UE11 – Parcours – Biologie génétique Cours n°13 6/03/2018	RT : Pierre Molins Maria Mouchtaq RL : Andrea Mesrobian

A survival kit to genetic epidemiology

I. Introduction : la démarche scientifique

II. La génétique épidémiologique

III. Variabilité du génome

A. Microbiologie
B. Homogénéité
C. Variabilité des maladies
D. La génétique mendélienne
E. La génétique complexe
F. Le principe de pénétrance
G. Réunion des deux écoles
H. Homogénéité allélique et homogénéité génique

IV. Chronologie de la médecine en terme d'information génétique

A. Génétique épidémiologique
B. Etude épidémiologique en faveur de la génétique dans la pathologie
C. Analyse de ségrégation
i. ségrégation simple
ii. ségrégation complexe
iii. calcul du résidu
D. Analyse de liaison
E. Analyse d'association

V. Genome Wide Association Study

I. Introduction

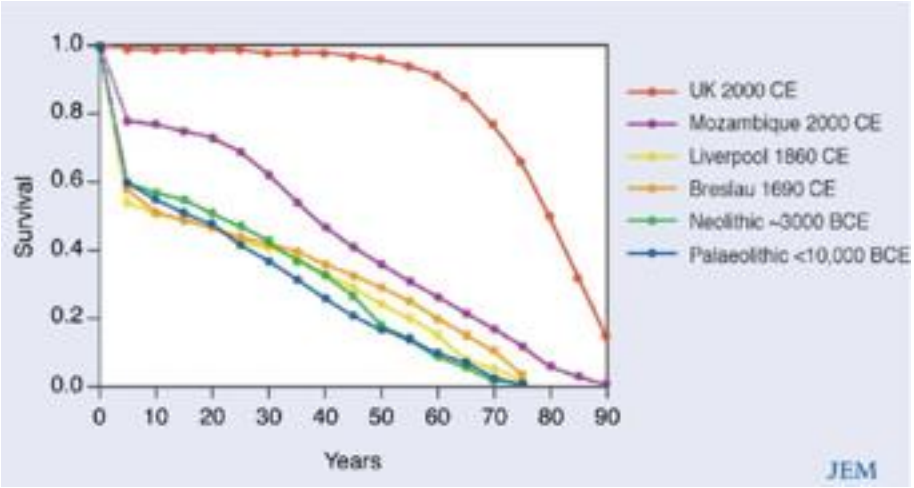
Dans l'étude des variants impliqués dans le phénotype d'une maladie, la démarche scientifique est toujours la même : on part de **génomes candidat**, de variant candidat, d'une cascade candidate. Dans le cas de la lèpre, la recherche se propose de trouver un gène candidat grâce aux données, ce qui nécessite de se questionner, en posant dans un premier temps une hypothèse nulle (ex : « le gène delta 2 ne joue aucun rôle dans l'apparition de la lèpre ».) Il faut par ailleurs maîtriser une grande quantité de données, or la gestion des big data pose un réel problème.

A la thèse s'ajoute des hypothèses H0 et H1, puis une accumulation de données afin de réfuter ou appuyer l'hypothèse. Si l'hypothèse est validée, il faut alors comprendre le mécanisme via lequel ce variant/ce gène prédispose à des pathologies.

Dans la génétique épidémiologique, on peut démontrer que dans 10 populations, un variant est associé à celui de la lèpre, mais il s'avère compliqué de démontrer le lien de causalité entre ce variant, la maladie et l'individu. Une solution est de mener des **études de cohortes**, de suivre des sujets pendant 50-60ans, et d'observer leur évolution.
L'avenir de la recherche réside donc dans le suivi de cohortes et surtout dans la mise en collaboration des banques de données. A partir de la grande quantité de données que sont les big data, il est alors possible de générer des hypothèses.

Afin de prouver le lien de causalité, il est nécessaire que la mutation précède l'apparition du phénotype. Toutefois, cela ne fait que montrer qu'il y'a une association, sans pour autant prouver quel élément est la cause de l'autre. En effet, il est possible que ce soit la mutation qui ait causé le phénotype, mais il est également envisageable que cette mutation ait été sélectionnée pendant 20 000 ans, et donc que c'est le phénotype qui a généré ce variant dans le génome.

II. La génétique épidémiologique



Ce graphique représente des courbes de survie de l'espèce en fonction de l'âge, à différentes époques de l'humanité. Il s'étend du paléolithique (10 000ans avant notre ère) jusqu'aux années 2000. Sur les 4 premières courbes (jusqu'à 1860, Liverpool), il y'a peu de variations, 50% de la population mourrait avant l'âge de 10ans : l'espèce sacrifie 50% de la population pour optimiser l'apparence du génome et assurer la survie de l'espèce.

Sur les courbes de survie au Royaume Uni en 2000, on remarque un décrochage total : jusqu'en 1860, 50% des gens mourrait avant 10ans tandis qu'en 2000 50% des gens vivent jusqu'à 80ans.

Il y a plusieurs explications possibles à ce décrochage: soit le génome humain s'est brutalement suradapté à son environnement *(peu probable),* soit on a découvert des moyens d'empêcher la **maladie** (infectieuses et métaboliques), la **famine** et la **guerre** de tuer les populations. Ces trois groupes d'événements ont modulé le génome humain.

La lutte contre la maladie est un système d'évolution dynamique qui implique un combat incessant : suite à l'adaptation de l'homme au microbe, un nouveau microbe plus résistant peut apparaître. Dans le cas des maladies métaboliques (formes génétiques de diabète, trouble de conduction hormonal…) les causes génétiques sont visualisables facilement.
Pour ce qui est de la guerre, celle-ci a laissé émerger des **phénotypes « psychiatriques »** avec des modifications génétiques plus difficiles à visualiser car les phénotypes sont plus compliqués à réaliser (avant de faire de la génétique il faut d'abord faire des phénotypes).

La découverte de la **vaccination**, l'émergence des règles **d'hygiène**, du **lavage** des mains, la découverte des antibiotiques a permis d'améliorer grandement le pronostic de l'infectio n nosocimiale. Sur le graphique, si on s'interesse à la courbe du Mozambique, on constate qu'elle se situe entre la courbe de Liverpool en 1860 et la courbe du Royaume Uni en 2000. Or la différence entre le Royaume Uni et le Mozambique est l'accès aux soins, et notamment la couverture vaccinale, ce qui explique la variabilité dans l'âge de la mort.
La question est maintenant de savoir comment ces variations se manifestent dans notre génôme.

III. La variation dans notre génome et ses conséquences

La recherche s'est d'abord orienté vers les phénotypes morbides puis vers des phénotypes plus « sains », c'est à dire en amont de l'expression génique de la pathologie.

Des chercheurs se sont intéressés au fait que certaines populations sont naturellement et massivement exposés à la tuberculose puisqu'ils respirent un air dans lequel tout le monde tousse, et pourtant ne sont pas positifs/ne s'infectent pas. L'infection est ici en amont des phénotypes causaux.

Des entreprises privées sur internet proposent de trouver le sport le plus adapté à chacun à partir de son génome en analysant un échantillon de salive, et de déterminer nos performances en termes de vélocité, dynamisme, explosivité, résistances aux fractures. En réalité, c'est une escroquerie.
En effet, si on prend l'exemple du foot, les footballers de haut niveau expriment le gène A à 30% tandis qu'il est exprimé à 10% pour le reste de la population et 5% chez les sédentaires. Il n'y a pas de lien direct. La probabilité de réussir au football est légèrement augmentée avec ce variant, cependant, de nombreuses personnes possèdent ce variant sans pour autant être douées au foot.

Les maladies infectieuses étaient par définition les maladies environnementales, ce qui impliquait que des maladies telles que la phénylcétonurie était environnementales, or celle-ci implique une enzyme. En réalité, rien n'est totalement environnemental et rien n'est totalement génétique. Dans les maladies infectieuses, le fait qu'un **microbe puisse être à l'origine d'une maladie génétique** constitue une des découvertes les plus importantes de l'histoire de la médecine.

Une question se pose alors : « Pourquoi, chez des personnes exposés à un même microbe, certaines vont mourir et d'autres ne vont pas s'infecter ? »

Cela peut s'expliquer par le fait que certaines personnes sont naturellement résistantes et d'autres qui sont naturellement susceptibles (d'où l'importance de la vaccination). Dans le cas de la tuberculose, on continue à faire du BCG (qui protège des tuberculoses extrapulmonaires de l'enfant et sauve ainsi de nombreuses vies) alors qu'il n'a aucun impact sur la tuberculose pulmonaire, qui est la forme contagieuse.
La question est donc de savoir si on doit continuer à effectuer le BCG, ce qui pose soit un problème technique soit un problème de concept. Deux mondes s'affrontent : un monde composé de 99% des personnes qui pensent à un problème technique, et un monde plus restreint pour qui c'est un problème de concept, puisque le BCG ne protège pas de la forme contagieuse qu'est la tuberculose pulmonaire.

A. Microbiologie

Fondée par Pasteur, la microbiologie est un champ clé *(centaine de milliers de chercheurs en microbiologie contre centaine de chercheurs en génétique des maladies infectieuses).* Pasteur a étudié la variabilité, les maladies infectieuses, le fait que des causes génétiques pouvaient être dues à des microbes/maladies infectieuses. A la fin du 19ème siècle, on était encore dans les courbes de 50% de mortalité avant 10ans.

La question qui se pose alors est : comment expliquer cette variabilité ?

B. Nécessité d'homogénéité

Tout d'abord, il est important de rappeler la nécessité d'**homogénéité**. Peu importe la taille de l'échantillon final: avoir 100 cas de tuberculose absolument homogènes *(touche les hommes, même forme génique, développée au même âge…)* est mieux que 5 000 cas de tuberculose extrêmement hétérogène au niveau du phénotype. Il faut donc passer du temps sur le phénotype, et l'objectif n'est pas la taille mais l'obtention d'un groupe d'individus le plus homogène possible.

L'homogénéité doit concerner **l'âge**, le **sexe**, la **forme génique** mais également le **stade** de la maladie. Ainsi, dans le cas de la tuberculose, la maladie est l'aboutissement de plusieurs étapes : exposition à la tuberculosis, infection, tuberculose latente, tuberculose pulmonaire, et pour construire un bon échantillon, il faut que les sujets soient au même stade.

Dans le domaine de la psychiatrie, la recherche est plus performante en terme de découverte de **causes partielles** (explique une partie mais pas tout), grâce à un **affinement de la définition phénotypique**. Des groupes sont devenus de plus en plus homogènes permettant la formation de sous-groupes de taille convenable.

Dans certaines pathologies rares, seuls 2 individus sont atteints dans le monde : cela constitue un groupe extrêmement homogène et la défaillance génétique est en générale majeure.

C. Variabilité des maladies

Comment expliquer la variabilité des maladies infectieuses ?
Dans les maladies infectieuses, plusieurs choses pourraient expliquer la variabilité :

- l'**exposition au microbe** (plus on est exposé plus on risque d'être infecté, avec certaines souches plus virulentes que d'autres (sucres simples, sucres composés…)
- les **facteurs de l'hôte** (facteurs non spécifiquement génétiques: l'immunodéficience augmente le risque d'infection), et facteurs génétiques qui spécifiques d'une pathologie ou d'un microbe donné).

Il est intéressant de parvenir à identifier ces facteurs génétiques puisqu'ils nous donnent une indication sur les cascades génétiques qui font que certains individus sont résistants et d'autres ultra-susceptibles).

La question qui se pose maintenant est : comment parvenir à identifier ces facteurs génétiques selon la pathologie que l'on étudie ?

D. Phénotype rare et génétique moléculaire (mendélienne)

Lorsque le **phénotype d'une maladie est très rare et très sévère,** cela nous donne une indication sur la sévérité du déterminant génétique mis en cause. Ces individus là seront peu nombreux mais riches en informations, et afin de les explorer il est judicieux de faire une étude **patient-based**. Elle consiste à explorer un patient ou une famille à la recherche d'altérations importantes des fonctions: c'est la **génétique moléculaire** ou **génétique mendélienne**. Le but est de parvenir à trouver des **mutations extrêmement rares**. Leur découverte suit un principe de tout ou rien : si le patient présente la mutation, c'est grave car elle peut avoir une conséquence clinique considérable, alors que la gravité est moindre s'il ne l'a pas.

E. Phénotype commun et génétique complexe (galtonienne)

A la génétique moléculaire mendelienne s'oppose la **génétique galtonienne** (même époque que Mendel). Ici on s'intéresse à un **phénotype très commun**, donc le but est d'identifier un **variant dont la récurrence est relativement élevée**. Si sa fréquence est élevée, c'est qu'il n'a pas été soumis à la pression de sélection naturelle, donc son **effet doit être faible.**

En travaillant sur un génotype commun, il est préférable d'agréger de nombreux individus, en effet, on cherche une information très fine, donc amalgamer de nombreux cas permet de mieux identifier cette information. *(ex: agréger 200 individus atteints de tuberculoses disséminées à 15ans; le fait d'avoir de nombreux individus permet d'obtenir assez de matériel génétique pour identifier les facteurs de survenue de cette tuberculose).*

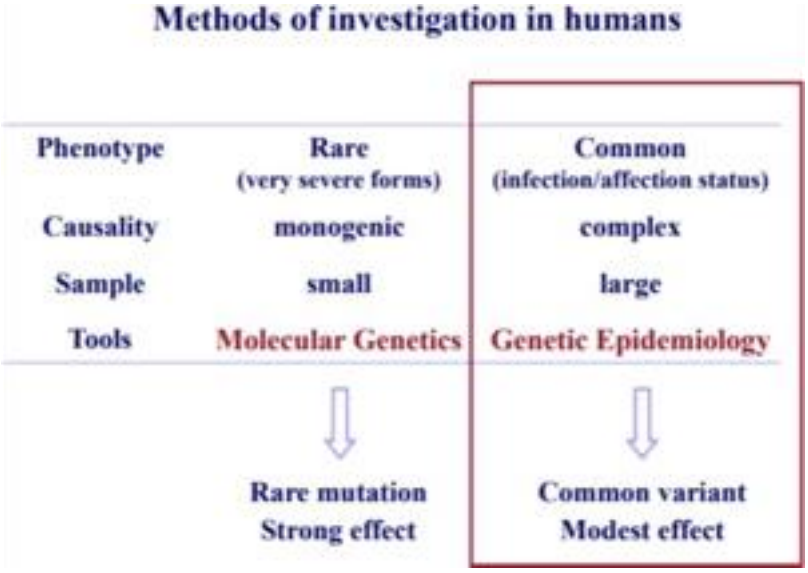
Le but est de trouver un **variant de fréquence relativement élevée** mais avec un **effet modeste** (ce qui s'oppose à la génétique mendélienne où si on est porteur alors on est malade). De plus, plusieurs gènes sont en cause et chacun a un effet, d'où la nécessité d'un échantillon relativement important.

F. Génétique mendélienne et pénétrance

La génétique mendélienne soulève la question de la **pénétrance**, qui peut être **incomplète**. La pénétrance incomplète peut être vue comme un **risque relatif**: c'est la mesure qu'on fait dans les maladies dites **communes à effet complexe**. D'une part on peut interpréter cet effet complexe comme étant dû à une multitude de gènes intervenant chacun avec un effet modeste ; d'autre part on peut le voir comme le résultat d'une pénétrance incomplète.

En réalité les deux vont de paire. Dans le problème de la mucoviscidose, la mutation est relativement fréquente et malgré tout sur ce fond génétique totalement homogène on va avoir des gens qui vont mourir à 6 ans et d'autres qui vont mourir à 40ans.

Dans cette variabilité, il est intéressant de rechercher des variants communs qui vont être des modulateurs.
Dans les maladies dites **communes à effet complexe,** sur 100 tuberculoses on ne peut pas affirmer que 100 tuberculoses seront homogènes. Ainsi, on a trouvé que, dans la tuberculose de l'enfant, 30-35% étaient dus à une mutation « perte de fonction » dans un gène.



Dans le cas des maladies communes, il ne faut donc pas d'abord chercher des éléments communs mais commencer par éliminer tout ce qui est expliqué par un variant rare dans un gène.
Parler de « commun variant », « commun disease » est faux, parler de « rare variant », « rare disease » est faux aussi, parce que ce qui est rare pourrait être dû à un variant très fréquent. **Les maladies fréquentes sont une somme de variants rares.**

G. Réunion des écoles mendélienne et galtonienne

Ces deux écoles sont restées éloignées pendant plus de 100ans, mais depuis une quinzaine d'années ces deux façons de penser commencent à converger.
Un laboratoire a créé en 2000 un groupe qui travaillait sur la génétique mendélienne et un autre sur la génétique complexe. L'idée était de voir si c'était une **dichotomie** absolue ou si au contraire il y avait une **continuité**. On a accumulé les données sur les tuberculoses pulmonaires dues à une mutation dans un gène et des données sur des formes sévères chez l'enfant inexpliquées par un seul gène. Ils ont donc travaillé à homogénéiser tout cela et la vraie révolution arrive juste : avant on faisait de la génétique sur de la génétique mendélienne et on partait des chromosomes (on en enlevait un, on mettait le microbe, ça produisait quelque chose, alors le chromosome n'était pas en cause et ainsi de suite). En génétique complexe on faisait des statistiques, on mettait du poids aux données.
Maintenant ce qui va obliger les gens à se rapprocher ou à se comprendre les uns les autres, c'est que quand on veut faire la génétique de quelque chose alors on prend 1000 personnes et on va faire un séquençage complet du génome.
Si on fait du mendélien alors on va chercher certaines choses et on va interroger les données d'une certaine façon, alors que si on est plutôt complexe alors on va interroger les données d'une autre façon. Si ça ne marche pas on peut changer notre façon de faire parce que le matériau de base de la génétique humaine est maintenant commun à tout le monde, c'est la séquence complète du génome humain.

Cela oblige donc les gens à parler d'un même langage. Et à partir de ce moment-là on peut travailler mais toujours à partir d'une hypothèse (qu'est-ce que l'on teste et quels sont les suppositions que l'on fait...). Dans la tuberculose pulmonaire d'un adulte on va donc regarder et chercher des variants perte de fonction dans les exons. Cela suppose qu'on a déjà éliminé tous le variants communs.

H. Homogénéité allélique/homogénéité génique vs hétérogénéité

Ce qui échappe beaucoup au monde de la génétique humaine en générale est que la contrainte porte sur le niveau d'homogénéité génétique. Si on travaille sur une famille où il y a deux cas d'une pathologie unique au monde, là on peut faire l'hypothèse d'homogénéité allélique (on cherche le même allèle chez les patients). En faisait cette hypothèse on peut faire des analyses.
Après parfois on peut avoir une maladie qui est peu fréquente mais qui est quand même un peu plus que très rare et à ce moment-là on va avoir une 100ène de familles et c'est difficile de faire une hypothèse d'homogénéité allélique (ça peut être des allèles différentes qui font une perte de fonction) on va alors faire une hypothèse d'homogénéité génique (le même gène est en cause dans toutes les familles).

Donc pour trouver ça on va faire une analyse de liaison, on va essayer d'identifier la région dans laquelle se trouve ce gène. Et après on va séquencer tous les patients et on va voir s'il y a ou non une région. On peut aller plus loin mais là où on arrive dans le paradoxe total est que quand on fait des études d'association pan génomiques (sur 1000 ou 10 000 cas et 10 000 témoins, on test 10 millions de variants communs et on regarde la fréquence est significativement différente chez les cas et les témoins), l'hypothèse qui est faite à partir de cas et contrôles tous indépendants est qu'on fait une hypothèse d'homogénéité allélique et on va tester un snip à un snip (ce qui est délirant).

Ceci n'est pas possible, donc on trouve, dans les séquences génétiques, des effets qui sont extrêmement faible (ça augmente le risque de 1,02). Ceci est miraculeux mais aussi aspécifique (on a trouvé le plus petit dénominateur commun génétique à plein de cas phénotypiquement hétérogènes mais en plus qui sont génétiquement hétérogènes dans la causalité). Il y a donc un cercle vicieux qui s'est mis en place : on ne trouve dans les études d'association pan génomiques que des effets de l'ordre de 1,1-1,2 donc très faibles. Donc quand on planifie une étude pan génomique ce qui se passe c'est que il faut au moins pour avoir de la puissance 20 000 cas. Si on met 20 000 cas on a un problème d'hétérogénéité.

Tout d'un coup les gens sont contents et surpris parce qu'ils voient les liens entre les pathologies (lien entre la tuberculose et maladie de Crohn). Ce qui est logique parce que ce que l'on a trouvé n'est absolument pas spécifique puisque tout est hétérogène. On a trouvé un mécanisme qui joue un rôle mais qui n'est pas forcément causal. Si on partait sur l'hypothèse du lien alors il faudrait aller au bout du concept et traiter la tuberculose avec les traitements de la maladie de Crohn.

Dans la génétique humaine il faut poser des hypothèses de ce que l'on veut tester (le génome, le gène, le variant…) et surtout des hypothèses en terme d'homogénéité génique. Et quand on écrit cela noir sur blanc on arrive à 100 000 cas de Crohn et 100 000 cas de pas Crohn, on fait une étude d'association pan génomique (on test 5 millions de variants). L'hypothèse de trouver quelque chose est que le même variant soit sur-partagé par tous les Crohn indépendants. C'est une hypothèse très violente et c'est pour cela que les GWAS ont disparu.

Les GWAS ont eu 8ans de durée de vie. Les gens n'ont pas fait l'effort de penser qu'un seul snip pouvait être indépendant du reste. Et l'hypothèse qui est faite quand on travaille comme ça elle est absurde, parce que c'est une hypothèse qu'on n'oserait pas faire dans une famille avec des cas extrêmement rares du fait de l'homogénéité nécessaire.

Mais là on a 40 familles dont certaines ont plusieurs allèles dans le même gène ou dans la même cascade. Maintenant les gens commencent à converger et à travailler ensemble mais ce n'est pas si fréquent. Si l'on met 20000 cas, on se retrouve dans un cas d'hétérogénéité. Ainsi on observe des liens entre les pathologies (par exemple : la tuberculose va avec la maladie de Crohn). En effet, tout étant hétérogène, il n'existe pas de spécificité. Auquel cas, cela reviendrais à traiter la maladie de Crohn avec des antituberculeux. Dans la génétique humaine, il faut poser les hypothèses sur le génome, gène ou variant mais surtout sur l'homogénéité génétique (+++). Ainsi, lorsque on fait une étude pan génomique, il faut faire l'hypothèse que le même variant soit partagé par tous les sujets. Cette hypothèse est absurde si l'on considère un grand nombre de sujets.

VI. Chronologie de la médecine en termes d'information génétique :

A. Génétique épidémiologique :

En 1995, on pensait pouvoir avoir de l'information génétique sur un million de variants ce qui n'était pas possible en terme techniques et financiers. On travaille donc sur le phénotype. De nos jours, l'information génétique est beaucoup plus accessible ce qui permet de faire de la génétique épidémiologique. Les programmes d'aujourd'hui doivent être utilisés à bon escient en posant les bonnes hypothèses.

Au début de la génétique épidémiologique, on a commencé à aborder les problèmes sous l'angle de la génétique. Dans un premier temps, il faut déterminer si la génétique joue un rôle dans la maladie grâce à des observations épidémiologiques nettement sur la transmission (récessif/dominant). Ensuite, on estime en recherchant un marqueur commun sur un groupe de malades atteint du même phénotype la localisation chromosomique. S'il y en a un, on sait que ce marqueur représente une zone du génome intéressante. Une fois que la région est repérée, on va tester les variants à travers des études d'association (voir s'il est présent chez des personnes malades/saines). Enfin, on va rechercher sa fonction avec de la génétique moléculaire. Ainsi, la génétique épidémiologique est la liaison entre l'épidémiologie et la génétique moléculaire.

B. Etudes épdémiologiques en faveur de la génétique

- la variabilité inter-individuelle : s'il n'y a pas de variabilité génotypique ça ne sert à rien de faire de la génétique. Exemple de variabilité lors du désastre de Lübeck où au lieu de vacciner avec le BCG ils ont vacciné avec le tuberculosis. Il y a donc eu des millions de morts mais il y a quand même eu des individus qui n'ont pas été infectés donc on peut penser que ces derniers étaient résistants du fait de leur génétique.
- le clustering : le fait d'avoir des cas rassemblés dans une même famille (rien ne prouve quoi que ce soit, ça va juste en faveur de...). Quand on vient de la même famille et qu'on habite près d'une zone de contamination on est plus à risque qu'une personne habitant plus loin. C'est intéressant de travailler dans des familles avec deux cas parce que dans ce cas on est plus certain d'une contribution génétique. Alors que si on prend des cas contrôles, ils sont totalement indépendants
- le risque de récurrence familiale : si on a un frère ou une sœur qui est atteint par la pathologie, on recherche le sur-risque de développer cette pathologie pour cet individu (= risque de récurrence) (fente palatine risque est de 50, sclérose en plaque de 14-15). Or on n'arrive pas à identifier les variants génétiques. La récurrence ne veut pas dire que c'est génétique (ça peut être dû au mode de vie comme à l'alimentation...) Pour calculer cette récurrence on prend 100 cas et 100 contrôles. On regarde tous les frères et sœurs de chaque cas, qui sont atteints ou pas. Il va y avoir une prévalence de la pathologie chez les frères et sœurs des cas. On fait la même chose chez les contrôles et on va voir la prévalence de la pathologie chez les frères et sœurs des contrôles. On fait donc le rapport et on trouve le risque de récurrence. En règle générale personne ne veut travailler avec des gens non malades donc on prend la prévalence de la maladie de la population au lieu des cas contrôles. On utilise généralement cette formule :

$$\lambda_S = \frac{\text{prevalence of the disease in siblings of cases}}{\text{prevalence of the disease in siblings of controls}}$$

- l'étude génétique de jumeaux : les monozygotes (partagent 100% du patrimoine génétique en commun à l'exception des mutations de novo toutes les 10^7 paires de bases) et les dizygotes (partagent 50% du patrimoine génétique, il n'y a pas d'effets cohortes). On regarde donc la concordance chez les monozygotes et les dizygotes. S'il y a une différence entre les deux frères jumeaux chez les monozygotes cela doit être dû aux facteurs environnementaux. Chez les jumeaux dizygotes cela peut être du soit à la génétique soit aux facteurs environnementaux.

On aura donc une concordance plus grande chez les jumeaux monozygotes que chez les dizygotes. Dans la lèpre 60% des monozygotes sont concordants et 20% seulement des dizygotes sont concordants (dans la tuberculose c'est de l'ordre de 30% et 10%). Or la lèpre est quelque chose qui n'est pas très fréquent.
On a donc un tableau avec le jumeau 1 et le jumeau 2 qui peuvent être malade ou non malade.
Pour ne pas se noyer dans les informations il faut donc faire une correction et donc éliminer tout ce qui est concordant non malade (case barrée dans le tableau). Trop de gens sont non malade donc cette correspondance ne nous intéresse pas. On regarde donc juste la concordance entre les malades versus pas malade.
Ici il y a les caractéristiques mais on est sans ADN, on a juste besoin de familles avec leur structure familiale (avec des jumeaux) ou de gens indépendants.

C. Analyse de ségrégation

i. Ségrégation simple

Pour caractériser ces facteurs génétiques on va regarder dans des arbres généalogiques sans ADN.
Si le phénotype s'agrège d'une façon qui rappelle un modèle connu (récessif, dominant, lié à l'X...) alors on pourra trouver le modèle génétique. On a donc besoin de familles pour suivre la ségrégation génétique des maladies.

ii. Ségrégation complexe

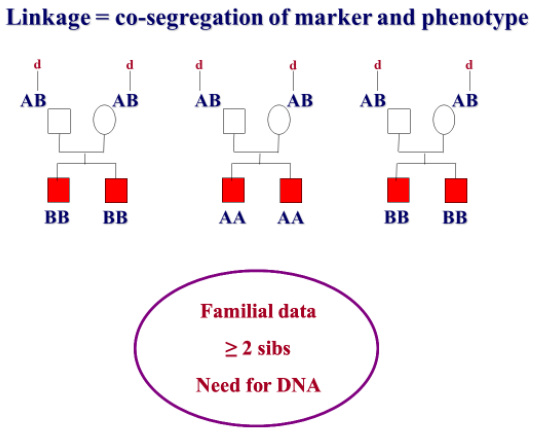
Le problème est que quand il y a un peu de complexité dans le domaine génétique (2-3 gènes, une interaction gène-environnement, une interaction inter-gène, une pénétrance incomplète, des effets variables...) c'est plus difficile de trouver le bon modèle.
Comment faire pour trouver le bon modèle :
Si on teste l'étude du tabac sur le phénotype : on fait un tableau avec atteint et non atteint. Si on a 100% des atteints qui fument et 0% des non atteints qui fument alors il n'y a pas besoin d'aller plus loin (le modèle est validé). Or ceci est très rarement le cas donc on va rajouter un gène puis plusieurs dans le tableau et on va voir si tabac + gène donne un pourcentage plus haut que tabac seul de tomber malade. Le problème c'est que l'on ne peut pas générer tous les modèles possibles, or on doit commencer par générer un modèle. On ne peut donc jamais explorer l'ensemble des modèles, et il y a plein de modèles qui expliquent aussi bien les données.
On a alors une mesure de combien cela explique les données. Il n'y a pas d'ADN du tout là-dedans.
S'il s'agit d'un gène de référence à pénétrance complète alors tous les non atteints ont les allèles n'amenant pas à la maladie. On peut donc donner des génotypes à tout le monde.
Il faut donc beaucoup de familles (d'autant plus qu'il y aura plein de variables à tester). On a donc des familles ou on a mesuré l'étendue de la tuberculine et on a ajusté les covariables (on doit tenir compte des allergies) et on doit tenir compte des résidus.

iii. Calcul du résidu

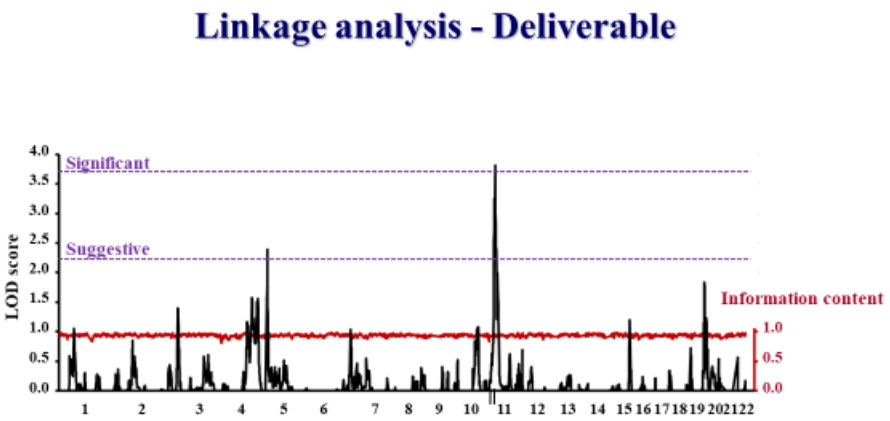
Si on a montré que le modèle qui explique le phénotype est que si on est fumeur alors cela augmente le risque de 4,4 et l'alcool augmentait de 8,2 (ce sont des valeurs qu'on a observé qui seront donc la prédiction du modèle). Si on mesure chez une personne un risque de 20 (si la donnée est mesurable par des tests fonctionnels quantitatifs) alors on peut confronter cette mesure avec la prédiction du modèle. Le résidu est la valeur observée moins celle qui est prédit. (Ici 20 – (8,2+4,4) = 7,4 = résidu)
Quand on mesure via le résidu on peut donc avoir des valeurs négatives (si la personne est moins atteinte que ce que le modèle le prédisait).
On peut faire un parallèle au test du ki2.
On peut donc quantifier la différence pour obtenir quelque chose pour travailler qui est plus simple.
Si la maladie est une donnée qualitative alors le malade sera 1 et le non malade 0. On calculera alors la probabilité d'être malade en fonction des différentes données du modèle (soit un chiffre entre 0 et 1). On peut de nouveau calculer le résidu.
Cette méthode du résidu est intéressante parce qu'elle permet d'éliminer toutes les sources de variations qui ne sont pas génétiques. C'est pour cela qu'avant de commencer une étude il faut passer un temps considérable à lire une bibliographie pour faire une liste la plus exhaustive possible de tous les facteurs de variations (pour que la variabilité qui reste soit vraiment inexpliquée).
Ceci est vraiment important parce que si on ne prend pas en compte par exemple la cigarette dans l'étude du cancer du poumon alors nous aurons une variabilité considérable qui est expliquée par une variable non prise en compte.
Lors de l'analyse de ségrégation on regarde sur un graphique la disposition des personnes en fonction de s'ils sont malades ou non. Parfois on voit clairement une différence entre deux groupes mais parfois seul 2% des personnes sont à un endroit complètement différent du graphique. Cette méthode d'analyse ne se fait plus mais peut encore la retrouver dans les bibliographies.
Cette méthode permettait de dire ce qui était dominant et pouvait alors aider pour l'analyse de liaison.
Pour l'analyse de ségrégation on a donc besoin de familles mais pas d'ADN.

D. Analyse de liaison :

Pour l'analyse de liaison on a maintenant besoin d'ADN (plus de phénotype). Elle revient à analyser une ou des régions du génome afin de trouver les zones intéressantes pour expliquer le phénotype. On regarde donc la ségrégation du phénotype mais on va en même temps regarder si il y a un marqueur qui vient toujours avec le phénotype atteint. La région devient donc un candidat « positionnel ».
Il faut donc des familles pour suivre les phénotypes et de l'ADN pour suivre les marqueurs (leur fonction n'a pas d'intérêt, il faut juste connaitre leur localisation). On fait l'analyse de liaison sur 3000/4000 ou 5000 marqueurs pour déterminer les localisations intéressantes lorsqu'il y a ségrégation avec un des phénotypes des familles.



On prend donc un cas familial très binaire avec au moins un deux atteints par famille. Ici, dans la première famille, les deux atteints ont reçu l'allèle B de leur père et de leur mère qu'on identifie avec d qui est un tag de la région donnée. Dans la deuxième famille, d vient avec l'allèle A et non B (d ne fait que taguer la localisation « malade »). Ainsi le marqueur n'est qu'un tag et le seul intérêt est qu'au sein des différentes familles les malades partagent plus que ne le voudrait le hasard les allèles d'un marqueur. Ce marqueur n'a aucune importance en soit. On répète l'opération afin de cribler tout le génome et on se retrouve avec un résultat de ce type :



Chaque marqueur tag une partie du génome et on mesure si ça s'agrège avec le phénotype. Si oui, on a une valeur importante. C'est la méthode de Lod score : Il faut des familles avec au moins deux atteints et on va regarder si les allèles sont distribués de façon aléatoire ou si dans une famille donnée un marqueur représente toujours 100% des allèles. Dans ce cas-là, cela veut dire que cette région que tague le marqueur a l'air de venir toujours avec le phénotype.

Pour l'analyse de liaison on a donc besoin de familles avec au minimum 2 individus atteints, ainsi que de l'ADN.

E. Analyse d'association :

Une fois que l'on a trouvé les régions à l'issu des analyses précédentes, on peut maintenant se focaliser sur une beaucoup plus petite région (3 milliards à 5 millions de pdb) (dans l'exemple précédent, la région portait sur le chromosome 11). Cette phase d'analyse consiste à comparer la distribution des génotypes pour un variant donné entre des cas indépendants et des cas contrôle. On analyse ainsi la fréquence de récurrence des différents variants entre les atteints et les contrôles.
Pour trouver le variant causal, il faut éliminer les variants redondants notamment en expérimentant sur des populations qui ont moins de SNPs comme en Afrique.
Pour une étude d'association il ne faut donc pas de familles mais il faut de l'ADN.
On observe dans un premier temps un résultat comme ça :

Hypothesis testing – Deliverable

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

Interpretations

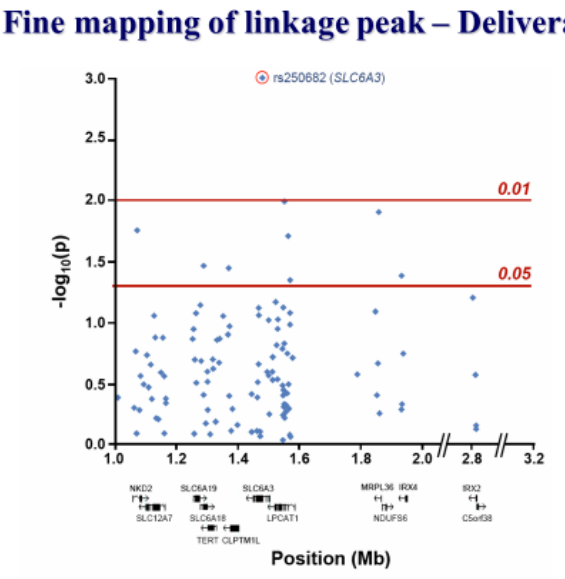
Type I error (false positive)

Allele B ⇒ phenotype = B is the causal allele

Allele B is in *linkage disequilibrium* with the causal allele

Population stratification ⇒ familial designs (TDT statistics)

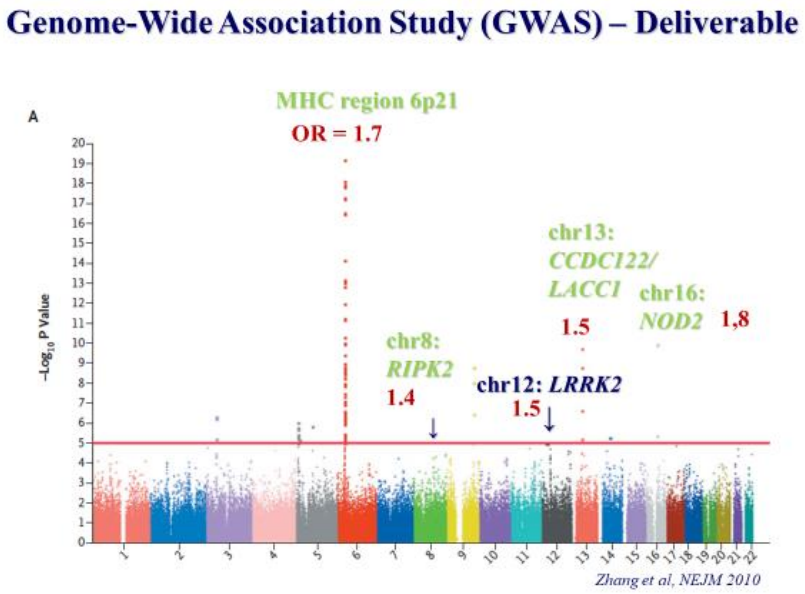
Si la récurrence entre le variant et les cas est significative, on peut alors dire qu'ils sont associés.
Dans un second temps, on regarde de manière plus précise sur la zone étudiée : plus on va dans le « petit », plus c'est significatif. On fait du fine mapping (c'est-à-dire faire de l'association uniquement sur 30Mb) et on regarde si ces variants sont sur ou sous représentés chez les atteints ou les malades et on regarde les différences.



Ici on peut identifier un variant en particulier (récepteur à la dopamine) qui joue donc un rôle de manière significative dans la tuberculose.

VII. Genome Wild Association Study

Dans ce type d'étude, on prend tout le génome et on teste toutes les variations du génome. On a donc un résultat avec des 100aines de milliers de points qui correspondent chacun à un variant classé en fonction de sa localisation dans le génome (les chromosomes sont en abscisse) et de son degré d'association avec la maladie (RR en ordonnée). On appelle ce genre de schéma un Manhattan bolt (en références au buildings).



Les GWAS sont bien car ils donnent des idées pour ensuite faire un travail de test sur ces gènes. Notons tout de même que le seuil de 5% est purement bancal et constitue un vice de pensée car plus on augment le nombre de cas, plus on trouve des hétérogénéités.

Conclusion

Avant on avait un phénotype, et la première question était « est ce que vous avez une idée du gène en cause ? », si oui alors il fallait tester l'hypothèse (l'idée pouvait venir des plantes, des animaux...).
Quand on n'avait pas d'idée ou que nos idées n'avaient pas marché alors on repartait pour générer des idées (la liaison était alors une bonne façon de trouver certains gènes auxquels nous n'aurions pas pensé (par exemple dans la lèpre on a trouvé un gène qui était connu pour être causal des
Parkinson juvéniles). Puis il y avait les études d'expressions (linkage d'expression) qui n'ont pas marché du tout (l'idée était de prendre des cas et des contrôles et de regarder le profil de transcription et de voir si quelque chose était distinguable (tous les gènes devenaient candidat pour jouer un rôle)). Le problème est qu'à la différence de l'ADN (qui est pareil partout), l'ARN est très variable.
Si on travaillait sur la tuberculose et qu'on faisait des profils d'expression, alors il fallait trouver l'endroit où l'on allait chercher l'information pour faire des profils d'expressions (du sang total, des ovaires...). De plus on avait envie de voir ce qui se passe quand on introduisait le microbe (pareil pour le diabète ou l'action du glucose ou de l'insuline in vivo sont intéressantes). On ne savait donc pas où regarder et on ne savait pas où stimuler donc cette méthode de recherche ne peut pas marcher.
Il y a peu de temps, on faisait un GWAS de manière systématique ce qui coutait très cher et qui ne permettait que de faire peu de choses. Aujourd'hui, on garde le GWAS pour les phénotypes communs, les phénotypes rares étant étudiés avec un séquençage entier de l'exome (le prix est de plus similaire). La stratégie du futur est donc de faire un séquençage d'exome complet pour tous les phénotypes. Il faut donc imaginer un test qui soit complet et qui fonctionne avec le mendélien et la génétique complexe.

Réflexion du prof sur l'évolution des méthodes :

Au départ on avait une idée, on la testait, elle ne marchait pas, on prenait une deuxième idée... La technologie était au service de pouvoir tester plusieurs idées plus vite. La technologie répondait donc à un besoin (= hypotesis testing). Puis on est arrivé à des hypotesis generating et des GWAS qui font que la technologie permette de générer des idées (ce qui est aussi bien).
Par contre maintenant on est à un stade de recherche qui devient massivement technologique.
C'est la technologie qui va décider de l'hypothèse que l'on va tester (par exemple : puces de méthylation et rôle ou non dans la lèpre ; bientôt on sortira la puce de glycosylation donc on testera avec cette puce pour voir ce que cela donne). Cela est très mauvais parce que cela coûte très cher et les gens ne pensent plus (plus de pensée brute).
Le problème est que l'objectif n'est pas de décrire indéfiniment les choses, mais il faut se lancer, essayer d'expliquer (même si c'est complètement faux).

On est maintenant inondé de données et on s'y perd. On va croire plus facilement à une variation dans un exon parce qu'historiquement c'est comme ça que c'était, et c'est la technologie qui nous amène à faire cela.

Pour en savoir plus :

Sham P, editor. Statistics in Human genetics. 1st ed. London: Arnold; 1998.
Rao DC, Gu GC, editors. Genetic Dissection of Complex Traits. 2nd ed. Academic Press; 2008. Ziegler A, Konig I., Pahlke F. A. Statistical Approach to Genetic Epidemiology. Wiley. 2010.
Thomas D. Statistical methods in genetic epidemiology. Oxford University Press. 2004.
Palmer L, Elston RC, Fallin D. Genetic epidemiology: fundamental concepts. Wiley 2007.

FICHE RECAPITULATIVE

I) La génétique épidémiologique

Jusqu'à la moitié du 19ème siècle, 50% de la population mourrait avant 10 ans. En effet, on a trouvé des moyens de lutter contre ce qui tuait les gens (les maladies infectieuses, les problèmes métaboliques et les guerres). La découverte de la vaccination, de l'hygiène, des antibiotiques etc nous ont aidés à ça.

II) La variation dans notre génome et ses conséquences

A. Microbiologie

Fondée par Pasteur qui étudiait la variabilité notamment dans les maladies infectieuses (qui pouvaient induire des causes génétiques).

B. Nécessité d'homogénéité

La taille de l'échantillon n'importe pas, c'est l'homogénéité qui compte (d'où l'intérêt des sous-groupes de population). On ne peut par exemple pas mettre des gens atteints de tuberculose dans un groupe sans faire de sous-groupes, car sinon on trouverait des hommes, des femmes, des malades de tout âge, ce qui rendrait ce groupe très hétérogène.

C. Explication de la variabilité des maladies

La variabilité dans les maladies infectieuses peut être due :
- à tout ce qui touche au microbe
- aux facteurs de l'hôte (génétiques ou non)

D. La génétique mendélienne

Exploration d'un patient avec un phénotype rare et sévère (étude patient based) pour trouver les mutations rares, c'est de la génétique moléculaire (= génétique mendélienne).

E. La génétique complexe

C'est la génétique galtonienne, on a un phénotype très commun donc on cherche un variant avec une fréquence élevée et un effet faible. Dans le cas de la tuberculose, on va chercher ce qui influence sa survenue.

F. La pénétrance

Elle peut être incomplète et pose problème en génétique mendélienne. Dans la variabilité des pathologies, il est intéressant de rechercher des variants communs modulateurs (ex : tuberculose, mucoviscidose).

G. Réunion des deux écoles

Aujourd'hui on fait un séquençage complet du génome d'un groupe d'individus, donc le matériau de base de la génétique est commun à tous ce qui facilite la recherche. Si on fait du mendélien on va analyser les données d'une certaine façon, et d'une autre façon en génétique complexe.

H. Homogénéité allélique/homogénéité génique vs hétérogénéité

L'homogénéité allélique/génique de l'échantillon est nécessaire à toute analyse. On va identifier la région avec le gène en cause pour ensuite séquencer les patients. Dans le cas d'un échantillon hétérogène, les résultats ne sont pas spécifiques donc on ne trouve pas forcément de lien de causalité entre 2 évènements.

III) Chronologie de la médecine en termes d'information génétique

Au 20ème siècle on parlait d'épidémiologie génique (travail en génétique sans ADN, sur le phénotype). Puis l'information génétique est devenue de plus en plus accessible donc on en a été inondé et les phénotypes se sont sophistiqués. L'accès est aujourd'hui très facile, il faut ainsi faire attention à l'interprétation de toutes ces données. On parle désormais de génétique épidémiologique.

A. Génétique épidémiologique

Frontière entre l'épidémiologie et la génétique moléculaire, d'où son ambiguïté car elle n'appartient vraiment à aucune de ces deux disciplines. Elle servirait à lire et interpréter les résultats d'un séquençage ciblé par exemple, grâce à des personnes qui connaissent toutes les méthodes, la pénétrance ...

B. Etude épidémiologique en faveur de la génétique dans la pathologie

- la variabilité inter-individuelle
- le clustering
- le risque de récurrence familiale : recherche du sur-risque chez frères/sœurs du malade, la récurrence ne veut pas dire génétique

Calcul de la récurrence :

$$\lambda_S = \frac{\text{prevalence of the disease in siblings of cases}}{\text{prevalence of the disease in siblings of controls}}$$

- l'étude génétique de jumeaux : étude de la concordance chez les monozygotes (100% patrimoine génétique en commun) et dizygotes (50% patrimoine génétique en commun).

Ainsi une différence entre jumeaux monozygotes est due aux facteurs environnementaux.

C. Analyse de ségrégation

- Ségrégation simple : Grâce à l'arbre généalogique, on repère un modèle connu (récessif, dominant, lié à l'X ...) ce qui permet de trouver le modèle génétique.
- Ségrégation complexe : Domaine génétique complexe (pénétrance incomplète, effets variables, interaction gène-environnement ou inter-gène, plusieurs gènes ...) rend le travail plus difficile. On ne peut pas tester tous les modèles, donc on mesure de combien ça explique les données.
- Calcul du résidu : Résidu = valeur observée moins celle qui est prédit

Avec le test du ki2, on calcule la probabilité d'être malade en fonction des données du modèle. Cette méthode permet d'éliminer les sources de variation non génétiques.

L'analyse de ségrégation nécessite des familles mais pas d'ADN. Elle consiste à regarder sur un graphique la disposition des personnes en fonction de s'ils sont malades ou non

D. Analyse de liaison

On crible le génome avec des marqueurs et on regarde dans les familles s'il y a un marqueur en particulier qui apparait si on est malade et qui est non présent chez les non malades. On regarde la ségrégation du phénotype et s'il y a un allèle qui ségrége toujours avec le phénotype. Dans ce cas, la région taguée peut jouer un rôle positionnel.

Méthode de Lod Score : famille avec au moins 2 atteints, on regarde si les allèles sont distribués de façon aléatoire ou si un marqueur représente toujours 100% des alleles (dans une famille donnée).

L'analyse de liaison nécessite une famille avec au moins 2 atteints et de l'ADN.

E. Analyse d'association

On isole des portions du génome, sachant qu'il est redondant, les variants le sont aussi. Ainsi on termine avec une région d'environ 300 kb. On regarde les variants de tous les gènes et on teste chez les cas et les contrôles si la fréquence est plus élevée.

Pour trouver le variant causal, il faut aller dans des populations moins redondantes (Afrique). L'analyse d'association nécessite de l'ADN mais pas de famille.

IV) Genome Wide Association Study

On teste toutes les variations du génome, on obtient un Manhattan Blot où chaque point représente un variant. Le rôle de l'individu s'est beaucoup appauvri dans la recherche, on ne génère plus d'idées. C'est désormais la technologie qui décide des hypothèses à tester.