

Méthodes d'analyse familiale pour mettre en évidence des variants génétiques associés à des phénotypes multifactoriels

Florence Demenais

INSERM EMI 00-06, Tour Evry 2, 523 Place des Terrasses de l'Agora, 91034 Evry Cedex

Résumé - La caractérisation de polymorphismes de l'ADN (SNP) sur l'ensemble du génome en nombre sans cesse croissant rend possible l'identification des variants génétiques impliqués dans le déterminisme de maladies multifactorielles par des études d'association. À côté des études cas-témoins classiquement utilisées pour mettre en évidence des associations maladie-marqueur, des méthodes d'analyse familiale ont été proposées pour s'affranchir des problèmes de stratification de population. Ces méthodes peuvent être regroupées en deux types d'approches principales : 1) des méthodes modèles-indépendantes qui ne font aucune hypothèse sur le modèle génétique sous-jacent au phénotype étudié ; 2) des méthodes modèles dépendantes dans lesquelles le modèle génétique est spécifié. Ces méthodes peuvent s'appliquer aussi bien à des phénotypes binaires que quantitatifs et à des structures familiales de taille variable selon les méthodes. Les propriétés de ces différentes approches sont présentées et discutées.

Introduction

Les progrès récents dans les technologies de biologie moléculaire et le séquençage du génome conduisent à caractériser des polymorphismes (variants) de l'ADN en nombre sans cesse croissant. Les marqueurs de choix pour étudier la variabilité du génome sont des SNP (Single Nucleotide Polymorphisms), sites du génome présentant des variations d'un seul nucléotide et observés chez de nombreux individus au sein d'une population. C'est ainsi que depuis avril 1999 s'est mis en place le SNP Consortium regroupant des laboratoires publics et des grands groupes de l'industrie pharmaceutique afin d'identifier entre 300 000 et 500 000 SNP et de réaliser une carte de ces polymorphismes facilement accessible aux chercheurs. En fait, ce consortium a déjà permis d'identifier à ce jour plus d'un million de SNP mais ces polymorphismes nécessitent d'être validés avant de pouvoir être largement utilisés.

Les analyses génétiques de maladies multifactorielles (par exemple, cancers, maladies cardiovasculaires, asthme, maladies neuro-psychiatriques), qui résultent des effets et interactions de nombreux facteurs génétiques et environnementaux, ont conduit à mettre en évidence des régions du génome pouvant contenir des gènes prédisposant à ces pathologies mais jusqu'à présent peu de variants génétiques impliqués directement dans le processus pathogénique ont été identifiés. La mise en évidence de SNP sur l'ensemble du génome offre la possibilité de caractériser ces gènes et le variant génétique (ou ensemble de variants) causal par des études d'association de la maladie avec ces polymorphismes. Ces études d'association reposent sur l'existence d'un déséquilibre de liaison (ou plus exactement déséquilibre gamétique), c'est à dire d'une association préférentielle entre allèles du variant génétique causal (SNP fonctionnel) et allèles de marqueurs génétiques (SNP marqueurs). Ces études d'association sont classiquement réalisées avec des gènes candidats (gènes connus dont la fonction suggère qu'ils peuvent jouer un rôle dans le processus physiopathologique) mais une approche « wide genome search » est actuellement envisagée. En effet, Risch et Merikangas (1996) ont montré que les études d'association avaient une puissance statistique supérieure aux analyses de liaison génétique en particulier pour les gènes ayant un effet modeste et ont suggéré que des études d'association sur l'ensemble du génome étaient possibles. Cependant, cet article, basé sur un certain nombre d'hypothèses, a soulevé de nombreuses controverses dans la littérature. Que ce soit par une approche basée sur des gènes candidats ou systématique sur l'ensemble du génome, les études d'association peuvent être effectuées par des comparaisons cas/témoins ou par des analyses familiales. Les études cas/témoins sont plus facilement réalisables mais l'existence d'une association maladie (phénotype) – marqueur peut non seulement être due à l'existence d'un déséquilibre de liaison entre le marqueur et le variant génétique causal situé à proximité du marqueur mais à des mécanismes de stratification de population

ou de mélange de populations ayant des fréquences différentes des allèles. Pour éviter ce problème de stratification de population, des études familiales ont été proposées avec de nombreux développements au cours de ces dernières années. Ces méthodes d'analyse

familiale peuvent être regroupées en deux types d'approches principales : 1) méthodes modèles-indépendantes qui ne font aucune hypothèse sur le modèle génétique sous-jacent au phénotype étudié ; 2) méthodes modèles dépendantes dans lesquelles le modèle génétique du variant fonctionnel recherché est spécifié. Ces méthodes peuvent s'appliquer aussi bien à des traits binaires que quantitatifs. Nous allons décrire ces différentes approches en essayant de discuter leurs avantages et limites respectifs et en les illustrant par des exemples.

Etudes d'association par des méthodes d'analyse familiale modèles-indépendantes

Traits Binaires

Une des premières méthodes d'analyse familiale qui a été proposée pour mettre en évidence à la fois une liaison génétique et une association est l'approche basée sur le Transmission Disequilibrium Test (TDT) (Spielman et al., 1993). Cette méthode considère des trios constitués de deux parents et d'un enfant atteint et compare le nombre de fois que des parents hétérozygotes pour le marqueur génétique étudié transmettent l'allèle supposé associé à la maladie à leurs enfants atteints au nombre de fois qu'ils transmettent l'autre allèle. Ce test, qui est un simple χ^2 , est significatif s'il existe à la fois une liaison génétique et une association entre le marqueur et le variant causal. Ce test s'applique à des marqueurs bi-alléliques et aux familles où les deux parents sont génotypés. Depuis, de nombreuses extensions du TDT ont été proposées. Pour les marqueurs multi-alléliques, on peut utiliser le TDT classique pour chaque allèle et répéter le test pour tous les allèles mais il est alors nécessaire de corriger le niveau de signification pour le nombre de tests effectués. Un test considérant l'ensemble des allèles simultanément a aussi été proposé : pour chaque parent hétérozygote M_iM_j , on compare le nombre de fois où l'allèle M_i est transmis au nombre de fois où l'allèle M_j est transmis aux enfants atteints. Ce test, qui suit un χ^2 avec $k(k - 1)/2$ degrés de liberté, manque de puissance et les cellules du tableau de contingence ont souvent des effectifs faibles (voir Spielman et Ewens, 1996 pour une discussion). Le TDT peut être appliqué à des familles comportant plusieurs enfants atteints en le répétant pour chaque enfant atteint ou en utilisant des tests s'appliquant à de fratries de même taille. Si l'échantillon comprend un grand nombre de familles avec plusieurs enfants atteints, la signification du TDT reflète plus une liaison génétique qu'une association allélique. Si l'un des deux parents n'est pas génotypé, il est préférable d'exclure ces familles pour éviter des biais. Quand on ne peut obtenir les génotypes parentaux, comme c'est souvent le cas pour des maladies à âge de début tardif, un autre test a été proposé qui utilise non plus les parents mais les germains non atteints des enfants atteints (Spielman et Ewens, 1998). Ce test, connu sous le nom de S-TDT (Sib_TDT), nécessite comme observation minimale une fratrie constituée d'un germain atteint et d'un germain non atteint et ayant des génotypes différents au niveau du marqueur. Le S-TDT compare les fréquences alléliques du marqueur chez les germains atteints et chez les germains non atteints, le niveau de signification de ce test étant obtenu par une procédure de permutation puisque les observations ne sont pas indépendantes. Ce test n'est significatif que s'il existe une liaison génétique. Quand les données incluent à la fois des familles avec parents génotypés et non-génotypés, il est possible de combiner le TDT et le S-TDT dans un test global (Spielman et Ewens, 1998). Des études de puissance ont montré que le TDT avait une puissance d'autant plus élevée que le déséquilibre entre le variant-marqueur et le variant causal est important et que les fréquences alléliques de ces variants sont similaires. Le S-TDT est généralement moins puissant que le TDT. De nombreuses extensions et des approches alternatives du TDT ont été proposées, en particulier une approche basée sur une statistique de scores qui a été généralisée pour prendre en compte simultanément différents types de témoins, parents des enfants atteints, germains non atteints et témoins issus de la population générale et des marqueurs multi-alléliques (Schaid et Rowland, 1998). L'étude de la puissance de cette méthode montre que les approches utilisant des parents comme témoins ou des témoins de populations ont des niveaux de puissance comparables tandis que celle basée sur les germains non atteints est moins puissante. Un variant génétique ayant un effet dominant sur le risque morbide est plus facile à mettre en évidence qu'un variant ayant un effet récessif, le nombre de familles nécessaire pouvant être réduit si on sélectionne des fratries avec au moins deux atteints. De plus, les tailles d'échantillons pour des criblages

systématiques du génome sont 4 à 5 fois plus élevées que celles requises pour l'analyse d'un seul marqueur (Schaid et Rowland, 1998).

Traits quantitatifs

Des approches similaires au TDT ont été développées pour l'analyse des traits quantitatifs (Allison, 1997). Cependant, ces méthodes étaient limitées à certaines structures familiales simples et n'intégraient pas les tests de liaison génétique pour la recherche de QTLs (Quantitative Trait Loci). Une approche intéressante est celle combinant les tests de liaison et d'association, telle que proposée par Fulker et al. (1999) pour des paires de germains et généralisée par la suite. Cette méthode introduit un test d'association dans la méthode d'analyse de liaison basée sur la décomposition de la variance et utilisant le maximum de vraisemblance. La distribution des traits quantitatifs observés dans la i ème fratrie est supposée suivre une distribution multivariée normale avec un vecteur de moyennes (μ_i) et une matrice de variance-covariance (Σ_i). La variance totale est la somme de la variance due au QTL recherché, la variance due à la ressemblance résiduelle entre germains (comprenant la variance polygénique additive et la variance due à l'environnement partagé par les germains) et la variance environnementale résiduelle propre à chaque membre de la fratrie. A une position donnée sur le génome, la covariance entre deux germains au niveau du QTL est exprimée en fonction du paramètre, π , la proportion d'allèles partagés par descendance mendélienne (IBD) à cette position, ceci permettant de tester la liaison génétique. L'association du QTL avec un marqueur candidat bi-allélique peut être modélisée par son effet sur la moyenne, tel que $y_{ij} = \mu + \beta g_{ij}$, où y_{ij} est le phénotype du j ème germain dans la fratrie i , g_{ij} est un score égal à $m_{ij} - 1$ (m étant le nombre d'un des deux allèles du marqueur M), et le paramètre β est égal à zéro sous l'hypothèse nulle d'absence d'association entre le phénotype et le marqueur. Cependant, le test de rapport de vraisemblance ($\beta = 0$ vs $\beta \neq 0$) peut être significatif en présence d'une stratification de population. Pour éviter de mettre faussement en évidence un déséquilibre de liaison, Fulker et al (1999) ont proposé que le score génotypique g_j soit décomposé en deux composantes orthogonales, la composante entre-famille (b) et la composante intra-famille (w), de telle sorte que $y_j = \mu + \beta_b b_i + \beta_w w_{ij}$. La composante entre famille prend en compte le phénomène de stratification de population tandis que la composante intra-famille est significative seulement s'il existe un déséquilibre de liaison. Cette approche permet donc de prendre en compte la liaison génétique dans la structure de covariance et les paramètres d'association par leurs effets sur la moyenne auxquels on peut ajouter les effets de covariables (facteurs de risque connus comme l'âge, le sexe, le mode de vie...). Les tests sont effectués par des rapports de vraisemblance en comparant différents modèles particuliers par rapport au modèle général. De manière intéressante, si le marqueur génétique testé est le QTL lui-même, il n'y a plus d'évidence pour la liaison génétique quand on prend en compte l'association dans le modèle. Donc, une liaison significative, en présence d'association, indique que le marqueur n'est pas le QTL mais est en déséquilibre avec lui. Cette approche a été récemment étendue à des familles nucléaires de taille variable et en prenant en compte les génotypes des parents au niveau du marqueur lorsqu'ils sont disponibles (Abecasis et al., 2000). Des études de puissance de cette méthode indiquent des niveaux de puissance relativement élevés pour mettre en évidence une association d'un marqueur avec un QTL expliquant 10% de la variance du trait quantitatif, ceci d'autant plus que le déséquilibre est important ($D/D_{max} > 50\%$, D étant l'écart entre la fréquence haplotypique et le produit des fréquences alléliques) et que les parents sont génotypés (Abecasis et al., 2000).

Etudes d'association par des méthodes d'analyse familiale modèles-dépendantes

A côté des méthodes précédentes, des approches spécifiant un modèle pour le variant génétique recherché ont été développées permettant de prendre en compte à la fois un déséquilibre de liaison et une liaison. Les modèles régressifs, proposés par G. Bonney, sont des modèles généraux qui spécifient une relation de régression entre le phénotype observé (binaire ou quantitatif) et des variables explicatives incluant l'effet du gène recherché, les phénotypes des antécédents de chaque sujet dans la famille pour prendre en compte des corrélations familiales non spécifiées (dues à d'autres gènes et/ou des facteurs environnementaux partagés) et des facteurs de risque mesurés (covariables). La formulation mathématique de ces modèles est relativement simple et flexible, la distribution multivariée des observations dans une famille étant décomposée en un produit de distributions univariées en conditionnant le phénotype de chaque sujet d'une famille sur le

phénotype de ses antécédents. Ces modèles considèrent classiquement la régression linéaire pour les traits quantitatifs et la régression logistique pour les traits binaires mais une formulation basée sur un modèle à seuil pour les traits binaires a aussi été proposée (Demenais, 1991) et a été étendue à l'analyse de traits polychotomiques. Cette formulation présente des caractéristiques intéressantes par rapport à la formulation logistique originale puisque les phénotypes des antécédents de chaque sujet de la famille peuvent être ajustés pour les effets de leurs propres génotypes et de leurs covariables. Ces modèles régressifs ont été étendus pour prendre en compte des cartes de marqueurs liés (Bonney et al., 1988) ; un déséquilibre de liaison entre gène de susceptibilité à la maladie (ou QTL) et marqueur bi-allélique a aussi été introduit dans ces modèles et le programme REGRESS (Demenais et Lathrop, 1994). Les analyses de la transmission conjointe de phénotypes et de marqueurs (analyses combinées ségrégation-liaison) reposent sur la méthode du maximum de vraisemblance et les tests d'hypothèses sont effectués par des rapports de vraisemblance. L'un des intérêts de cette approche par rapport aux méthodes précédentes est de pouvoir estimer l'effet du variant causal et le déséquilibre de liaison entre ce variant et le(s) marqueur(s). De plus, les modèles régressifs permettent de prendre en compte les effets de facteurs environnementaux pouvant interagir avec le QTL. Ces modèles peuvent être appliqués à des familles de structure variable (familles nucléaires et généalogies). Cette approche apparaît puissante puisque l'analyse de données simulées a permis de mettre en évidence un QTL rendant compte de 5% de la variabilité d'un trait quantitatif multifactoriel (Martinez et al., 1995). Des études plus systématiques de la puissance des modèles régressifs en présence de déséquilibre de liaison sont en cours dans notre équipe. La sélection de familles par des sujets atteints peut augmenter la puissance de détection du QTL et ces modèles apparaissent assez robustes au biais de sélection des familles, surtout quand le déséquilibre de liaison est complet. La présence d'un déséquilibre de liaison gène de susceptibilité-marqueur facilite la mise en évidence d'interactions gène x environnement et la prise en compte de telles interactions peut aussi augmenter la puissance de détection du déséquilibre de liaison et donc conduire à l'identification du variant causal. Des extensions de ces modèles à l'analyse multivariée de phénotypes corrélés pourrait aussi faciliter la mise en évidence de variants ayant un effet pleiotropique sur ces phénotypes.

Application de ces méthodes à l'analyse de maladies multifactorielles et de phénotypes intermédiaires associés

Le TDT est une méthode à l'heure actuelle largement utilisée pour rechercher des variants génétiques associés à des maladies multifactorielles. Ces études ont été réalisées principalement au niveau de gènes candidats et ont souligné la difficulté de mettre en évidence le variant causal. Il apparaît d'ailleurs que ce n'est pas un seul variant mais plutôt une combinaison de variants (haplotypes) qui pourraient être impliqués, comme l'ont suggéré des études récentes de l'asthme (Ober et al., 1998) et de traits quantitatifs associés aux maladies cardiovasculaires (Keavney et al., 1998). A titre d'exemple, nous illustrerons ces différentes méthodes par l'analyse d'associations de polymorphismes situés au niveau d'un gène candidat, le gène *NOS1* (Nitric Oxide Synthase), avec l'asthme et des traits binaires et quantitatifs associés à cette pathologie (Demenais et al., 2001). Ce gène est un gène candidat situé sur le chromosome 12q, région rapportée liée à l'asthme et à des phénotypes intermédiaires par plusieurs criblages du génome dont celui réalisé dans un sous-ensemble de 107 familles de l'étude française EGEA à laquelle nous participons. Les sujets de ces familles ont été génotypés pour trois polymorphismes du gène *NOS1* (exon 29, intron 2 et région promotrice). L'analyse des traits binaires (asthme, tests cutanés aux allergènes) a été effectuée par le TDT et l'analyse des traits quantitatifs (taux d'Immunoglobulines E, taux d'éosinophiles) par des analyses de variance et des analyses combinées ségrégation-liaison basées sur les modèles régressifs. Les analyses basées sur le TDT ont mis en évidence une association du variant 187 bp de l'intron 2 avec des tests cutanés aux allergènes. Les analyses basées sur les modèles régressifs ont suggéré un rôle du variant « 18 répétitions » de l'exon 29 sur la variabilité des éosinophiles, en particulier chez les sujets non-asthmatiques. Ces résultats nécessitant d'être confirmés dans l'échantillon total des familles EGEA et dans d'autres populations.

Conclusion

La réalisation d'une carte dense de SNP couvrant l'ensemble du génome couplée au développement de méthodes statistiques prenant en compte différents facteurs, génétiques et environnementaux, impliqués dans

la variabilité des phénotypes étudiés peuvent faciliter l'identification des déterminants génétiques des maladies multifactorielles. Cependant, étant donné l'effet souvent faible de ces variants et la complexité des mécanismes impliqués, la caractérisation de ces déterminants reste une tâche difficile. L'analyse systématique de tous les SNP au niveau d'un gène donné et l'analyse simultanée de différents gènes intervenant dans le même processus physiologique peuvent faciliter la caractérisation de ces variants génétiques causaux. La collection d'échantillons de taille importante et l'accès à des plateaux techniques permettant un génotypage rapide et peu coûteux de milliers de marqueurs sont aussi des conditions nécessaires pour mener à bien ces études. Les analyses statistiques doivent bien entendu être couplées à des études fonctionnelles pour confirmer le rôle des variants mis en évidence dans le processus pathogénique.

Références bibliographiques

- Abecasis G.R., Cardon L.R., Cookson W.O.C., 2000. A general test of association for quantitative traits in nuclear families. *Am J Hum Genet*, 66, 279-292.
- Allison D.B., 1997. Transmission-disequilibrium tests for quantitative traits. *Am J Hum Genet*, 60, 676-690.
- Bonney G.E., Lathrop G.M., Lalouel J.M., 1988. Combined linkage and segregation analysis using regressive models. *Am J Hum Genet*, 43, 29-37.
- Deménaix F.M., 1991. Regressive logistic models for familial diseases: a formulation assuming an underlying liability model. *Am J Hum Genet*, 49, 773-785.
- Deménaix F.M., Lathrop G.M., 1994. REGRESS: a computer program including the regressive approach into the LINKAGE package. *Genet Epidemiol*, 11:285.
- Deménaix F., Boussaha M., Meunier F., Dizier M.H., 2001. Search for linkage and association of neuronal nitric oxide synthase gene (NOS1) with asthma and asthma-related phenotypes in 107 French EGEA families. *Am J Resp Crit Care Med*, 163, A206.
- Fulker D.W., Chemy S.S., Sham P.S., Hewitt J.K., 1999. Combined linkage and association sib-pair analysis for quantitative traits. *Am J Hum Genet*, 64, 259-267.
- Keavney B., McKenzie C.A., Connell J.M.C., Julier C., Ratcliffe P.J., Sobel E., Lathrop M., Farrall M., 1998. Measured haplotype analysis of the angiotensin-I converting enzyme gene. *Hum Mol Genet*, 7, 1745-1751.
- Martinez M., Abel L., Deménaix F., 1995. How can maximum likelihood approaches indicate the role of a candidate gene in a quantitative trait? *Genet Epidemiol*, 12, 789-794.
- Ober C., Leavitt S.A., Tsalenko A. et al., 2000. Variation in the interleukin 4-receptor alpha gene confers susceptibility to asthma and atopy in ethnically diverse populations. *Am J Hum Genet*, 66, 517-526.
- Risch N., Merikangas K., 1996. The future of genetic studies of complex human diseases. *Science*, 273, 1516-17.
- Schaid D.J., Rowland C., 1998. Use of parents, sibs, and unrelated controls for detection of associations between genetic markers and disease. *Am J Hum Genet*, 63, 1492-1506.
- Spielman R.C., McGinnis R.E., Ewens W.J., 1993. Transmission test for linkage disequilibrium: the insulin region and insulin dependent diabetes mellitus (IDDM). *Am J Hum Genet*, 52, 506-516.
- Spielman R.S., Ewens W.J., 1996. The TDT and other family-based test for linkage disequilibrium. *Am J Hum Genet*, 59, 983-989.
- Spielman R.S., Ewens W.J., 1998. A sibship test in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet*, 62, 450-458.

