotype quantitatif. On va essayer de quantifier ces **observations**, les cas appartiennent plus ou moins à des **familles** (3 cas appartiennent à la famille A, 5 cas à la famille B…) et on peut quantifier l'intensité de ces corrélations familiales par un indicateur :
le risque de récurrence familiale (= (prévalence chez individus ayant frère ou sœur atteint)/ (prévalence chez individus n'ayant pas un frère ou sœur atteint)
Autrement dit, on cherche le rapport de la prévalence chez la fratrie des cas et de la prévalence chez la fratrie des témoins.

Ex : on choisit 100 cas, et on regarde parmi les frères et sœurs s'il y a des atteints

Sur les 200 frères et sœurs de cas atteints, on en a 100 atteints : prévalence de la maladie chez les frères et sœurs des cas = 50%
Sur les 100 frères et sœurs de témoins, on en a 10 atteints : prévalence de la maladie chez les frères et sœurs des contrôles = 10%
Lambda s=0,5/0,1=5

Plus le risque lambda est élevé plus l'implication génétique est probable, mais il n'y a pas de lien de causalité car l'environnement et l'éducation peuvent changer entre les familles : cela quantifie juste la concentration familiale.
Exemple : un cluster de cas au sein d'une famille peut être dû à l'alimentation.
Cela s'étudie lors d'études **au sein de familles**.

C) Etudes de jumeaux

On va comparer **la concordance** pour un phénotype chez des **jumeaux monozygotes et dizygotes.** C'est un modèle très performant mais on a malheureusement peu de registres de jumeaux en France. Les jumeaux monozygotes dérivent du même œuf, donc ont le patrimoine génétique en commun à des mutations de novo près, tandis que les dizygotes sont des frères et sœurs normaux (50% du patrimoine génétique en commun), qui partagent la même vie. Si la génétique est impliquée, la concordance observée pour le phénotype concerné pour les jumeaux monozygotes est supérieure à celle des jumeaux dizygotes.

Etant donné qu'on travaille souvent sur des **traits plutôt rares**, on ne considère pas la concordance non atteint/non atteint car il y en aurait trop : ça n'est pas représentatif et ça absorberait tout (beaucoup de monozygotes non atteint/non atteint et beaucoup de dizygotes non atteint/non atteint). La concordance serait alors très élevée, sans pour autant qu'il y ait une implication génétique.

Ex : si on prend la lèpre, chez les monozygotes, on a 70% de deuxième jumeau atteint si le premier est atteint, et 15% de deuxième jumeau dizygote atteint si le premier est atteint. Il y a ici une contribution génétique mise en évidence.

Cela s'étudie lors d'études **au sein de familles**.

D) Etudes d'adoptés

Deux jumeaux monozygotes ayant un fond d'ADN identique mais un environnement différent (cela s'étend à tous les enfants). On regarde si les enfants adoptés ressemblent plus aux parents biologiques ou aux parents adoptifs. Dans les pays scandinaves, il y a des registres de jumeaux et des registres d'adoption très ouverts, on a accès à l'information des parents biologiques. Si pour un trait phénotypique, les enfants ressemblent plus à leurs parents adoptifs, c'est **environnemental,** et s'ils ressemblent plus aux parents biologiques, c'est plus en faveur d'une **origine génétique**.

Exemple : le cancer a majoritairement une cause environnementale tandis que pour les maladies cardio-vasculaire, l'origine est 50-50 entre l'environnement et la génétique, et pour les décès dus à une maladie infectieuse, c'est majoritairement d'origine génétique car il y a 5 fois plus d'enfants décédant d'une maladie infectieuse si un de ses parents biologiques est mort d'une maladie infectieuse que si son parent adoptif est mort d'une maladie infectieuse.
Cela s'étudie lors d'études **au sein de familles** également.

E) Etudes de ségrégation

On regarde la ségrégation familiale des cas : on recueille des **familles**, on regarde qui est atteint et qui ne l'est pas. On regarde si la ségrégation de la maladie/phénotype qui nous intéresse est compatible avec une contribution génétique. Est-ce que sur l'arbre généalogique, la répartition des sujets atteints fait penser à un type autosomique dominant, récessif… ?

On va expliquer le phénotype d'un individu par des **variables environnementales** (ex : co-variable 1 : tabac ; co-variable 2 : exposition solaire…) et des **variables génétiques** (gène a, b…)
Si tous les cas sont T+ (tous tabagiques) et tous les autres sont T-, si on rentre la variable tabac dans le modèle, phénotype = tabac et on explique tout, puisque tous ceux qui sont malades fument et les non malades ne fument pas.

Si certains fument sans être atteint, on recherche une autre variable (exposition solaire) et si ça n'explique pas encore tout, on recherche une troisième variable (un gène, un modèle environnemental…). Le problème est que le périmètre exploré est limité car on n'explore pas toutes les hypothèses mais seulement celles qu'on teste.
Ce sont des **études sans ADN** car on ne s'intéresse qu'à la **ségrégation des phénotypes au sein des familles.**

III- Etudes avec ADN

A) Analyse de liaison= locus+++

1. Le principe

L'hypothèse de départ est que le **génotype de l'autre joue un rôle dans notre génotype**, ce qui implique une variabilité du phénotype que l'on va essayer de relier au génome. Il faut donc poser des hypothèses.

Les arguments en faveur de la génétique sont du type : des cas familiaux qui semblent plus concentrés au sein d'une même famille ; si un frère est atteint, le risque de transmettre est-il augmenté ?, etc.
Les hypothèses initiales sont alors :

H0 : le génome de l'autre ne joue aucun rôle
H1 : le génome de l'autre joue un rôle

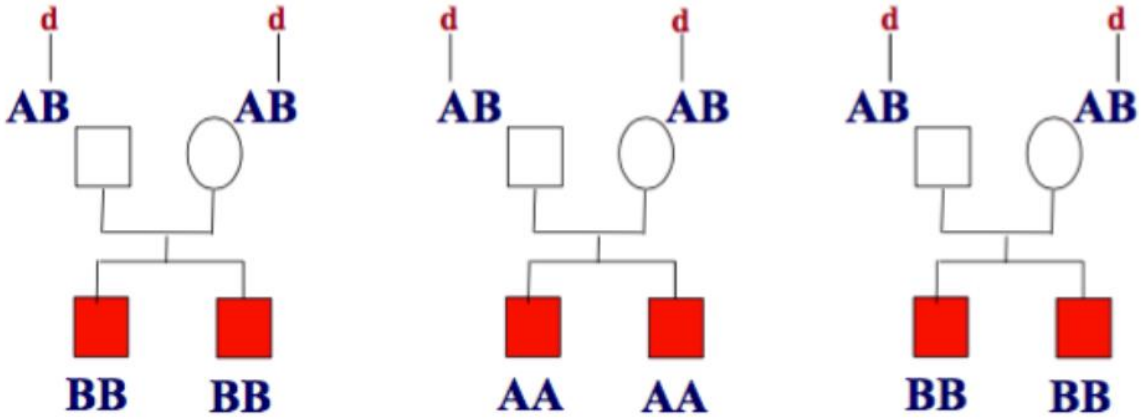
En injectant de l'ADN dans l'expérience, la dimension à explorer est restreinte de 3 milliards de pb à 30 millions de pb par **l'analyse de liaison** puis à 300 000 pb par le **GWAS**.

Cependant, on n'aura pas le variant causal de la pathologie, seulement des idées. Le génotypage utilise 1 million de variants parce que l'évolution de la connaissance du génome a permis de résumer les 60 millions de variations communes à 1 million. Le génotypage d'1 million de variants permet de dire que l'on fait un génome wild.
L'analyse de liaison est une première étape afin de diminuer l'espace que l'on veut étudier.
On ne fait plus vraiment d'études sans ADN parce que maintenant, on part du principe que le génome joue toujours un rôle.

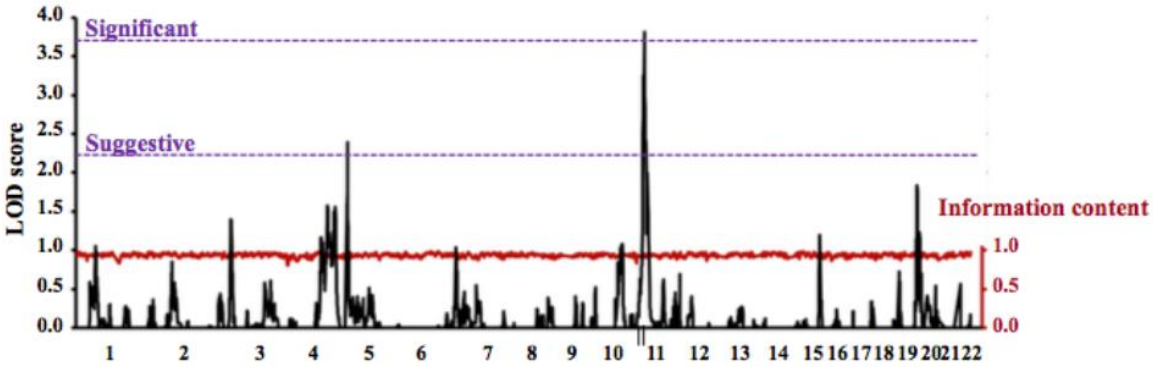
2. Étude sur les familles

L'étude de liaison est une étude qui nécessite des familles : frères et sœurs, cousins ... Sachant qu'il y a déjà **2 personnes malades dans la famille**, la probabilité que ces cas soient des cas génétiques est augmentée par rapport à des cas sporadiques.
On veut dans chaque famille au **moins 2 enfants atteints** de la pathologie étudiée. C'est une **donnée imposée**. On force les populations à se ressembler au point de vue du phénotype mais on ne peut pas forcer leur génome à se ressembler. Ayant des enfants qui ont le même phénotype, ont-ils une **région du génome où ils se ressemblent plus** que ne le voudrait le hasard ? Les frères et sœurs partagent 50% du patrimoine génétique, est-ce qu'en les forçant à se ressembler, la ressemblance est supérieure à ces 50% ?

Dans beaucoup de cas, ce n'est pas le cas, mais dans certaines régions du génome, oui.
Ces régions du génome sont plus candidates à contenir quelque chose qui pourrait développer le phénotype d'intérêt.
On ne prend pas les non-atteints parce qu'on ne sait jamais s'ils sont réellement non- atteint ou seulement pas encore, et qu'une erreur comme celle là a un coût très important dans l'analyse.
Pour chaque marqueur, on regarde s'il y a eu co-ségrégation avec le phénotype (transmission au niveau de ce marqueur des parents aux enfants avec la maladie) ou pas.



Ici, tous les parents sont hétérozygotes AB, ce qui est nécessaire pour l'expérience, sinon la co-ségrégation n'est pas possible.
Dans les première et troisième familles, l'allèle B est celui transmis aux enfants. Dans la seconde famille, c'est l'allèle A du marqueur. La probabilité d'avoir le **locus causal** est quasi nulle mais on ne devrait pas en être trop loin. Ce n'est pas l'allèle A ou l'allèle B du marqueur qui est la cause de la maladie, mais on peut supposer que dans les familles 1 et 3, **l'allèle causal doit être localisé sur le même chromosome** que l'allèle B du marqueur. De même avec l'allèle A pour la famille 2.



On s'attendait à ce que les frères et sœurs partagent 50% en commun et dans ce cas-là, ils sont à 100% du matériel génétique hérité en commun des parents.

On a en abscisse, les chromosomes, et en ordonnée le score. **Plus il est élevé**, plus c'est en faveur du fait que **cette région du génome co-ségrège de façon non-aléatoire** ; les enfants partagent une région du génome plus que ne le voudrait le hasard.

B. Études d'Association = Allèles +++

Elles permettent de **réduire encore le champ**.
On prend plein de cas et plein de témoins, si possible indépendants les uns des autres. On les génotype pour des variants. Dans le cas où il a été fait juste avant une étude de liaison, on prend des variants pris dans les gènes des chromosomes pour lesquels on a restreint l'étude.

1. Première méthode

On peut aussi génotyper tout le génome, comme pour la maladie d'Alzheimer, car il est plutôt compliqué d'avoir les parents, mais cela n'a pas beaucoup de spécificité génétique.
Par exemple, dans le cas du cancer du poumon, on regarde ce qui fait varier la probabilité de l'avoir : garçon/fille, alcoolique/non alcoolique, grand/petit, caucasien/non caucasien, fumeur/non-fumeur... À ça, on ajoute juste comme potentiel d'explication des variables qui sont des génotypes particuliers.

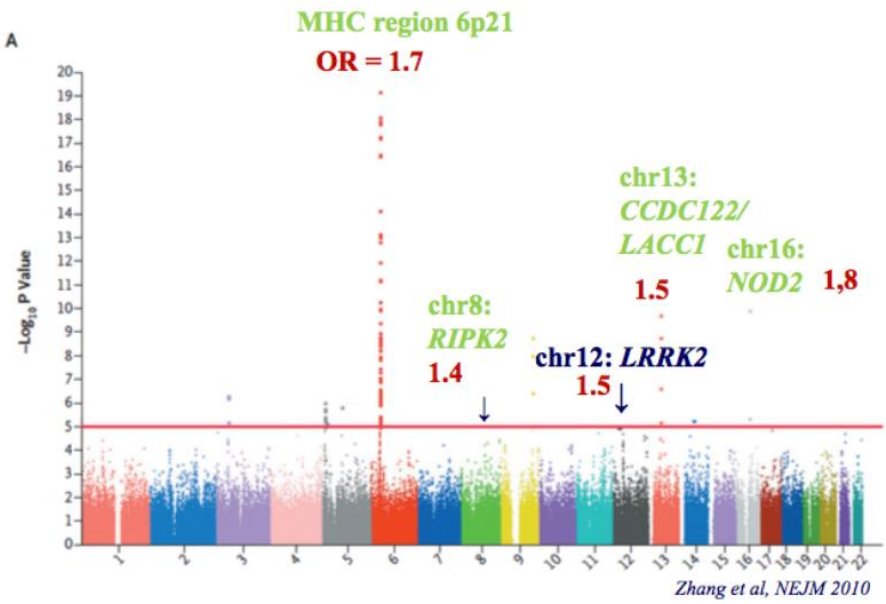
Dans le cas où **aucune étude de liaison n'ait été faite** avant, il peut être instructif de savoir si dans
la famille il y a déjà eu des cas. On peut ainsi restreindre l'analyse à ceux qui ont déjà eu d'autres cas dans la famille. Les cas sont donc plus génétiques.

	cases	controls	Odds Ratio	P-value
AA	100	200	1.00	
AB+BB	200	100	4.00	<0.001

Ici, il y a deux fois plus de cas AB+BB que de contrôles (B est considéré comme dominant ici).
L'effet ici est de 4, c'est-à-dire, si on est malade, on a 4 fois plus de risque d'être porteur de AB ou BB qu'un contrôle. L'étude étant rétrospective, on ne peut pas faire de risque relatif : on interprète ici **des odds ratio**.
L'odd ratio peut être interprété comme un risque relatif **uniquement si la prévalence de la pathologie est très petite**. On ne peut pas approximer un odd ratio en risque relatif dans le cas de
l'obésité en France parce que sa prévalence est beaucoup trop importante. L'odd ratio va surestimer le risque.

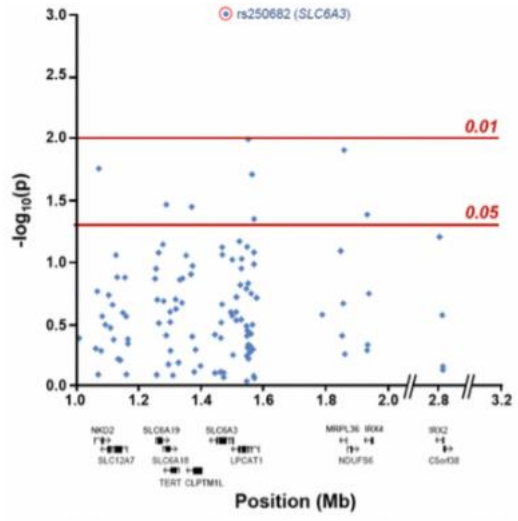
2. Deuxième méthode

On a fait une étude de liaison juste avant, on sait que l'on peut **se concentrer sur certaines régions**, que l'on va **saturer en variants**. C'est une approche **génocentrique**. On va, comme pour l'étude de liaison voir s'il y a un variant différent des autres qui pourrait expliquer la pathologie.



3. Troisième méthode

On ne passe pas par l'étape de ce qui est sans ADN ni pas l'étude de liaison et on va directement prendre 10 000 témoins, génotyper 1 million de marqueurs chez tout le monde, les **tester 1 par 1** et voir s'il y a une **surreprésentation d'un des allèles**, si c'est significativement différent entre les
cas et les témoins. Cela forme un **Manhattan Plot**. Il y a à la fois une mise en évidence d'association ainsi qu'une estimation de l'effet.
Ici, pour le chromosome 6, le risque de développer la lèpre est augmentée de 1,7.



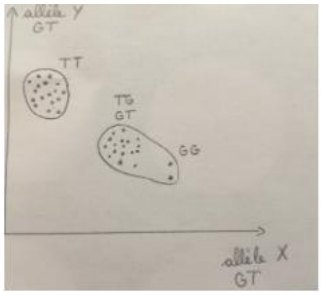
IV. Génotypage et séquençage

Le cours précédent s'était arrêté aux études sans ADN, les analyses de ségrégations (observations épidémiologiques). Ces analyses n'ont maintenant plus de réel intérêt si ce n'est méthodologique, on ne s'en sert plus du tout. Les autres études avec ADN sont les **analyses de liaison** et les **études d'association.**

S'il n'y a pas de variabilité, il n'y a pas d'intérêt à aller chercher un quelconque rôle du génome de l'individu car dans ce cas-là, tout le monde est égal devant ce qui se passe. Ainsi, il faudrait plutôt empêcher la menace, que de compenser avec une ligne de défense différente entre ceux qui résistent et ceux qui ne résistent pas. Dans la génétique épidémiologique, il faut donc des gens atteints, des gens non atteints... L'information qu'on essaie de disséquer est la **variabilité entre les individus.**
Les études de génétique ont commencé avec des **analyses sans ADN** car on ne savait que faire avec l'ADN, étant donné qu'il y avait très peu de marqueurs. À la fin des années 90, on disposait des
systèmes rhésus, ABO, des Ig, soit environ un marqueur par chromosome. Puis l'information génétique est devenue colossale, la technologie et la connaissance ont évolué. On sait alors que le **génome est redondant.**

A. Génotypage

Dans le génotypage, les puces interrogent 1 million de variants.
On s'intéresse à deux allèles, Y et X. L'information s'agrège, formant des petits groupes dans l'idéal. Ces groupes sont les groupes GT ou TG et TT.
L'algorithme d'attribution d'un génotype à une position donnée pour un marqueur donné va faire une moyenne en prenant en compte les intensités de X et Y pour tous les individus afin d'en définir des groupes.
En revanche, si les **allèles sont rares** (moins de 1%), comme pour les homozygote GG, il y aura très peu d'information à ce sujet et l'algorithme va avoir tendance à introduire ces allèles rares dans un groupe d'allèle plus fréquent (GT). C'est la masse qui distribue le génotype avec un poids important pour les génotypes les plus fréquents.
Comme le génome est redondant, on peut interroger un nombre plus restreint de variants plutôt que tous les variants du génome. Seulement, les **variants rares** ne représentent pas assez d'information, si bien qu'ils sont déclarés comme étant un **génotype inconnu**.
Finalement, un génotype est une **mesure indirecte** par hybridation avec la sonde : l'intensité permet de dire quelles sont les bases présentes.



B. Séquençage

Pour le séquençage d'un variant à une position donnée (la croix), on va d'abord séquencer plusieurs morceaux d'ADN, puis les reséquencer de nombreuses fois, et enfin l'algorithme va déterminer, selon le marquage jaune/rouge, le génotype.

Ici, la **couverture est de 7** (il y 7 fragments d'ADN qui ont été séquencés) et il y a à peu près autant de rouge (G) que de jaune (T), il s'agit donc d'un GT.

Le séquençage **ne dépend pas des autres individus** de l'étude. Les sondes sont définies à l'avance,
mais il n'y **pas d'a priori sur le nombre de variations** que l'on cherche. Il y aura autant de variants que présents dans l'échantillon.

QCM

QCM 1 :
En faveur d'une contribution génétique :

A. Variabilité entre les individus
B. Risque récurrence familiale bas
C. Risque de récurrence familiale élevé
D. Concordance MZ <concordance DZ
E. Concordance MZ >concordance DZ
F. Aucune de ces réponses n'est exacte

Réponse :
QCM 1 : A C E
QCM 2 : C D E F
QCM 3 : A B C
QCM 4 : B D

QCM 2 :
Il faut de l'ADN pour :
A-Pour calculer lambda s
B-Pour une étude de ségrégation complexe
C-Pour une liaison génétique
D-Pour une étude d'association
E-Pour faire du séquençage
F-Pour faire du génotypage

QCM 3 :
Il faut des familles pour :
A-Pour calculer lambda s
B-Pour une étude de ségrégation complexe
C-Pour une liaison génétique
D-Pour une étude d'association
E-Pour faire du séquençage
F-Pour faire du génotypage

QCM 4
Concernant une étude d'association :
A-Il faut des familles
B-Il faut de l'ADN
C-Identifie le variant causal
D-Peut être du hasard
E-Peut être confirmé par une étude de liaison

FICHE RECAPITULATIVE

I. Introduction

La génétique épidémiologique démarre à partir d'une idée formalisée sous la forme d'une **hypothèse**. Ho : la génomique de l'hôte n'influence pas le trait d'intérêt

H1 : la génomique de l'hôte influence le trait d'intérêt

Il faut trouver quels **tests** existent pour démontrer ou être en faveur de l'hypothèse qu'il y a un rôle des facteurs génétiques de l'hôte dans votre phénotype d'intérêt. Il faut donc distinguer :
-**sans ADN/avec ADN**
-sans familles/avec familles

II. Etudes sans ADN

A. Variabilité inter individuelle

Pour parler de génétique, il faut qu'il y ait de la **variabilité entre les individus**. S'il n'y a pas de variabilité, il n'y a pas d'intérêt à aller chercher un quelconque rôle du génome de l'individu car dans ce cas-là, tout le monde est égal devant ce qui se passe et il n'y a rien à expliquer. L'information s'étudie lors d'études **sans familles**, chacun est indépendant des autres.

B. Concentration familiale

On observe la concentration familiale de l'étude avec des cas si on est sur un phénotype 0/1 ou des valeurs élevées s'il s'agit d'un phénotype quantitatif. On va essayer de quantifier ces **observations**, les cas appartiennent plus ou moins à des **familles.**
risque de récurrence familiale (= (prévalence chez individus ayant frère ou sœur atteint)/ (prévalence chez individus n'ayant pas un frère ou sœur atteint)

C. Etudes de jumeaux

On va comparer **la concordance** pour un phénotype chez des **jumeaux monozygotes et dizygotes.** Etant donné qu'on travaille souvent sur des **traits plutôt rares**, on ne considère pas la concordance non atteint/non atteint car il y en aurait trop.

D. Etudes d'adoptés

Deux jumeaux monozygotes ayant un fond d'ADN identique mais un environnement différent. Si pour un trait phénotypique, les enfants ressemblent plus à leurs parents adoptifs, c'est **environnemental,** et
s'ils ressemblent plus aux parents biologiques, c'est plus en faveur d'une **origine génétique**.

E. Etudes de ségrégation

On regarde la ségrégation familiale des cas : on recueille des **familles**, on regarde qui est atteint et qui ne l'est pas. On va expliquer le phénotype d'un individu par des **variables environnementales** (ex : co-variable 1 : tabac ; co-variable 2 : exposition solaire…) et des **variables génétiques** (gène a, b…).
Ce sont des **études sans ADN** car on ne s'intéresse qu'à la **ségrégation des phénotypes au sein des familles.**

III. Etudes avec ADN

A. Analyse de liaison

1. Le principe

L'hypothèse de départ est que le **génotype de l'autre joue un rôle dans notre génotype**, ce qui implique une variabilité du phénotype que l'on va essayer de relier au génome. Il faut donc poser des hypothèses.
Les hypothèses initiales sont alors :
H0 : le génome de l'autre ne joue aucun rôle
H1 : le génome de l'autre joue un rôle
On ne fait plus vraiment d'études sans ADN parce que maintenant, on part du principe que le génome joue toujours un rôle.

2. Étude sur les familles

L'étude de liaison est une étude qui nécessite des familles : frères et sœurs, cousins ... Sachant qu'il y a déjà **2 personnes malades dans la famille**, la probabilité que ces cas soient des cas génétiques est augmentée par rapport à des cas sporadiques.
On veut dans chaque famille au **moins 2 enfants atteints** de la pathologie étudiée. C'est une **donnée imposée**. => Ayant des enfants qui ont le même phénotype, ont-ils une **région du génome où ils se ressemblent plus** que ne le voudrait le hasard ?
On a en abscisse, les chromosomes, et en ordonnée le score. **Plus il est élevé**, plus c'est en faveur du fait que **cette région du génome co-ségrège de façon non-aléatoire** ; les enfants partagent une région du génome plus que ne le voudrait le hasard.

B. Analyse d'association

Elles permettent de **réduire encore le champ**. On prend plein de cas et plein de témoins, si possible indépendants les uns des autres. On les génotype pour des variants. Dans le cas où il a été fait juste avant une étude de liaison, on prend des variants pris dans les gènes des chromosomes pour lesquels on a restreint l'étude.

1. Première méthode

On peut aussi génotyper tout le génome, comme pour la maladie d'Alzheimer, car il est plutôt compliqué d'avoir les parents, mais cela n'a pas beaucoup de spécificité génétique.
Par exemple, dans le cas du cancer du poumon, on regarde ce qui fait varier la probabilité de l'avoir : garçon/fille, alcoolique/non alcoolique, grand/petit, caucasien/non caucasien, fumeur/non-fumeur... À ça, on ajoute juste comme potentiel d'explication des variables qui sont des génotypes particuliers.

Dans le cas où **aucune étude de liaison n'ait été faite** avant, il peut être instructif de savoir si dans
la famille il y a déjà eu des cas. On peut ainsi restreindre l'analyse à ceux qui ont déjà eu d'autres cas dans la famille. Les cas sont donc plus génétiques.

2. Deuxième méthode

On a fait une étude de liaison juste avant, on sait que l'on peut **se concentrer sur certaines régions**, que l'on va **saturer en variants**. C'est une approche **génocentrique**. On va, comme pour l'étude de liaison voir s'il y a un variant différent des autres qui pourrait expliquer la pathologie.

3. Troisième méthode

On ne passe pas par l'étape de ce qui est sans ADN ni pas l'étude de liaison et on va directement prendre 10 000 témoins, génotyper 1 million de marqueurs chez tout le monde, les **tester 1 par 1** et voir s'il y a une **surreprésentation d'un des allèles**, si c'est significativement différent entre les
cas et les témoins. Cela forme un **Manhattan Plot**.

IV. Génotypage et séquençage

A. Génotypage

Dans le génotypage, les puces interrogent 1 million de variants.

Comme le génome est redondant, on peut interroger un nombre plus restreint de variants plutôt que tous les variants du génome. Seulement, les **variants rares** ne représentent pas assez d'information, si bien qu'ils sont déclarés comme étant un **génotype inconnu**.
Finalement, un génotype est une **mesure indirecte** par hybridation avec la sonde : l'intensité permet de dire quelles sont les bases présentes.

B. Séquençage

Pour le séquençage d'un variant à une position donnée (la croix), on va d'abord séquencer plusieurs morceaux d'ADN, puis les reséquencer de nombreuses fois, et enfin l'algorithme va déterminer, selon le marquage jaune/rouge, le génotype.
Le séquençage **ne dépend pas des autres individus** de l'étude. Les sondes sont définies à l'avance,
mais il n'y **pas d'a priori sur le nombre de variations** que l'on cherche. Il y aura autant de variants que présents dans l'échantillon.

Mot d'amour de la RL : BRAVO d'avoir survécu à ce cours ! (trop intéressant mais bien trop long à mon goût !!) J'espère que vous profitez bien de vos vacances de la JOIE ! :D