Méthodes d'analyse génétique de traits quantitatifs corrélés: application à l'étude de la densité minérale osseuse

Aude Saint Pierre

UNIVERSITÉ PARIS XI
FACULTÉ DE MÉDECINE PARIS-SUD

Année 2011

N° attribué par la bibliothèque

THESE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE PARIS XI

Champ disciplinaire : Biostatistique

Ecole Doctorale 420 : Santé Publique

présentée et soutenue publiquement le 3 janvier 2011
par

Aude SAINT PIERRE

Titre :

Méthodes d'analyse génétique de traits quantitatifs corrélés : application à l'étude de la densité minérale osseuse

Directrice de thèse : Maria MARTINEZ

JURY

Madame	Françoise Clerget-Darpoux	Présidente
Monsieur	Edouardo Manfredi	Rapporteur
Monsieur	Bertram Müller-Myhsok	Rapporteur
Madame	Emmanuelle Génin	Examinateur
Madame	Pascale Leroy	Examinateur
Madame	Maria Martinez	Directrice de thèse

Remerciements

En premier lieu, je tiens à remercier Maria Martinez pour m'avoir encadré durant cette thèse. Elle a su faire preuve de beaucoup de compréhension et parfois, lorsque je faisais mes premier pas en génétique, d'une grande patience face à mon manque de connaissance. J'ai ainsi pu découvrir grâce à elle, le petit milieu des généticiens, que ce soit lors de nos multiples congrès ou durant nos longues heures de discussion. Grâce à sa rigueur scientifique et ses compétences statistiques en génétique, elle a su me donner l'envie de poursuivre dans ce domaine de recherche. Je la remercie pour son soutien scientifique sans lequel je n'aurais jamais pu réaliser tout ce travail, pour sa disponibilité et pour toute la confiance qu'elle m'a accordée durant ces années de thèse. Elle m'a appris énormément de choses que ce soit dans le contexte professionnel ou bien au-delà, je lui exprime toute mon estime.

Je suis également très reconnaissante envers Eduardo Manfredi et Bertram Müller-Myhsok pour l'intérêt qu'ils ont bien voulu accorder à cette thèse en acceptant d'évaluer ce travail. Je les remercie très chaleureusement pour leur extrême compréhension face à tous les évènements qui ont retardé l'envoi du manuscrit.

J'adresse mes remerciements sincères à Françoise Clerget, Emmanuelle Génin et Pascale Le Roy pour avoir accepté de participer à ce jury.

Je tiens à exprimer mes remerciements à Marie Paul Roth et Hélène Coppin pour m'avoir accueilli dans l'unité INSERM 563. Dans cette ambiance aux thèmes multiples et variés, j'ai énormément apprécié, en tant que statisticienne, de pouvoir naviguer avec beaucoup de liberté au sein de leur équipe.

Certains m'ont particulièrement aidé ou encouragé dans l'élaboration de ce travail. Je pense à Nora, Mathieu, qui coule des beaux jours au soleil, et bien sûr Mohamad. Je ne vous remercierais jamais assez pour votre optimisme et votre soutien ainsi que pour m'avoir toujours accompagné par tous les temps lors de nos multiples pauses café ! Vous y êtes pour beaucoup dans ce travail, autant sur le plan humain que scientifique. Du rire aux larmes … personne ne sait autant que toi, Mohamad, ce qu'il y a d'invisible entre ces lignes. Il est parfois des personnes dont la simple présence apaise et tu en fais partie. Je te souhaite de poursuivre cette longue aventure avec autant de brio que ce que tu m'as déjà montré. Dommage que vous soyez arrivés si tard au sein de l'équipe, le temps nous as manqué, mais j'espère que l'histoire ne s'arrêtera pas là…

Je décerne une mention spéciale à la petite famille réunionnaise, Florence & co, pour tous les bons moments passés. Les balades en montagne et les longues discussions sur la varangue autour d'une « dodo » resteront pour moi des moments inoubliables. Je suis heureuse de continuer à vous voir malgré la distance…peut être au Burkina la prochaine fois.

Je remercie également Julie pour les nombreuses soirées autant survoltées que mémorables, Caroline pour m'avoir fait découvrir l'Indonésie, Fanny, Christophe & Cristina pour tous les bons moments passés ensemble.

Un grand merci à la famille d'Olivier qui m'a toujours apporté son soutien : Thérèse & Philippe, Bénédicte & Marino ainsi qu'à leurs deux petits monstres.

Je voudrais remercier spécialement ma famille qui m'a toujours soutenu. Mes parents qui ont toujours cru en moi et qui ont su m'encourager et me soutenir dans les moments les plus difficiles. Mes deux frères Guillaume et Philippe, qui fort de leurs expériences de « thésards » pimentés à la sauce enseignant-chercheur, ont su me conseiller et m'aider en urgence dès que j'en avais besoin. Ces deux-là ne seraient sûrement pas ce qu'ils sont sans les deux plus agréables belle-sœur Céline et Isabelle à qui je souhaite les plus belles aventures dans de nombreux voyages au bout du monde. Merci Joseph pour tes compétences inimaginables aussi bien en statistique qu'en politique. Merci Michel de nous faire profiter de tes bois à champignons et de ton luxuriant potager quel que soit la saison.

Ces remerciements ne seraient pas complets si je n'évoquais pas la présence et le soutien inconditionnel d'Olivier. Il m'a supporté au quotidien et m'a toujours encouragé dans les moments difficiles qui jalonnent une thèse, et d'ailleurs celle-ci n'existerait probablement pas s'il n'avait pas toujours été présent. Je ne te remercierai jamais assez pour tout ce que tu as fait. Tu en as profité pour devenir un excellent cuisinier, il ne te reste plus qu'à découvrir la gastronomie italienne !

La rédaction d'une thèse donne la chance de pouvoir remercier toutes les personnes qui ont rendu ce travail possible. Familles, collègues, ou amis qui m'avait si souvent écouté et encouragé, si j'en suis arrivée là, c'est aussi grâce à vous. Puissent ces quelques lignes exprimer l'ampleur de ma reconnaissance et vous rendre un peu de ce bonheur que vous m'avez transmis.

Aude

Résumé

La plupart des maladies humaines ont une étiologie complexe avec des facteurs génétiques et environnementaux qui interagissent. Utiliser des phénotypes corrélés peut augmenter la puissance de détection de locus de trait quantitatif. Ce travail propose d'évaluer différentes approches d'analyse bivariée pour des traits corrélés en utilisant l'information apportée par les marqueurs au niveau de la liaison et de l'association. Le gain relatif de ces approches est comparé aux analyses univariées. Ce travail a été appliqué à la variation de la densité osseuse à deux sites squelettiques dans une cohorte d'hommes sélectionnés pour des valeurs phénotypiques extrêmes. Nos résultats montrent l'intérêt d'utiliser des approches bivariées en particulier pour l'analyse d'association. Par ailleurs, dans le cadre du groupe de travail GAW16, nous avons comparé les performances relatives de trois méthodes d'association dans des données familiales.

Mots-clés

Génétique statistique, trait complexe, analyse bivariée, échantillons sélectionnés, puissance statistique, analyse pangénomique, analyse de liaison, analyse d'association, QTL, DMO, ostéoporose.

Coordonnées du laboratoire d'accueil

INSERM U563 - CPTP
CHU Purpan, BP 3028
31024 Toulouse CEDEX
FRANCE

Title

Statistical methods for genetic analysis of correlated quantitative traits: application to the study of bone mineral density.

Abstract

The majority of complex diseases in humans are likely determined by both genetic and environmental factors. Using correlated phenotypes may increase the power to map the underlying Quantitative Trait Loci (QTLs). This work aims to evaluate and compare the performance of bivariate methods for detecting QTLs in correlated phenotypes by linkage and association analyses. We applied these methods to data on Bone Mineral Density (BMD) variation, measured at the two skeletal sites, in a sample of males selected for extreme trait values. Our results demonstrate the relative gain, in particular for association analysis, of bivariate approaches when compared to univariate analyses. Finally, we study the performances of association methods to detect QTLs in the GAW16 simulated family data.

Keywords

Statistical genetics, complex trait, bivariate analysis, sample selection, statistical power, genome-wide analysis, linkage analysis, association analysis, QTL, BMD, osteoporosis.

Contact information of the hosting lab

INSERM U563 - CPTP
CHU Purpan, BP 3028
31024 Toulouse CEDEX
FRANCE

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

Un des objectifs de l'épidémiologie génétique est de caractériser la composante génétique de la variabilité de traits chez l'homme. Les traits complexes tels que l'ostéoporose, certains cancers ou le diabète sont des problèmes de santé publique à cause de la gravité des symptômes et du grand nombre de personnes touchées dans la population générale. De nombreuses recherches portent donc sur le développement de méthodes d'identification des loci génétiques impliqués dans la variabilité des traits complexes. L'objet de la thèse porte sur la comparaison des performances de différentes méthodes pour l'analyse génétique de traits quantitatifs.

Encouragé par le succès de la cartographie génétique pour des traits mendéliens, tels que la maladie de Huntington, certaines formes de cancers du sein ou de la maladie d'Alzheimer, les efforts se sont tournés vers la cartographie de gènes pour des traits complexes dont l'étiologie est multifactorielle. Les traits mendéliens, ou monogéniques, sont caractérisés par l'effet de mutation(s) rare(s) dans la population générale ayant des effets forts sur la variabilité du trait. Les traits complexes sont plutôt expliqués par plusieurs gènes interagissant entre eux et avec des facteurs de l'environnement. Ce ne sont pas des mutations délétères rares, mais plutôt des variants, généralement fréquents, dont la présence n'est ni nécessaire ni suffisante au développement de la maladie chez un individu.

Il existe deux approches méthodologiques, basées fondamentalement sur le même principe, permettant de détecter des facteurs génétiques de maladies : les études de liaison et les études d'association. Ces approches sont des outils puissants et complémentaires pour caractériser la composante génétique des maladies. Généralement, on utilise dans un premier temps les études de liaison pour localiser des régions chromosomiques pouvant contenir un gène expliquant une part de la variabilité du trait. Puis, dans un deuxième temps, les études d'association pour préciser plus finement l'emplacement du gène.

Ces dernières années on a assisté à une véritable explosion des études d'association à grande échelle, permettant une recherche non exhaustive des polymorphismes génétiques pouvant être impliqués dans les mécanismes biologiques à l'origine de traits complexes. Ceci s'est fait en parallèle avec les progrès techniques de la biologie, qui ont largement contribués à la diminution à la fois du coût et du temps de génotypage. Bien qu'une grande partie des régions chromosomiques identifiées ne contiennent pas de gènes codant

pour des protéines (moins de 5% du génome code pour des protéines), ces études ont permis de révéler de nouveaux gènes-candidats potentiels pour de nombreuses maladies.

La plupart du temps, plusieurs traits reliés au phénotype d'intérêt sont collectés pour un même individu. Ces traits sont souvent corrélés entre eux et une part de cette corrélation pourrait être expliquée par des facteurs génétiques communs. Analyser conjointement ces traits peut s'avérer plus puissant que l'analyse séparée de chacun des traits pour la recherche de loci de susceptibilité. Plusieurs méthodes d'analyses multivariées pour des traits corrélés ont été développées. Deux axes principaux existent, l'analyse de liaison multivariée peut se faire soit par une analyse conjointe des traits, soit sur des méthodes de réduction de la dimension telle que l'analyse en composantes principales. Les méthodes bivariées sont souvent utilisées dans le cadre de la recherche de loci par la liaison génétique, mais elles sont moins répandues dans le cadre de la recherche par l'association.

Au cours de cette thèse, nous avons cherché à évaluer l'intérêt que peut présenter l'analyse jointe de traits quantitatifs corrélés en mesurant le gain apporté par l'analyse jointe relativement aux analyses univariées.

Ce manuscrit s'organise principalement autour de quatre grands chapitres.

Le chapitre 1 permet de poser les fondements statistiques de la génétique quantitative nécessaires à la compréhension de ce manuscrit.

Nous présentons à cette occasion les différentes méthodes d'analyse possibles pour la recherche de liaison ou d'association. Nous introduisons également dans ce chapitre la part de la corrélation phénotypique observée entre des traits quantitatifs et expliquée par la composante génétique ; notions de base qui serviront pour toutes les méthodes d'analyses bivariées utilisées par la suite. Nous faisons également un état de la littérature sur la génétique de l'ostéoporose et de ses traits associés, en particulier la densité osseuse.

La suite du manuscrit détaille plus spécifiquement le travail de recherche réalisé au cours de cette thèse. Dans chaque partie, les méthodes utilisées sont présentées.

Le chapitre 2 présente l'ensemble du travail réalisé dans la recherche de QTL par des méthodes de liaison. Nous présentons les résultats du criblage du génome pour la liaison de la variabilité interindividuelle de la densité minérale osseuse (DMO), dans les données familiales NEMO (NEtwork on Male Osteoporosis ; J.M. Kaufam, U Gent, Belgique ; MC de Vernejoul, U606, Lariboisère ; M. Martinez, U563, Toulouse). Pour chaque participant de cette étude, l'état de l'os (DMO) a été évalué à deux sites différents du squelette : au rachis lombaire (LS, pour Lumbar Spine) et au col du fémur (pour Femoral Neck). L'originalité de la base de données NEMO est d'utiliser un échantillon de familles sélectionnées par des hommes ayant une faible DMO au site LS ou FN. Les analyses ont été conduites à l'aide de tests de liaison non-paramétriques (basés sur la décomposition de la variance) et pour chaque phénotype (LS ou FN) indépendamment l'un de l'autre.

Comme ces deux traits sont fortement corrélés, nous avons émis l'hypothèse qu'une part de cette corrélation est induite par des effets de facteurs génétiques communs. Ce travail a donc été étendu à l'analyse de liaison bivariée (analyse jointe des deux phénotypes LS et FN). Les deux principales approches de liaison bivariées reposent sur les phénotypes d'intérêts (LS et FN) ou sur des combinaisons linéairement indépendantes de ces phénotypes. Les paramètres du test de la liaison génétique, utilisant les phénotypes d'intérêt, sont les variances et la covariance entre les traits.

La non indépendance des paramètres de liaison induit une réduction de l'espace des paramètres sous l'hypothèse nulle. Ce phénomène fait que la distribution asymptotique du test de liaison bivarié est complexe et non connue à ce jour. Plusieurs distributions théoriques ont été proposées dans la littérature, mais pour la plupart, la validité n'est pas prouvée. Nous avons donc comparé, par simulations, ces différentes distributions dans nos données.

Le chapitre 3 présente l'ensemble du travail pour la recherche de loci par association. Nous commençons par introduire les principaux problèmes liés à l'analyse d'association à grande échelle (GWAS) pour la variation de traits complexes. Nous présentons ensuite les résultats obtenus dans le cadre de notre étude d'association GWAS en population. L'échantillon NEMO est constitué d'hommes non apparentés selectionnés pour des valeurs basses ou élevées de la DMO. Deux approches ont été développées : les traits LS et FN ont été analysés simultanément ou indépendamment. Ces résultats nous ont conduits à penser que l'analyse d'association jointe pouvait être une approche pertinente pour les études à grande échelle. Pour évaluer la puissance des tests d'association joints, nous avons conduit une étude de simulation dans des sujets en population recensés selon des critères variables vis-à-vis de la valeur des phénotypes étudiés. Ces résultats suggèrent que l'analyse jointe peut être une approche puissante pour détecter des loci de traits quantitatifs.

Le chapitre 4 aborde l'étude de test d'association dans des données de sujets apparentés. Ce travail a été développé dans le cadre du groupe de travail GAW (Genetic Analysis Workshop 16). Afin de tenir compte des relations d'apparentés entre les individus, une approche intéressante pour tester l'association est d'utiliser le modèle de décomposition de la variance. Une des méthodes largement utilisée est celle utilisant une décomposition orthogonale des scores génotypiques. Afin d'augmenter le nombre d'individus effectivement pris en compte pour tester l'association, plusieurs méthodes alternatives existent. Nous avons comparé les performances relatives de trois méthodes d'association pour des traits quantitatifs.

Chapitre 1

1. Génétique de traits quantitatifs : définitions et notations

1.1. Relation trait – gène

Nous allons présenter dans cette partie comment modéliser l'effet d'un gène sur un trait quantitatif Y. Nous considérons des échantillons d'individus indépendants que nous étendrons à des échantillons d'individus apparentés (famille nucléaire ou généalogie).

Y est une variable aléatoire suivant une distribution de moyenne $E(Y)$ et de variance σ_Y^2. On suppose un gène G expliquant une part de la variabilité du trait Y tel que :

$$Y = G + E$$

où E est une variable aléatoire résiduelle

Les variables G et E sont supposées indépendantes. G à deux allèles A1 et A2 tel que :

- l'allèle A1 soit de fréquence p
- l'allèle A2 soit de fréquence q et $p + q = 1$

Avec 2 allèles, il y a 3 génotypes possibles. Sous l'hypothèse d'Hardy-Weinberg, les 3 génotypes A2A2, A1A2 et A1A1, sont présents dans la population avec les probabilités respectives q^2, 2pq, p^2.

Nous notons $G_k, k = 1,2,3$ les trois génotypes :

$$\begin{array}{lll} G_3 = A2A2 & & P(G=G_3) = P_{G_3} = q^2 \\ G_2 = A1A2 & \text{tel que} & P(G=G_2) = P_{G_2} = 2pq \\ G_1 = A1A1 & & P(G=G_1) = P_{G_1} = p^2 \end{array} \qquad \textbf{(1.1)}$$

L'équilibre d'Hardy-Weinberg (HW) est un principe fondamental en génétique qui soutient que les fréquences génotypiques à un locus donné restent constantes de génération en génération si les quatre conditions suivantes sont respectées :

- la population est de taille infinie
- les unions sont aléatoires, le choix d'un partenaire ne dépend pas de son génotype (hypothèse de panmixie)
- il n'y a pas de sélection dans la population
- il n'y a ni mutation, ni migration dans la population

L'espérance, notée M , et la variance, notée $\sigma_G^2(Y) = \sigma_G^2$, de la loi de la variable aléatoire Y conditionnellement au génotype G sont modélisées à partir de quatre paramètres qui sont la fréquence de l'allèle A1 (p) et les trois **moyennes génotypiques** $\mu_{G_k} = E(Y \mid G = G_k) \quad k = 1,2,3$. On a :

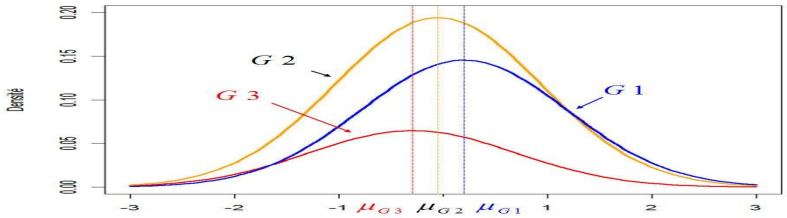
$$\begin{aligned} M &= \sum_{k=1}^{3} P_{G_k} \mu_{G_k} \\ &= p^2 \mu_{G_1} + 2pq\mu_{G_2} + q^2 \mu_{G_3} \end{aligned} \qquad \textbf{(1.2)}$$

$$\begin{aligned} \sigma_G^2 &= \sum_{k=1}^{3} P_{G_k} \left(\mu_{G_k} - M\right)^2 \\ &= p^2(\mu_{G_1} - M)^2 + 2pq(\mu_{G_2} - M)^2 + q^2(\mu_{G_3} - M)^2 \end{aligned} \qquad \textbf{(1.3)}$$

Nous représentons l'effet d'un gène sur un trait en suivant l'approche classique décrite par Falconer (Falconer and Mackay 1996).

Les distributions du trait Y dans la population conditionnellement au génotype G_k sont montrées dans la figure 1.

Figure 1 : Distribution de Y conditionnellement aux génotypes G1 (bleu), G2 (orange) et G3 (rouge) de moyennes respectives $\mu_k, k = 1,2,3$.

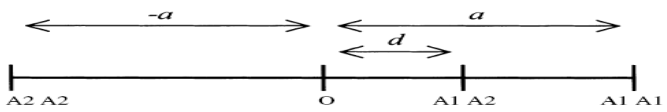


Effets additifs et d'interactions des allèles

Afin de simplifier les calculs, on se ramène par une translation aux moyennes génotypiques $(\mu_{G3}, \mu_{G2}, \mu_{G1}) = (-a, d, a)$ (Figure 2). On pose :

- $\mu_{G3} = -a$ la moyenne du trait chez les homozygotes A2A2
- $\mu_{G2} = d$ la moyenne du trait chez les hétérozygotes A1A2.
- $\mu_{G1} = a$ la moyenne du trait chez les homozygotes A1A1

Figure 2 : Illustration graphique des moyennes génotypiques $(\mu_{G_k})_{k=1,2,3}$.



d reflète les effets d'interactions entre les allèles A1 et A2. C'est-à-dire que l'effet d'un allèle dépend de l'état de l'autre allèle. On distingue quelques cas typiques. Par exemple :

- Si $d > 0$ alors l'effet de l'allèle A1 est plus important si l'autre allèle est A1 que si c'est l'allèle A2.
- Si $d > a$: l'effet de l'allèle A1 est plus important si l'autre allèle est A2 que si c'est l'allèle A1.
- Si $d = 0$ alors l'effet d'un allèle est indépendant de l'autre. On dit que les effets sont additifs.

A partir de ces nouvelles notations, on peut réécrire M et σ_G^2 des équations (1.2) et (1.3) :

$$M = a(p-q) + 2pqd$$

et $$\sigma_G^2 = p^2(a-M)^2 + 2pq(d-M)^2 + q^2(-a-M)^2$$

Remarque : M est l'addition de deux termes : le premier terme $a(p-q)$, attribué à l'effet des homozygotes et le second terme $2pqd$, attribué aux hétérozygotes A1A2. S'il n'y a pas d'interaction entre les effets des allèles A1 et A2 ($d=0$) :

$$M = a(p-q)$$

et $\sigma_G^2 = 2pqa^2$

Supposons maintenant qu'un individu soit porteur de l'allèle A1, on cherche à calculer la moyenne du trait Y chez cet individu sachant qu'il porte déjà l'allèle A1. On note cette quantité : $E(Y \mid Allèle1 = A1)$.

Pour l'autre allèle ($Allèle2$) de son génotype G_k, il reçoit l'allèle A1 avec une probabilité p et l'allèle A2 avec une probabilité q :

$P(G = G_1 \mid Allèle1 = A1) = p$, $P(G = G_2 \mid Allèle1 = A1) = q$ et $P(G = G_3 \mid Allèle1 = A1) = 0$

On a donc :

$$\begin{aligned} E(Y \mid Allèle1 = A1) &= \sum_{i=1}^{3} P(G = G_k \mid Allèle1 = A1) E(Y \mid Allèle1 = A1, G = G_k) \\ &= P(G = G_1 \mid Allèle1 = A1)\mu_{G_1} + P(G = G_2 \mid Allèle1 = A1)\mu_{G_2} \\ &= pa + qd \end{aligned}$$

De la même manière, on obtient : $E(Y \mid Allèle1 = A2) = -aq + pd$

En standardisant ces deux équations par rapport à la moyenne dans la population on obtient l'**effet moyen de l'allèle** A_j (j=1,2) de la forme $\alpha_j = E(Y \mid Allèle1 = Aj) - M$

Avec :

$$\begin{aligned} \alpha_1 &= q(a + d(q-p)) \\ \alpha_2 &= -p(a + d(q-p)) \end{aligned}$$

Si on pose $\alpha = a + d(q-p)$, on voit que :

$$\begin{cases} \alpha = \alpha_1 - \alpha_2 \\ \alpha_1 = q\alpha \\ \alpha_2 = -p\alpha \end{cases}$$

A partir de ces effets moyens alléliques, on peut réécrire M et σ_G^2 (équations (1.2) et (1.3)) :

$$M = -(\alpha_1 + \alpha_2) + d(1-2pq)$$

$$\sigma_G^2 = p^2(a-M)^2 + 2pq(d-M)^2 + q^2(-a-M)^2$$

et $$= 2pq\alpha^2 + (2pqd)^2$$

$$= \sigma_A^2 + \sigma_D^2$$

En absence d'épistasie, la variance génétique s'écrit comme la somme de la variance des effets additifs σ_A^2 et de la variance des effets de dominance (effets d'interactions) σ_D^2.

On a modélisé l'effet du gène en fonction de quatre paramètres (α_1, α_2, d, p) résultant des effets moyens additifs, des effets moyens d'interactions et de la fréquence de l'allèle.

On montre dans le tableau 1 la décomposition des effets des allèles en termes des effets additifs, notés α_A, et en termes des effets d'interactions des allèles, notés α_D.

Tableau 1 : Effets additifs et d'interactions des allèles

G_i	P_{G_i}	μ_{G_i}	
		α_A	α_D
A2A2	q^2	$2x(\alpha_2) = -2p\alpha$	$-2p^2d$
A1A2	$2pq$	$\alpha_1 + \alpha_2 = (q-p)\alpha$	$2pqd$
A1A1	p^2	$2x(\alpha_1) = 2q\alpha$	$-2q^2d$

Les moyennes des effets additifs et des effets de dominance sont nuls :

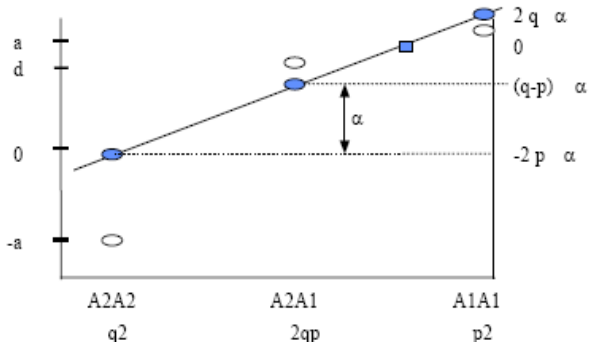
$$E(\alpha_A) = p^2 \times 2q\alpha + 2pq(q-p) \times \alpha - q^2 \times 2p\alpha = 0$$

De la même manière $E(\alpha_D) = 0$.

Il n'y a pas non plus de covariance entre les effets des allèles.

$$\begin{aligned} \text{cov}(\alpha_A, \alpha_D) &= E(\alpha_A \alpha_D) \\ &= -4q^3p^2\alpha d + 4p^2q^2(q-p)\alpha d + 4q^2p^3\alpha d \\ &= 0 \end{aligned}$$

Figure 3 : Représentation graphique des effets moyens $(\mu_{G_k})_{k=1,2,3}$ (cercles blancs) et des effets moyens additifs α_A (cercles bleus) sachant le génotype G_k.



Cas particulier : On peut écrire M, σ_A^2, σ_D^2 et σ_G^2 dans le cas où d ($=\mu_{G_2}$) prend des valeurs particulières.

μ_{G_2}	M	σ_A^2	σ_D^2	σ_G^2
$-a$	$a(p-q)-2apq$	$8qp^3a^2$	$2pqa^2$	$2pqa^2(4p^2+1)$
0	$a(p-q)$	$2pqa^2$	0	$2pqa^2$
a	$a(p-q)+2apq$	$8pq^3a^2$	$2pqa^2$	$2pqa^2(4q^2+1)$

Les traits binaires sont un exemple particulier, dans lequel on considère qu'un individu exprime ou non le caractère à partir d'un certain seuil. Les paramètres du modèle sont la fréquence de l'allèle p, le vecteur des pénétrances (f_{DD}, f_{Dd}, f_{dd}) tel que :

$$\begin{cases} f_{DD} = P(\text{Malade}|\text{DD}) \\ f_{Dd} = P(\text{Malade}|\text{Dd}) \\ f_{dd} = P(\text{Malade}|\text{dd}) \end{cases}$$

Ressemblance entre apparentés

Nous venons de caractériser l'effet d'un gène G sur un trait Y dans une population à l'équilibre d'HW par sa moyenne M et sa variance σ_G^2. Nous présentons maintenant l'extension dans une population d'individus apparentés. Deux individus sont apparentés s'ils ont un ancêtre commun.

La ressemblance phénotypique entre deux individus apparentés est mesurée par la covariance du trait. S'il existe un gène expliquant une part de cette variabilité

phénotypique commune, du fait du lien d'apparentement entre ces individus, une part de cette covariance est expliquée par une composante génétique commune.

Soit (G_i, G_j) les variables aléatoires du gène G chez deux individus apparentés i et j, la **covariance génétique** est :

$$\begin{aligned}\text{cov}(G_i, G_j) &= E\left[E(Y \mid G_i)E(Y \mid G_j)\right] - M^2 \\ &= E\left[(\alpha_{A1} + \alpha_{D1} + M)(\alpha_{A2} + \alpha_{D2} + M)\right] - M^2 \\ &= \text{cov}(\alpha_{A1}, \alpha_{A2}) + \text{cov}(\alpha_{D1}, \alpha_{D2})\end{aligned}$$

- $\text{cov}(\alpha_{A1}, \alpha_{A2})$ est non nulle si les deux individus ont un allèle en commun issu d'un ancêtre commun et correspond au **coefficient d'apparentement** (*kinship* coefficient), noté Φ_{ij}, entre deux individus i et j. C'est la probabilité pour que deux allèles tirés au hasard au même locus, l'un chez i et l'autre chez j, soient identiques par descendance (**IBD** : *Identity By Descent***)**.
  Si on note π_{ijk} la probabilité a priori pour qu'une paire d'individus apparentés i et j partagent k = 0, 1 ou 2 allèles identiques par descendance (IBD), on a : $\Phi_{ij} = \frac{1}{2} \times E\left(1/2 \times \pi_{ij1} + \pi_{ij2}\right)$.
  On note $2\text{x}\Phi_{ij}$ le coefficient de parenté (*Relationship* coefficient) entre deux individus i et j. On a $\text{cov}(\alpha_{A1}, \alpha_{A2}) = \sigma_A^2 \text{x} 2\Phi_{ij}$.

- $\text{cov}(\alpha_{D1}, \alpha_{D2})$, fonction de la variance des effets de dominance, est non nulle si la paire d'allèles à un locus est identique par descendance (IBD). La probabilité de cet événement est notée $P2$ et $\text{cov}\left(\alpha_{D1}, \alpha_{D2}\right) = \sigma_D^2 \times P2$.

Dans le tableau 2 nous montrons les valeurs des coefficients Φ_{ij} et de $P2$ ainsi que la covariance génétique pour les principaux types d'apparentés.

Tableau 2 : Covariance génétique entre deux individus selon le type d'apparenté

Paires d'apparentés	Φ_{ij}	$P2$	$\text{cov}(G_i, G_j)$
Jumeaux monozygotes	$1/2$	1	$\sigma_A^2 + \sigma_D^2$
Germains	$1/4$	$1/4$	$1/2\ \sigma_A^2 + 1/4\ \sigma_D^2$
Parent - Enfant	$1/4$	0	$1/2\ \sigma_A^2$
Oncle - Neveux	$1/8$	0	$1/4\ \sigma_A^2$
Cousins germains	$1/16$	0	$1/8\ \sigma_A^2$

La composante de la variance des effets de dominance est en général plus petite que la composante additive. Les seuls types d'apparentés informatifs pour son estimation sont les germains. Elle est souvent supposée nulle lorsque l'on travaille dans des familles non nucléaires.

On définit le **coefficient de consanguinité** (*imbreeding* coefficient), propre à l'individu, par la probabilité pour que les deux allèles que possède un individu en un locus donné soient identiques par descendance (IBD).

Modèle plus complexe pour Y

De manière générale, les traits quantitatifs ne sont pas simplement expliqués par l'effet d'un seul facteur mais par plusieurs. On prend aussi en compte des facteurs familiaux commun résultant de l'effet de nombreux facteurs qui peuvent être par exemple des gènes et qu'on appelle l'**effet polygénique**.

La valeur du trait Y pour un individu s'écrit sous la forme suivante :

$$Y = G + C + E \qquad \textbf{(1.4)}$$

où G, C et E sont les composantes des effets génétiques; polygéniques et résiduels de moyennes nulles et de variances σ_G^2, σ_C^2 et σ^2. On suppose que G, C et E sont non corrélées $\text{cov}(G,E) = \text{cov}(G,C) = \text{cov}(C,E) = 0$.

On peut également prendre en compte l'effet des covariables observées.

La moyenne et la variance de Y sont :

$$\begin{cases} E(Y) = E(G) \\ \sigma_Y^2 = \sigma_G^2 + \sigma_C^2 + \sigma^2 \end{cases}$$

On définit l'**héritabilité** par la part de la variance du trait expliquée par la composante polygénique:

$$h^2 = \frac{\sigma_C^2}{\sigma_Y^2}$$

La covariance entre des apparentés i et j est de la forme :

$$\begin{aligned} \text{cov}(Y_i, Y_j) &= \text{cov}(G_i + C_i + E_i, G_j + C_j + E_j) \\ &= \text{cov}(G_i, G_j) + \text{cov}(C_i, C_j) \qquad si\ \text{cov}(E_i, E_j) = 0 \\ &= \sigma_A^2 \times \Pi_{ij} + \sigma_C^2 \times 2\Phi_{ij} \end{aligned}$$

où $\Pi_{ij} = \frac{1}{2} \times \pi_{ij1} + \pi_{ij2}$ est la matrice des coefficients de parenté c'est-à-dire la probabilité de partager un allèle IBD entre les apparentés i et j au locus du gène et π_{ijk} est la probabilité de partager k=0,1 ou 2 allèles IBD. Si l'on note Ω_{ij} la matrice de variance-covariance phénotypique entre des individus (i, j) apparentés, on a :

$$\Omega_{ij} = \begin{cases} \sigma_A^2 + \sigma_C^2 + \sigma^2 & \text{si i = j} \\ \sigma_A^2 \times \Pi_{ij} + \sigma_C^2 \times 2\Phi_{ij} & \text{si i} \neq \text{j} \end{cases} \qquad \textbf{(1.5)}$$

où Φ_{ij} est le coefficient d'apparentement entre des apparentés i et j ; σ_A^2 est la variance des effets additifs du gène G ; σ_C^2 est la variance des effets polygéniques C et σ^2 est la variance des effets résiduels aléatoires E. On suppose que les variables G, C et E sont non corrélées.

1.2. Marqueurs génétiques

Des marqueurs génétiques sont des séquences polymorphes d'ADN dont les positions exactes sur le génome sont connues. On dit qu'un locus est polymorphe s'il existe au moins deux allèles dans la population. Parmi les marqueurs les plus utilisés on trouve les Short Tandem Repeat Polymorphisms (STRPs ou microsatellites), ce sont des enchaînements de quelques bases qui se répètent un certain nombre de fois. Un autre type

de marqueurs couramment utilisé est le polymorphisme d'une seule base, les Single Nucleotide Polymorphisms (SNPs). Ils sont de manière générale bi-alléliques.

Un marqueur peut apporter de l'information à deux niveaux :

- **au niveau familial**, si les allèles des deux loci se transmettent de façon non indépendante au cours des générations. On évalue l'indépendance de transmission des allèles à l'aide d'une **analyse de liaison** génétique. Les marqueurs utilisés pour l'analyse de liaison sont en général des marqueurs très polymorphes comme par exemple les microsatellites.

- **au niveau de la population**, s'il existe une association préférentielle (le déséquilibre de liaison) entre les allèles du variant génétique causal et les allèles de marqueurs génétiques. On évalue la force de cette association à l'aide d'une **analyse d'association**. On utilise en général des SNPs (Single Nucleotide Polymorphisms).

Liaison génétique et recombinaison génétique

Considérons deux loci A et B. Ces deux loci peuvent être sur la même paire de chromosomes ou sur des paires de chromosomes différents. Si les deux loci sont situés sur des chromosomes différents ou lorsqu'ils sont situés sur le même chromosome mais éloignés l'un de l'autre alors on dit que les loci sont **indépendants.** En revanche, si ces deux loci sont proches l'un de l'autre, et donc lorsqu'ils sont localisés dans la même région chromosomique, on dit que les deux loci sont **liés génétiquement.** La liaison génétique reflète cette distance entre les loci. Cette distance génétique est le résultat du phénomène de recombinaison ou crossing-over.

La **recombinaison** est un phénomène résultant du mélange de matériel génétique qui se produit par enjambement entre chromosomes. Elle survient au cours de la méiose, le processus de formation des gamètes mâles, les spermatozoïdes, et des gamètes femelles, les ovules. Chaque chromosome a alors la possibilité d'échanger une partie d'ADN avec son chromosome homologue. Plus les deux loci sont proches, moins il y a de chances qu'il y ait une recombinaison entre les deux.

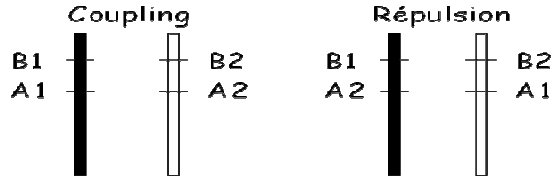
On représente la probabilité de recombinaison sur une distance donnée par l'unité de distance génétique : centiMorgan (cM). 1 cM correspond à environ 1% de recombinaison c'est-à-dire une recombinaison en moyenne pour 100 méioses.

L'équivalence entre la distance génétique et distance physique entre deux loci varie selon l'espèce considérée. Chez l'homme, on admet : $1cM \approx 1Mb$ (Méga Base).

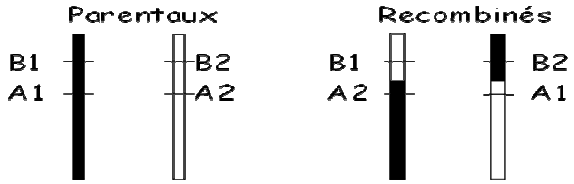
On note A1/A2 les allèles de A et B1/B2 ceux de B. On peut distinguer deux cas selon les combinaisons des allèles sur les deux chromosomes de la paire. Ces deux cas correspondent à deux phases alléliques possibles:

Les allèles A1 et B1 sont sur le même chromosome de la paire, on dit que A1 et B1 sont en « coupling » ; ou encore phasés.

Les allèles A1 et B1 sont chacun sur un chromosome différent, A1 et B1 sont alors en « répulsion ».



Supposons que les allèles A1 et B1 soient en « coupling ». Les gamètes A1B1 et A2B2 produits par cet individu sont dits parentaux ; les gamètes A2B1 et A1B2 sont dits recombinés. Il s'est produit entre les loci A et B un nombre impair de phénomènes de recombinaison. Plus les loci A et B sont proches, moins on observera de gamètes de type A1B2 (ou A2B1).



L'arrangement des allèles à ces deux loci définit un haplotype.

Lorsque les deux loci sont liés génétiquement, la probabilité d'observer les différents gamètes dépend du phénomène de recombinaison mesuré par le paramètre θ, taux de recombinaison entre les loci A et B.

θ est la proportion de gamètes recombinés sur l'ensemble des gamètes transmis par le parent :

$$\theta = \frac{\text{nombre de gamètes recombinés}}{\text{nombre de gamètes transmis}}$$

Association allélique

Le phénomène d'association allélique correspond à une combinaison préférentielle entre les allèles de deux ou plusieurs loci. Considérons deux loci, l'un A1/A2 (avec P_{A2} la fréquence de l'allèle mineur A2) et l'autre B1/B2 (avec P_{B2} la fréquence de l'allèle mineur B2). Si l'on note P_{A2B2} la fréquence de la combinaison A2B2, alors le phénomène d'association allélique signifie que :

$$P_{A2B2} \neq P_{A2} \times P_{B2}$$

Comment se produit l'association allélique ?

Plusieurs événements dans l'histoire génétique des populations sont susceptibles de créer de l'association entre les allèles à différents loci :

– L'évolution démographique de la population :

- La migration : lorsqu'il y a mélange de deux populations et que celles-ci ont des fréquences alléliques différentes pour les deux loci concernés, il y a alors apparition d'association.

- La réduction momentanée de la taille de la population peut également créer de telles situations. C'est un phénomène de dérive génétique.

– La sélection naturelle, favorisant certaines combinaisons alléliques.

– L'apparition d'une mutation sur un chromosome peut entraîner une association. Lorsqu'une mutation apparaît, une association préférentielle va se créer entre la mutation et les allèles aux autres polymorphismes situés sur ce même brin de chromosome. Le maintien de cette association au cours du temps dépend de la distance génétique entre ces loci. La relation est :

$$D_n = (1-\theta)^n D_0$$

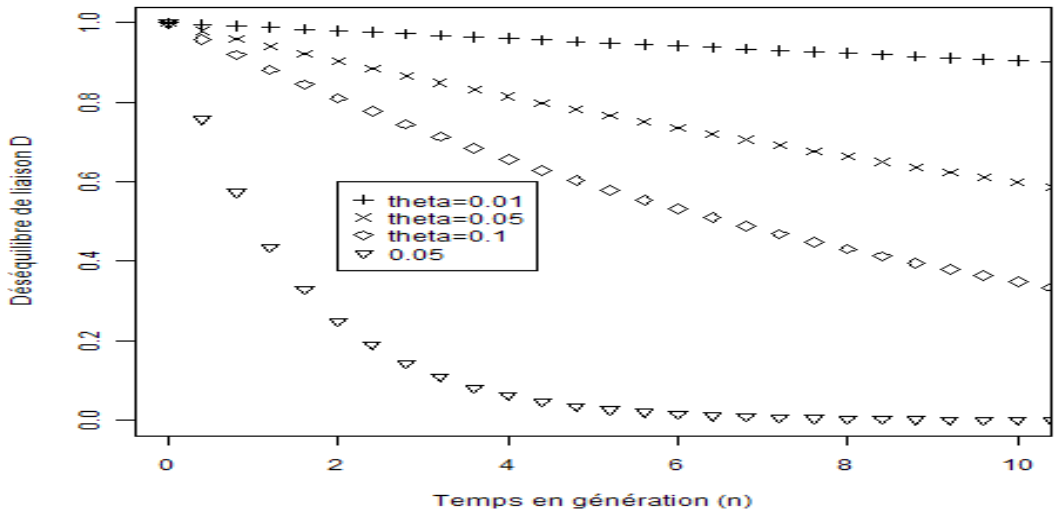
où D_0 est la valeur initiale de l'association et D_n la valeur de cette association à la $n^{ième}$ génération.

C'est cette situation où il y a association et liaison que nous dénotons par déséquilibre de liaison (DL).

Quand les loci sont indépendants ($\theta = 0.5$), l'association initial D_0 décroît de moitié à chaque génération. Au contraire, plus les deux loci sont liés, plus les

recombinaisons sont rares ($\theta \approx 0$) et plus le déséquilibre se maintient au cours du temps (Figure 4).

Figure 4 : Évolution du DL en fonction du nombre de génération (n) pour différentes valeurs du taux de recombinaison θ



Différents paramètres permettent de mesurer l'association entre deux loci.

Le déséquilibre de liaison

Le déséquilibre de liaison D correspond à la différence entre la proportion d'une combinaison donnée d'allèles (haplotype) et celle attendue sous l'hypothèse d'indépendance des allèles:

$$D = P_{A2B2} - P_{A2} \times P_{B2}$$

Ainsi, plus D est élevé et plus les loci sont en déséquilibre de liaison, c'est-à-dire qu'il existe une association non aléatoire entre les allèles des deux loci. Lorsque $D = 0$, il y a équilibre.

Des standardisations de D ont été proposées afin d'avoir des coefficients compris entre -1 et 1. Les plus utilisées sont le coefficient de corrélation r^2 et le D' de (Lewontin 1964):

$$D' = \frac{D}{D_{\max}} \text{ où } D_{\max} = \begin{cases} \min\{P_{A1} \times P_{B2}; P_{A2} \times P_{B1}\} & si\ D > 0 \\ \min\{P_{A1} \times P_{B1}; P_{A2} \times P_{B2}\} & si\ D < 0 \end{cases}$$

Lorsque D' = -1 ou 1, pratiquement, cela se traduit par l'absence d'une des quatre combinaisons possibles, et on parle de **déséquilibre complet**. Cela n'implique pas que deux loci portent exactement la même information. Pour cette raison, on utilise plutôt un indice de corrélation (r^2), lié à la quantité d'information que fournit un locus sur l'autre :

$$r^2 = \frac{D^2}{P_{A1} \times P_{A2} \times P_{B1} \times P_{B2}}$$

On note que $r^2 \leq D'$. Si D'=1 et les deux marqueurs ont des fréquences alléliques identiques alors r^2 = 1. Ont dit que les marqueurs sont complètement corrélés, et dans ce cas ils apportent la même information. En pratique cela se traduit par la présence de seulement deux des quatre combinaisons possibles et par l'égalité des proportions alléliques ($P_{A2} = P_{B2}$). On parle alors de **déséquilibre parfait**.

1.3. Méthodes d'analyses

Pour mettre en évidence l'effet d'un gène, on peut utiliser ou non l'information apportée par un marqueur génétique. Dans le cas où l'on ne dispose pas des génotypes aux marqueurs, on cherche à expliquer la manière dont le phénotype se répartit à l'intérieur des familles en caractérisant l'effet du gène. Ceci se fait par une analyse de ségrégation que nous allons brièvement décrire un peu plus loin. La connaissance des génotypes aux marqueurs est un outil puissant pour caractériser la composante génétique des traits complexes. A partir de l'information génotypique aux marqueurs, on peut regarder s'il existe une corrélation entre le trait et les marqueurs étudiés. S'il existe une relation entre le trait et le marqueur, dans le cas d'un trait quantitatif, on parlera d'un locus de trait quantitatif (**QTL** pour Quantitative Trait Locus). Généralement, on utilise dans un premier temps les études de liaison pour identifier des régions du génome pouvant contenir un gène expliquant une part de la variabilité du trait ou de la maladie, puis dans un deuxième temps les études d'association pour préciser plus finement l'emplacement du gène. Nous allons brièvement décrire le principe de ces méthodes pour identifier des régions chromosomiques en utilisant l'information apportée par les marqueurs aux niveaux de la liaison et de l'association.

Analyses de ségrégation

L'analyse de ségrégation constitue une des premières étapes permettant de caractériser l'effet d'un gène majeur dans un échantillon de familles, sans aucune information apportée par les marqueurs génétiques. L'analyse de ségrégation vise à mettre en évidence l'effet d'un gène (gène majeur) transmis de façon mendélienne parmi l'ensemble des facteurs génétiques et environnementaux causant les concentrations familiales du phénotype étudié.

Ces analyses utilisent comme unité d'échantillonnage un groupe d'individus apparentés : la famille. Le principe général est de déterminer, par des hypothèses statistiques emboîtées, le mode de transmission expliquant le mieux les distributions familiales observées du trait étudié.

Nous allons décrire brièvement le modèle régressif (Bonney 1984) car c'est celui qui a été utilisé dans le cadre de l'étude NEMO.

Nous reprenons les notations définit par le système (1.1). Le gène majeur G à deux allèles A1 (de fréquence p) et A2. Les effets du gène sont caractérisés par 4 paramètres qui sont la fréquence p et les trois moyennes génotypiques ($\mu_{A1A1}, \mu_{A1A2}, \mu_{A2A2}$).

Pour un individu ayant ses parents dans l'échantillon, la probabilité de son génotype est conditionnelle au génotype de son père et de sa mère. Cette probabilité est fonction des trois taux de transmission : $(\tau_{A1A1}, \tau_{A1A2}, \tau_{A2A2})$. Ce sont les probabilités de transmettre l'allèle A1 à un enfant conditionnellement aux génotypes du parent qui est soit A1A1, A1A2 ou A2A2. Sous l'hypothèse de **transmission mendélienne**, ces probabilités sont égales à $(1, 1/2, 0)$. Dans le modèle général de transmission, elles peuvent prendre n'importe quelle valeur entre 0 et 1.

Le trait observé résulte de l'effet de G et des corrélations familiales (noté CF) entre époux (ρ_{FM}), entre le père et l'enfant (ρ_{FO}), entre la mère et l'enfant (ρ_{MO}) et entre les germains (ρ_{SS}). La variance des effets résiduels est notée σ^2. Elle est égale à la variance totale du trait moins la variance de l'effet de G.

Vraisemblance du modèle

L'ensemble des phénotypes Y observés chez les individus d'une même famille F forme une unité d'observation. Les paramètres du modèle générale sont α=(p, μ_{A1A1}, μ_{A1A2}, μ_{A2A2}, ρ_{FM}, ρ_{FO}, ρ_{MO}, ρ_{SS}, τ_{A1A1}, τ_{A1A2}, τ_{A2A2},) et nous donnons ici la vraisemblance de ce modèle pour une famille F de s enfants :

$$L(\alpha \mid F) = P(F \mid \alpha) = P(Y \mid \alpha).$$

$$P(Y|\alpha) = P(Y_f, Y_m, Y_1, Y_2, ...|\alpha)$$
$$= \sum_{k=1}^{3} P(Y_f|G_{f_k}) \times P(G_{f_k}) \times \sum_{j=1}^{3} P(Y_m|G_{m_j}) \times P(G_{m_j}) \times P(Y_m|Y_f)$$
$$\times \prod_{e=1}^{s} \left[\sum_{l=1}^{3} P(Y_e|G_{e_l}) \times P(G_{e_l}|G_{f_k}, G_{m_j}, \alpha) \times P(Y_e|Y_m, Y_f, Y_{e-1}) \right]$$

où k, j et l (1, 2, 3) représente le génotype du père f, de la mère m et des enfants e (1,..,s).

Tests d'hypothèses

Différents sous modèles correspondant à différentes hypothèses sont construits en fixant la valeur de certains paramètres. Les différentes hypothèses sont présentées dans le tableau ci-dessous.

Modèle	p	μ_{A1A1}	μ_{A1A2}	μ_{A2A2}	σ^2	ρ_{FM}	ρ_{FO}	ρ_{MO}	ρ_{SS}	τ_{A1A1}	τ_{A1A2}	τ_{A2A2}
I. Sporadique	-	*	=	=	*	(0)	(0)	(0)	(0)	-	-	-
II. #CF sans $G	-	*	=	=	*	*	*	*	*	-	-	-
III. G s ans CF	*	*	*	*	*	(0)	(0)	(0)	(0)	(1)	(1/2)	(0)
IV. G + CF	*	*	*	*	*	*	*	*	*	(1)	(1/2)	(0)
V. Modèle général de transmission	*	*	*	*	*	*	*	*	*	*	*	*
VI. Absence de transmission	*	*	*	*	*	*	*	*	*	*	=	=

() : Paramètre fixé ; = vaut la valeur du paramètre précédent
- : Paramètre non estimé ; * : Paramètre estimé
: Corrélations familiales ; $: Gène majeur

Les sous-modèles sont testés par le test du rapport des vraisemblances maximales.

La caractérisation de l'effet d'un gène se fait en testant les différentes hypothèses suivantes :

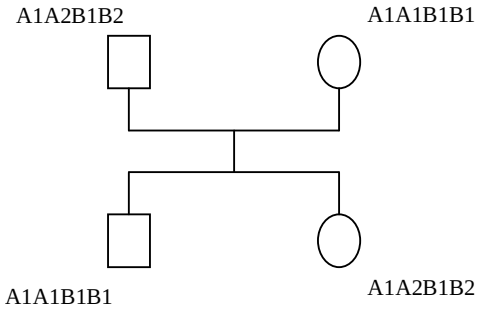
– La présence de corrélations familiales est mise en évidence si le modèle (I) est rejeté par rapport au modèle (II).
– S'il existe des corrélations familiales, la présence d'un gène majeur est recherchée. Celui-ci est mis en évidence si le modèle (II) est rejeté par rapport au modèle (IV).

– Dans le cas où un facteur majeur est détecté, l'existence de corrélations familiales en plus de ce facteur est testée en comparant le modèle (III) au modèle (IV).
– Pour s'assurer que le facteur majeur est bien un gène, transmis de façon mendélienne, deux tests sont effectués :
    - Comparaison de (IV) et (V)
    - Comparaison de (VI) et (V).

Si le modèle de transmission mendélienne (IV) n'est pas rejeté par rapport au modèle général de transmission (V) et si le modèle d'absence de transmission (VI) est rejeté par rapport au modèle général de transmission (V), on pourra conclure à la présence d'un gène majeur.

1.3.1. Méthodes de liaison

Prenons l'exemple simple de la transmission à deux marqueurs. Cela revient à supposer qu'à partir du phénotype on peut en déduire le génotype du locus causal sans ambiguïté. Soit les génotypes à deux marqueurs génétiques, l'un A1/A2 et l'autre B1/B2 pour la famille suivante:



La mère est double homozygote : elle ne produit que des gamètes A1B1 avec probabilité 1. Le père est double hétérozygote, il y a donc 2 phases alléliques possible équiprobable (1/2) :

A1B1/A2B2 $\rightarrow$ A1 et B1 sont en phase

A1B2/A2B1 $\rightarrow$ A1 et B2 sont en phase

On a :

Gamètes	A1B1	A2B2	A1B2	A2B1
A1 et B1 en phase	$\frac{1-\theta}{2}$	$\frac{1-\theta}{2}$	$\frac{\theta}{2}$	$\frac{\theta}{2}$
A1 et B2 en phase	$\frac{\theta}{2}$	$\frac{\theta}{2}$	$\frac{1-\theta}{2}$	$\frac{1-\theta}{2}$

Les enfants sont A1A1B1B1 et A1A2B1B2 avec probabilité $\left(\frac{1-\theta}{2}\right)^2$ si A1 et B1 sont en phase et $\left(\frac{\theta}{2}\right)^2$ si A1 et B2 sont en phase.

Donc la probabilité d'observer cette famille sachant le paramètre θ est :

$$P(F \mid \theta) = \frac{1}{2}\left(\frac{1-\theta}{2}\right)^2 + \frac{1}{2}\left(\frac{\theta}{2}\right)^2 = \frac{1}{8}\left(1 - 2\theta + 2\theta^2\right)$$

On voit que si θ = ½, alors, quelque soit la phase, tous les types de gamètes ont la même probabilité ¼. On dit que les 2 loci sont **indépendants**, c'est-à-dire qu'il n'y a pas liaison.

Vraisemblance de la liaison génétique

L'ensemble des phénotypes Y et des génotypes au marqueur M observés chez les individus d'une même famille F, à n enfants, forme une unité d'observation.

La probabilité de l'observation sachant la liaison génétique (par exemple $\theta = \theta_1$ ou $\theta = 1/2$) entre les deux loci et les paramètres du modèle (α) est :

$P(F \mid \theta) = P(Y, M \mid \theta, \alpha)$. La vraisemblance de θ est : $L(\theta \mid F) = P(F \mid \theta)$

$$\begin{aligned} P(Y, M \mid \theta, \alpha) &= P\left(Y_f, M_f, Y_m, M_m, Y_1, M_1, Y_2, M_2 \ldots \mid \theta, \alpha\right) \\ &= P\left(Y_f, M_f\right) P\left(Y_m, M_m\right) P\left(Y_1, M_1, Y_2, M_2 \ldots \mid Y_f, M_f, Y_m, M_m, \theta, \alpha\right) \\ &= \sum_{k=1}^{3} P\left(Y_f \mid G_{f_k}, \alpha\right) P\left(G_{f_k}\right) P\left(M_f\right) \times \sum_{j=1}^{3} P\left(Y_m \mid G_{m_j}, \alpha\right) P\left(G_{m_j}\right) P\left(M_m\right) \\ &\quad \times \prod_{e=1}^{n} \left[\sum_{l=1}^{3} P\left(Y_e \mid G_{e_l}, \alpha\right) P\left(G_{e_l}, M_e \mid G_{f_k}, M_f, G_{m_j}, M_{m,} \theta\right) \right] \end{aligned} \qquad \textbf{(1.6)}$$

Pour un phénotype quantitatif, les paramètres α sont la fréquence de l'allèle p et les 3 moyennes génotypiques $\left(\mu_{G_1}, \mu_{G_2}, \mu_{G_3}\right)$ et la probabilité du phénotype Y sachant le génotype G_k est la densité : $f\left(\mu_{G_k}, \sigma^2\right)$. Pour un phénotype qualitatif les paramètres sont la fréquence de l'allèle p et les 3 pénétrances, et $P\left(Y \mid G_k\right), k = 1,2,3$ est le vecteur des pénétrances.

Il existe deux principales approches pour la recherche de liaison de phénotypes quantitatifs ou qualitatifs, qui sont les analyses de liaison dites paramétriques et les analyses de liaison dites non-paramétriques.

Méthode du LOD score

L'analyse de liaison paramétrique modélise exactement la co-transmission du phénotype et des marqueurs dans les familles. Pour cela, les paramètres du modèle génétique sont supposés connus. Une des méthodes les plus communément utilisée pour tester la liaison génétique entre deux ou plusieurs loci est la méthode des LOD scores (Morton 1955).

L'hypothèse nulle H_0, d'une transmission indépendante des deux loci est définie par $\theta = 1/2$.

On considère H_1, l'hypothèse alternative suivante : $\theta = \theta_1 < 1/2$, c'est-à-dire, la transmission des deux loci n'est pas indépendante et le taux de recombinaison entre les deux loci est θ_1.

Soit $L_i(\theta = 1/2)$ et $L_i(\theta = \theta_1)$ la vraisemblance des données sous l'hypothèse H_0 et H_1 respectivement pour la famille i.

La statistique du **LOD score** en θ_1 de cette famille $Z_i\left(\theta_1\right)$, est donné par le logarithme en base 10 du rapport de ces deux vraisemblances et vaut:

$$Z_i\left(\theta_1\right) = \log_{10}\left(\frac{L_i(\theta = \theta_1)}{L_i(\theta = 1/2)}\right)$$

Le LOD score en θ_1, $Z\left(\theta_1\right)$ d'un échantillon de k familles est la somme des LOD scores de chaque famille i :

$$Z\left(\theta_1\right) = \sum_i Z_i\left(\theta_1\right)$$

Critère de décision :

La procédure de test est de type séquentiel. On accumule de l'information (des familles), jusqu'au moment ou il sera possible de trancher entre les hypothèses H_0 et H_1. La valeur du LOD score indique les probabilités relatives d'observer l'échantillon sous H_1 et sous H_0. Ainsi, un LOD score de 3 signifie que la probabilité d'observer l'échantillon est 1 000 fois plus grande sous H_1 que sous H_0.

Les seuils de décision sont fixés à -2 et +3, c'est-à-dire que si :

- si $Z(\theta_1) > 3$ on rejette H_0 et on conclut à la liaison pour $\theta = \theta_1$
- si $Z(\theta_1) \leq -2$ on exclut la liaison génétique à θ_1
- si $-2 < Z(\theta_1) \leq 3$ on ne peut trancher entre ces deux hypothèses, il faut continuer d'accumuler de l'information.

Remarque :

On peut montrer que les familles sont informatives, (c'est-à-dire que la vraisemblance est une fonction non constante de θ) pour la liaison génétique lorsque l'un des parents est double hétérozygote et lorsqu'elles contiennent au moins deux enfants.

En pratique, les études ne rapportent pas la valeur du LOD score étant donné une valeur de θ, mais le maximum du LOD score : $Z(\theta')$, où θ' est la valeur qui maximise la vraisemblance du modèle de liaison. Dans ces cas, la statistique $Z(\theta')$ suit sous l'hypothèse nulle, un mélange de χ^2 à 0 et 1 degré de liberté.

Des erreurs sur la valeur des paramètres génétiques augmentent le biais sur l'estimation de θ et diminuent la puissance de détecter une liaison (Clerget-Darpoux, Bonaiti-Pellie et al. 1986). De plus, nous ne savons pas spécifier a priori un tel modèle dans le cas des maladies multifactorielles. C'est pour cela que des méthodes (méthodes dites « model-free ») ne faisant aucune hypothèse sur le modèle génétique ont été développées.

Méthodes non-paramétriques, « model-free »

Le principe général de ces méthodes de liaison est de rechercher s'il existe une corrélation entre la ressemblance au trait et la ressemblance au marqueur entre apparentés. Dans ces approches, le test de liaison ne dépend, généralement que d'un seul paramètre. La similarité des allèles au marqueur est représentée par la proportion du nombre d'allèles partagés par descendance. On représente ce nombre par le statut IBD (Identity by Descent). Dans le cas de germains, le nombre d'allèles IBD est 0, 1 ou 2 (Tableau 2).

Parmi les méthodes de liaison « model-free » qui ont été développées, une des plus connues est celle proposée par Haseman & Elston (Haseman and Elston 1972), basée sur des paires de germains (frères/sœurs). Nous la décrivons ici brièvement. Elle a été « revisitée » afin de tenir compte de la covariance phénotypique entre les germains (Elston, Buxbaum et al. 2000). D'autres méthodes permettent d'utiliser tous les types de familles, comme les méthodes basées sur la décomposition de la variance (Amos 1994). Cette dernière approche est décrite plus en détail dans le chapitre 2.1.1.2

Trait quantitatif

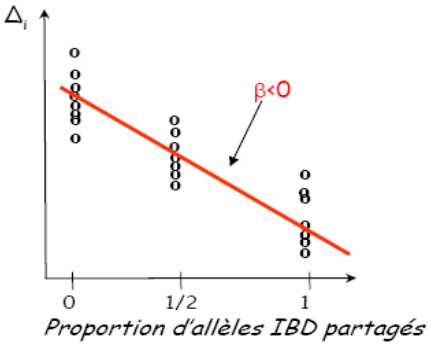
Le principe de la méthode de Haseman & Elston est de rechercher une relation entre la différence des valeurs phénotypiques et la proportion d'allèles IBD au marqueur étudié, chez des paires de germains. Le modèle est une simple régression linéaire. Brièvement, soit y_1 , y_2 et π_i , les valeurs observées au trait et la proportion d'allèles IBD chez deux frères de la famille i. Le modèle de régression est :

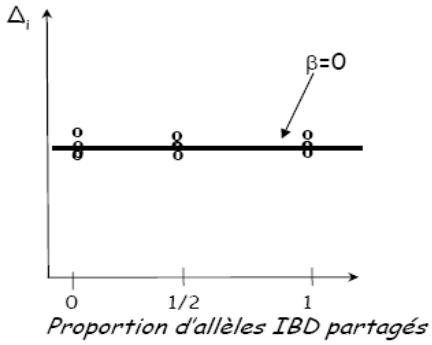
$$\Delta_i = (y_1 - y_2)^2 = \mu + \beta\ \pi_i + e$$

où e est un effet aléatoire.

Sous l'hypothèse nulle de non liaison, il n'y a pas de relation entre Δ_i et π_i et donc $\beta=0$. On considère l'hypothèse alternative de liaison, $\beta < 0$, c'est-à-dire Δ_i diminue lorsque la proportion d'allèles IBD augmente.
Le test de liaison suit sous l'hypothèse nulle un χ^2 à 1 degré de liberté.

Les situations attendues sous l'hypothèse alternative et sous l'hypothèse nulle sont montrées respectivement dans les figures ci-dessous.





Conditionnellement au marqueur, Haseman et Elston ont montré que l'estimateur de β est $\hat{\beta} = -2(1-2\theta)^2\sigma_A^2$ où θ est le taux de recombinaison entre le QTL (Quantitative Trait Locus) et le marqueur et σ_A^2 est la variance des effets additifs du QTL. A partir de l'estimation de β on ne peut distinguer les effets dus à la liaison entre le QTL et le marqueur et les effets du QTL. Si la liaison est complète ($\theta = 0$) : $\hat{\beta} = -2\sigma_A^2$.

Plus généralement, pour toutes les méthodes de liaison « non-paramétriques », le paramètre de liaison est une fonction du taux de recombinaison θ entre le QTL et le marqueur et de la variance des effets du QTL σ_A^2.

Trait qualitatif

La méthode des paires de germains atteints (« affected sib-pairs ») consiste à comparer la distribution d'allèles IBD observée à celle attendue a priori chez des germains quelque soit leur phénotype. La distribution IBD attendue est 1/4,1/2,1/4 pour k=0,1 ou 2 allèles IBD. En revanche, si le marqueur est génétiquement lié au QTL, on s'attend à un déficit de paires de germains partageant 0 allèles IBD.

Le test de liaison est un test de χ^2 de conformité.

D'autres méthodes sont plus générales et permettent de tester la liaison lorsque l'IBD est ambiguë (à cause d'un manque d'informativité du marqueur et/ou parce que les parents ne sont pas tous génotypés) comme la méthode du Maximum Likelihood Score de Risch (1990), ou celle de (Kong and Cox 1997).

Remarque :

Il existe une autre manière de choisir les paires informatives qui est de prendre des paires **discordantes** où un des germains est malade tandis que l'autre est sain. Puisque les paires sont choisies parce que les individus sont différents du point de vue du phénotype, on s'attendra à trouver un déficit de paires où l'IBD=2 et une augmentation de paires où l'IBD=0. On peut aussi combiner les échantillons de paires discordantes et concordantes. Sous l'hypothèse nulle, les proportions d'allèles IBD sont similaires dans les deux échantillons. Pour les maladies à étiologie complexe la plupart des études de liaison se sont conduites dans des échantillons d'apparentés atteints, par des approches de type « affected sib-pairs » ou « affected-only ».

1.3.2. Méthodes d'association

L'analyse d'association permet d'identifier des variants génétiques impliqués dans la variabilité du trait à partir d'observations de données en population ou de données de sujets apparentés. Le principe consiste à regarder s'il existe une association préférentielle entre les allèles du variant génétique causal et les allèles de marqueurs génétiques.

Trait quantitatif - Données de sujets non apparentés

L'étude de l'association entre un marqueur et un trait quantitatif peut se faire par différentes approches classiques, comme la régression linéaire ou l'ANOVA. Le principe général est de stratifier l'échantillon selon les valeurs (génotypes) au marqueur et de tester si la moyenne du phénotype varie entre ces groupes. La relation entre un phénotype et un marqueur bi-allélique s'écrit :

$$y_j = \mu + \beta \times g_j + e$$

où, y_j est la valeur du phénotype au j^{ème} individu ; g_j est le score génotypique de l'individu j (par exemple, égal au nombre d'allèle mineur au marqueur moins 1 sous un modèle additif), β est le coefficient de régression de l'effet du marqueur M, et e est un effet aléatoire. Sous l'hypothèse nulle, $\beta = 0$, il n'y a pas d'association entre le marqueur et le trait quantitatif. On considère l'hypothèse alternative suivante : $\beta \neq 0$. Le test du rapport des vraisemblances maximales suit sous l'hypothèse nulle un χ^2 à 1 degré de liberté.

Supposons qu'une partie de la variance de Y est expliquée par un gène G. Si G est M, le marqueur étudié ou si G est en déséquilibre de liaison avec M, alors les moyennes de Y varient par groupe génotypique de M. Dans le premier cas, on parle ***d'association directe***, dans le second cas, on dit que ***l'association est indirecte***. Plus la relation de déséquilibre de liaison entre G et M est forte plus la relation phénotype-marqueur est apparente.

Trait quantitatif - Données de sujets apparentés : Ces méthodes sont présentées plus en détails dans le paragraphe 4.1.

Trait qualitatif - Données de sujets non apparentés

L'association entre une maladie est un marqueur peut être testée par des approches classiques, comme la régression logistique, le χ^2 d'homogénéité, etc.

Soit un échantillon de N cas et N témoins. On peut comparer la distribution des allèles, ou des génotypes, chez les cas et chez les témoins par un χ^2 d'homogénéité.

Test de la distribution allélique :

	A1	A2	
cas	n_{11}	n_{12}	$2N$
témoins	n_{21}	n_{22}	$2N$
	n_{A1}	n_{A2}	$4N$

La statistique du χ^2 est :

$$\chi^2 = \sum_{i=1}^{2} \frac{\left(n_{1i} - \frac{n_{Ai} * 2N}{4N} \right)^2}{\frac{n_{Ai} * 2N}{4N}} + \sum_{i=1}^{2} \frac{\left(n_{2i} - \frac{n_{Ai} * 2N}{4N} \right)^2}{\frac{n_{Ai} * 2N}{4N}}$$

La statistique suit, sous l'hypothèse nulle d'absence d'association, une loi de χ^2 à 1 degré de liberté.

Si la distribution des allèles est significativement différente entre les cas et les témoins, il y a association entre le marqueur et la maladie.

La force de la relation entre la maladie et le marqueur est définit par le rapport des côtes noté OR (Odds Ratio) :

$$OR = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}}$$

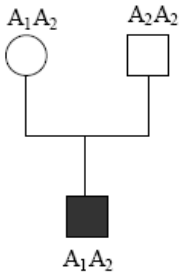
Test de la distribution génotypique :

	A1A1	A1A2	A2A2	
cas	n_{11}	n_{12}	n_{13}	N
témoins	n_{21}	n_{22}	n_{23}	N
	n_{A1A1}	n_{A1A2}	n_{A2A2}	$2N$

On peut calculer de la même manière la valeur de la statistique du χ^2, qui suit sous H_0 un χ^2 à 2 degrés de liberté.

Trait qualitatif - Données de sujets apparentés

Ici nous présentons brièvement le test du TDT (Transmission Disequilibrium Test) ; (Spielman, McGinnis et al. 1993). Le principe général consiste à considérer les allèles transmis par les parents à l'enfant malade comme étant des allèles « cas » et les allèles non transmis comme étant des allèles « témoins ». Ce test utilise des trios constitués des deux parents et d'un enfant atteint et s'appuie sur l'observation des allèles transmis par les parents hétérozygotes à l'enfant atteint.



La table de contingence pour n trios se présente ainsi :

		Allèles non transmis	
		A1	A2
Allèles transmis	A1	-	*b*
	A2	*c*	-

où,

- *b* est le nombre de fois où un parent hétérozygote A1A2 a transmis l'allèle A1 à son enfant atteint (et correspond aussi au nombre de fois où l'allèle A2 n'a pas été transmis).
- *c* est le nombre de fois où un parent hétérozygote A1A2 a transmis l'allèle A2 à son enfant atteint (et n'a donc pas transmis A1).

On calcule alors la statistique suivante basée sur une statistique du χ^2 pour des séries appariées (McNemar 1947) :

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

qui suit sous l'hypothèse nulle une loi de χ^2 à 1 degré de liberté.

L'hypothèse nulle est composite : DL=0 (absence d'association) ou θ=1/2 (absence de liaison). Le rejet de l'hypothèse nulle permet donc de conclure à l'existence à la fois à l'association et à la liaison entre le marqueur et le phénotype étudié.

Performances relatives de ces méthodes

Historiquement, les analyses de ségrégation et les méthodes de liaison paramétriques ont eu du succès pour mettre en évidence des effets de gènes majeurs et pour localiser les régions du génome de gènes impliqués dans des maladies monogéniques à transmission mendélienne. Cependant, dans le cas de traits complexes, le modèle de transmission résulte probablement de l'effet de plusieurs gènes interagissant avec des facteurs environnementaux, il n'existe donc plus de relation simple entre les génotypes et les phénotypes.

L'avantage des modèles de liaison « model-free » est qu'ils ne nécessitent pas de spécifier la correspondance phénotype – génotype. Ce type d'approche permet de mettre en évidence une corrélation entre la ressemblance au trait et la ressemblance au marqueur entre apparentés. Cependant, ces méthodes ne permettent pas de préciser les effets génétiques ni de localiser le gène vis-à-vis du marqueur. Malgré un certain succès des études de liaison « model-free » comme par exemple pour le diabète de Type 1 (Davies, Kawaguchi et al. 1994) ou la maladie de Crohn (Hugot, Chamaillard et al. 2001), la plupart des résultats dans la recherche de gènes impliqués dans la variabilité de phénotypes complexes sont peu consistants. La puissance de détection d'une liaison dépend du modèle génétique sous-tendant la maladie. Ainsi, les analyses de liaison ont peu de puissance pour détecter des variants fréquents. En effet, dans ce cas, les apparentés atteints peuvent être porteurs du variant fonctionnel sans l'avoir reçu du même ancêtre (sans être IBD). Pour des phénotypes complexes, il est plus réaliste de penser que les effets génétiques sont faibles et que les variants causaux sont relativement fréquents (hypothèse « common disease – common variant »)(Risch 2000; Cardon and Bell 2001; Hirschhorn and Daly 2005). La puissance d'une analyse de liaison est également réduite en présence d'hétérogénéité génétique non allélique, c'est-à-dire si plusieurs variants fonctionnels, localisés dans des régions différentes du génome, sont impliqués dans la variation du trait. Pour l'ensemble des familles de l'échantillon, ce ne sont alors pas forcément les mêmes variants causaux qui se transmettent à travers les générations entre les différentes familles.

Suite aux travaux de Risch et Merikangas (Risch and Merikangas 1996), les études d'association ont connues un regain d'intérêt. Leur étude a montré le gain apporté par les études d'association relativement aux études de liaison en termes de taille d'échantillon, en particulier pour des gènes ayant un effet modeste. Ils ont aussi suggéré que des études

d'association sur l'ensemble du génome étaient possibles. Cependant, comme nous l'avons souligné, ce gain de puissance est fortement dépendant de la force du déséquilibre de liaison (DL) entre le marqueur et le gène causal et de leurs fréquences alléliques respectives.

Le déséquilibre d'association ne réfère pas toujours au déséquilibre de liaison car il peut apparaître suite à des évènements tels que les mélanges de population. Les tests d'association dans des données de sujets non apparentés sont plus facilement réalisables. Cependant, l'existence d'une association trait – marqueur peut non seulement être due à l'existence d'un déséquilibre de liaison entre le marqueur et le locus de susceptibilité situé à proximité du marqueur, mais aussi à des mécanismes de stratification de population. Si les groupes constitués ne sont pas homogènes, une stratification de l'échantillon pourra générer une fausse association significative entre la variabilité du trait et le marqueur. Ce phénomène de stratification est plus détaillé au chapitre 3.

Notons que le biais résultant d'un mauvais appariement entre les cas et les témoins ou d'un problème de mélange de population augmente avec la taille de l'échantillon. Ainsi, dans le cadre des études d'association par criblage du génome (GWAS) les biais peuvent ne pas être négligeables, puisque ces études sont généralement conduites dans des échantillons de grandes tailles.

L'analyse d'association en population se heurte à des problèmes de robustesse. Comme nous le verrons (chapitre 3.1), des approches ont été développées dans le but de corriger ou contrôler les biais potentiels de ces analyses.

Une autre approche est de réaliser l'étude d'association sur des données familiales. Un des tests les plus connus est le TDT (Spielman, McGinnis et al. 1993). L'intérêt principal du TDT est qu'il exploite les deux informations, familiale et populationnelle, en testant simultanément les corrélations induites par la liaison génétique et l'association. Ce test est robuste à la stratification de la population car le test est conditionné par le génotype des parents. Ce test peut être appliqué à des familles comportant plusieurs enfants atteints en le répétant pour chaque enfant atteint, mais il faut alors corriger la signification des tests en raison de la non-indépendance des paires au sein d'une même fratrie. Ces tests d'association sur des trios supposent que les deux parents sont génotypés, dans le cas contraire il est préférable d'exclure ces familles pour éviter des biais. Dans le cas de maladie à début d'âge tardif, ce design d'échantillonnage est alors difficile à réaliser car les parents ne peuvent être génotypés. Pour ces phénotypes, il est plus facile de recueillir des données de sujets non apparentés. Un autre test (*Sib TDT*) a été proposé qui utilise comme témoins les germains non atteints des enfants atteints (Spielman and Ewens 1998). Quand les familles incluent à la fois des parents génotypés et non génotypés, il est possible de combiner ces deux derniers tests.

Pendant longtemps, l'hypothèse implicite utilisée dans l'analyse génétique de traits complexes, communs dans la population, était que les variants impliqués dans ces phénotypes avaient des effets modestes sur la variation des traits, mais étaient

relativement fréquent dans la population générale (hypothèse « common variants-common disease »). Une hypothèse alternative plus récente, pour des traits complexes, est que de nombreux variants sont impliqués dans la variation des traits complexes, chacun avec des effets relativement forts, mais très peu d'individus sont porteurs du variant (hypothèse « common disease-rare variants »). Dans ces cas, les études d'association ont très peu de puissance pour détecter l'association avec des variants rares c'est à dire classiquement pour des fréquences inférieures à 1% (Frazer, Murray et al. 2009; Schork, Murray et al. 2009; Cirulli and Goldstein 2010).

1.4. Analyse génétique de traits corrélés

Les causes de covariance des caractères peuvent être génétiques et/ou environnementales. La principale source de co-variation génétique est la ***pléiotropie***, c'est à dire, lorsqu'un QTL (G) exerce une action simultanément sur les deux caractères (Figure 5.a).

La covariance génotypique entre deux QTLs, notés G et G', est la moyenne des produits des valeurs génotypiques :

$$\mathrm{cov}(G,G') = q^2 \times \mu_{G_3}\mu'_{G_3} + 2pq \times \mu_{G_2}\mu'_{G_2} + p^2 \times \mu_{G_1}\mu'_{G_1}$$

Le tableau 3 montre comment le produit des valeurs génotypiques de deux caractères se décompose en quatre produits :

$$\begin{aligned}\mathrm{cov}(G,G') &= \mathrm{cov}(\alpha_A + \alpha_D, \alpha_{A'} + \alpha_{D'}) \\ &= \mathrm{cov}(\alpha_A,\alpha_{A'}) + \mathrm{cov}(\alpha_D,\alpha_{D'}) + \mathrm{cov}(\alpha_A,\alpha_{D'}) + \mathrm{cov}(\alpha_{A'},\alpha_D)\end{aligned}$$

où α_A, $\alpha_{A'}$ sont les effets additifs de G et G' et α_D, $\alpha_{D'}$ sont les effets de dominance de G et G'.

On montre que $\mathrm{cov}(\alpha_A,\alpha_{D'}) = 0$ et de même $\mathrm{cov}(\alpha_{A'},\alpha_D) = 0$

Ainsi $\mathrm{cov}(G,G') = \mathrm{cov}(\alpha_A,\alpha_{A'}) + \mathrm{cov}(\alpha_D,\alpha_{D'})$

La covariance génotypique est la somme de la covariance des valeurs génétiques, ou **covariance génétique additive**, et de la covariance des résidus de dominance, ou **covariance de dominance.**

On montre que :

$$\mathrm{cov}(\alpha_A, \alpha_{A'}) = 2pq(\alpha_1 - \alpha_2)(\alpha'_1 - \alpha'_2)$$
$$\mathrm{cov}(\alpha_D, \alpha_{D'}) = 4p^2q^2dd'$$

où α_1, α_2 sont les effets moyens des allèles (paragraphe 1.1)

On voit que ces covariances sont les moyennes géométriques des 2 variances correspondantes, puisque :

$$\mathrm{cov}(\alpha_A, \alpha_{A'}) = \sqrt{\sigma^2_A} \times \sqrt{\sigma^2_{A'}}$$
$$\mathrm{cov}(\alpha_D, \alpha_{D'}) = \sqrt{\sigma^2_D} \times \sqrt{\sigma^2_{D'}}$$

La covariance génétique est souvent exprimée sous la forme d'un coefficient de corrélation. La corrélation entre les valeurs génétiques additives de deux caractères : $\rho_A = \dfrac{\mathrm{cov}(\alpha_A, \alpha_{A'})}{\sqrt{\sigma^2_A} \times \sqrt{\sigma^2_{A'}}}$, est appelée la corrélation génétique ρ_G entre les deux caractères. Dans le cas d'un locus à effets pléiotropiques, ρ_G = ±1.

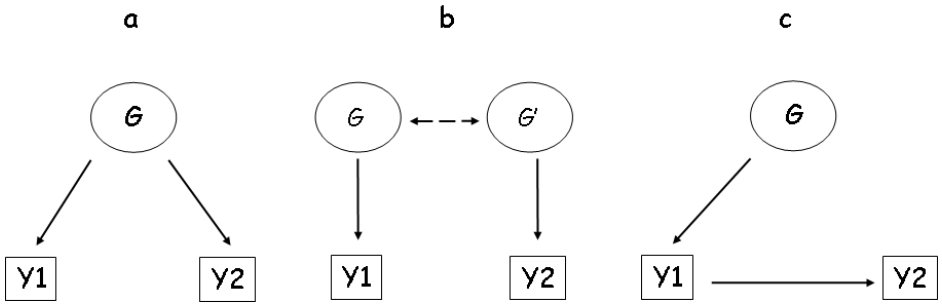
Tableau 3 : Décomposition du produit de deux valeurs génotypiques

G_i		**A1A1**	**A1A2**	**A2A2**
P_{G_i}		p^2	$2pq$	q^2
Valeurs génotypiques (GG')		$\mu_{G_1}\mu'_{G_1}$	$\mu_{G_2}\mu'_{G_2}$	$\mu_{G_3}\mu'_{G_3}$
Valeurs génétiques	additive AA'	$4\alpha_1\alpha'_1$	$(\alpha_1+\alpha_2)(\alpha'_1+\alpha'_2)$	$4\alpha_2\alpha'_2$
	dominance DD'	$4q^4dd'$	$4p^2q^2dd'$	$4p^4dd'$
Effets croisés	AD'	$-4q^2\alpha_1d'$	$2pq(\alpha_1+\alpha_2)d'$	$-4p^2\alpha_2d'$
	A'D	$-4q^2\alpha'_1d$	$2pq(\alpha'_1+\alpha'_2)d$	$-4p^2\alpha'_2d$

La force de la corrélation génétique reflète l'influence du (des) gène(s) commun(s) sur la variabilité des caractères. Certains gènes peuvent contribuer de façon similaire sur chacun des caractères, par exemple l'allèle A1 augmente les valeurs des traits. Au contraire, d'autres gènes peuvent contribuer de façon opposée, par exemple, A1 augmente la valeur du premier trait et diminue celle du deuxième trait. Dans la première situation, la corrélation génétique ρ_G est positive alors qu'elle est négative dans la seconde situation. La pléiotropie ne génère donc pas nécessairement des corrélations génétiques détectables.

Il existe d'autres situations, hors la simple pléiotropie, où des corrélations génétiques non nulles peuvent être générées. La première situation est celle de deux QTLs influençant l'un le premier caractère, l'autre le second, et qui ne sont pas indépendants à cause de la liaison. Cette situation correspond à la ***co-incidence***. Les deux QTLs sont génétiquement liés, situés dans la même région chromosomique, et très proches l'un de l'autre. Les génotypes à ces deux QTLs ne sont donc pas indépendants au sein des familles (Figure 5.b). Une autre situation est celle où G n'explique directement qu'un seul des 2 phénotypes, mais ceux-ci ont des relations de causalité entre-eux (Figure 5.c).

Figure 5 : Sources de corrélation génétique pour des traits Y1 et Y2.



Décomposition de la covariance entre traits

Soit $\rho_P = \dfrac{\mathrm{cov}(y_1, y_2)}{\sqrt{\sigma^2_{y1}\sigma^2_{y2}}}$ la corrélation phénotypique entre Y1 et Y2 et σ^2_{yj} la variance de Y_j.

Sous le modèle défini par l'équation (1.4), la covariance phénotypique entre les traits Y1 et Y2 est :

$$\begin{aligned}\mathrm{cov}(y_1, y_2) &= \mathrm{cov}(y_1, y_2 \mid G) + \mathrm{cov}(y_1, y_2 \mid C) + \mathrm{cov}(e_1, e_2) \\ &= \rho_G \sqrt{\sigma^2_{G1}\sigma^2_{G2}} + \rho_C \sqrt{\sigma^2_{C1}\sigma^2_{C2}} + \rho \sqrt{\sigma^2_1 \sigma^2_2}\end{aligned}$$

Avec ρ_G, ρ_C et ρ les corrélations des effets de G, de C et de E sur les traits Y1 et Y2 et $\sigma^2_{Gj}, \sigma^2_{Cj}$ et σ^2_j, les variances respectives des effets du gène G, de la polygénique C et de la résiduelle E sur le trait Y_j qui peut elle-même s'écrire sous la forme :

$$\begin{aligned}\sigma^2_j &= E\left(y^2_j\right) - E\left(y^2_j \mid G\right) - E\left(y^2_j \mid C\right) \\ &= \sigma^2_{yj} - \sigma^2_{Gj} - \sigma^2_{Cj} = \sigma^2_{yj}\left(1 - h^2_{Gj} - h^2_{Cj}\right)\end{aligned}$$

où $h^2_{Gj} = \frac{\sigma^2_{Gj}}{\sigma^2_{yj}}$ et $h^2_{Cj} = \frac{\sigma^2_{Cj}}{\sigma^2_{yj}}$ sont les parts de variance phénotypique du trait Y_j expliquées par les composantes G (h^2_{Gj}) et C (h^2_{Cj}).

Finalement, la corrélation phénotypique peut se réécrire sous la forme :

$$\rho_P = \rho_G \sqrt{h^2_{G1} h^2_{G2}} + \rho_C \sqrt{h^2_{C1} h^2_{C2}} + \rho \sqrt{\left(1 - h^2_{G1} - h^2_{C1}\right)\left(1 - h^2_{G2} - h^2_{C2}\right)} \quad (1.7)$$

Lorsque le gène G n'est pas connu $\rho_P = \rho_C \sqrt{h^2_{C1} h^2_{C2}} + \rho \sqrt{\left(1 - h^2_{C1}\right)\left(1 - h^2_{C2}\right)}$.

Cette équation est souvent utilisée pour estimer l'impact de facteurs familiaux communs sur la covariation des traits pouvant suggérer l'existence de gènes à effets pléiotropiques.

Plusieurs méthodes d'analyses génétiques ont été développées pour l'étude de traits corrélés. Les méthodes bivariées ont l'avantage de tenir compte des corrélations phénotypiques entre les traits. C'est ce type de méthodes que nous avons appliquées dans ce travail de thèse pour l'étude génétique de traits quantitatifs. Ces méthodes d'analyses bivariées de liaison et d'association sont détaillées dans les paragraphes 2.1.1 et 3.2.1.

1.5. Génétique de la densité osseuse : généralités et revue de la littérature

L'ostéoporose se définit comme une fragilisation de la matrice osseuse. L'ostéoporose ne s'accompagne habituellement d'aucun signe, mais sa présence augmente le risque de fracture. Les sièges classiques des fractures d'origine ostéoporotique sont :

- Les fractures vertébrales (ou tassements vertébraux) dont les symptômes sont le plus souvent négligés par les patients.
- Les fractures de l'extrémité inférieure de l'avant-bras.
- Les fractures de l'extrémité supérieure du fémur, provoquant le plus de complications. Un peu moins d'un quart des fractures surviennent chez les hommes (International Osteoporosis Fundation).

Le risque de fracture est inversement corrélé à la densité minérale osseuse (DMO).

Biologie de l'os : L'os est une structure en perpétuel renouvellement. Le tissu osseux se présente sous deux aspects bien distincts représentés dans la figure 25 en annexe :

(i) Une partie périphérique : l'os cortical ou os compact est une partie osseuse dense, dure et résistante.

(ii) Une partie centrale : l'os trabéculaire ou os spongieux est la partie interne de l'os ; il est constitué de petites travées. Il est le siège de la moelle osseuse qui fabrique les globules.

On admet que les vertèbres sont formées de 50 % d'os trabéculaire et de 50 % d'os cortical. Le col du fémur, quant à lui, est constitué de 30 % d'os trabéculaire et de 70 % d'os cortical.

- La matrice osseuse se renouvelle perpétuellement. Ce phénomène de remodelage a lieu par petites zones et comprend une phase de résorption osseuse d'une durée d'une à deux semaines, suivie d'une phase de formation osseuse d'une durée d'environ trois mois. L'os trabéculaire, bien que moins abondant quantitativement, se renouvelle environ cinq fois plus rapidement que l'os cortical. Pour cette raison, l'ostéoporose se manifeste cliniquement dans des sites où il existe une proportion relativement importante d'os trabéculaire, c'est-à-dire principalement au niveau du rachis, de la hanche et de l'avant-bras où, à terme, elle conduit à la fracture (Cohen-Solal and de Vernejoul 2004)

- Les cellules osseuses, ostéoclastes et ostéoblastes, sont responsables respectivement de la dégradation et de la synthèse de la matrice osseuse. Les ostéoblastes prolifèrent le long des travées osseuses et se différencient progressivement en exprimant des gènes du collagène de type I et les gènes de différentes protéines non collagéniques. Ces protéines ne sont pas forcément spécifiques du tissu osseux.

- Les capacités prolifératives de ces cellules diminuent avec l'âge, induisant un déficit de la formation osseuse à l'origine de l'ostéoporose. Une matrice osseuse, qui n'est pas suffisamment renouvelée, peut avoir une mauvaise résistance biomécanique. A l'inverse, un remaniement osseux excessif peut être cause de fragilisation osseuse, la maladie de Paget en est un exemple.

En conclusion, si la perte osseuse demeure un processus normal du vieillissement, l'ostéoporose résulte, quant à elle, d'une balance osseuse exagérément négative. De plus, chez la femme, l'épaisseur de l'os cortical diminue de façon significative à partir de la ménopause et se poursuit tout au long du vieillissement en raison d'une carence en œstrogènes après la ménopause. Cette perte est nettement moins importante chez l'homme.

Densité minérale osseuse : L'évolution de la densité minérale osseuse en fonction de l'âge et du sexe est représentée dans la figure 26 en annexe. La densité osseuse augmente pour atteindre son pic vers l'adolescence. Ensuite, cette densité diminue, mais chez la femme, comme pour l'épaisseur de l'os, cette diminution est plus rapide après la ménopause.

- **Mesure de la densité minérale osseuse** : Le diagnostic de l'ostéoporose repose sur la mesure de la densité minérale osseuse (DMO) par ostéodensitométrie. Cet examen consiste à exploiter le contenu minéral osseux (CMO) exprimé en grammes de calcium hydroxyapatite afin d'en calculer la densité minérale osseuse (DMO) exprimée en grammes de calcium hydroxyapatite par cm² (g/cm^2). La méthode de référence pour apprécier la qualité de l'os est aujourd'hui l'absorptiométrie biphotonique à rayons X (DXA). Elle consiste à émettre des photons en direction d'un os. Plus l'os est dense, moins nombreux sont les photons qui atteignent le détecteur. On mesure alors ce qui reste du rayonnement après sa traversée de l'os, ce qui renseigne sur sa densité. La mesure densitométrique peut être réalisée au rachis lombaire, à la hanche, à l'avant-bras (le plus souvent au radius) ou au corps entier.

  Les sites habituels de la mesure sont le rachis lombaire et la hanche, qui sont les sièges fréquents de fractures ostéoporotiques. La figure 27 en annexe montre les zones de mesures de la DMO pour le rachis lombaire (photo de gauche) et la hanche (photo de droite). La mesure au rachis lombaire est le résultat global de la mesure aux lombaires L1 à L4 ou L2 à L4. La mesure au rachis lombaire peut être surestimée par l'arthrose, fréquente après 70 ans. La DMO à la hanche peut être mesurée au col du fémur (n°1), au trochanter (n°2), à la région inter-trochantérienne (n°3). Les logiciels fournissent aussi la mesure au Ward (n°4) ou la longueur de l'axe du col (n°5).

- **Variables de la mesure de la DMO** : La mesure de DMO n'est comparable qu'entre des sujets de même sexe et même âge. La construction de variables de DMO standardisées, par rapport à une population de référence, permet de comparer la DMO entre individus.

Le T-score utilise comme population de référence la population des femmes jeunes (au pic de la masse osseuse) normales. Le T-score est la différence entre la mesure réalisée chez un individu et la moyenne de la DMO dans cette population de référence, exprimée en fractions d'écarts-types. C'est à partir de cette valeur standardisée qu'est définie l'ostéoporose :

- **Normal** (T-score > -1) : la densité osseuse mesurée est supérieure à la moyenne des densités osseuses de référence moins un écart-type

- **Ostéopénie** (-2.5 < T-score < -1) : la densité osseuse mesurée est située entre la moyenne de référence moins un écart-type et la moyenne moins 2.5 écarts-types.

- **Ostéoporose** (T-score < -2.5) : la densité osseuse mesurée est inférieure ou égale à la moyenne de référence moins 2.5 écarts-types
- **Ostéoporose sévère** (T-score < -2.5) et présence d'une ou plusieurs fractures de fragilité.

Le Z-score est la différence entre la valeur mesurée chez l'individu et la moyenne théorique pour une population de même âge et de même sexe exprimé en fraction d'écart type.

Donc la seule différence entre ces deux scores standardisés, le T-score et le Z-score, vient de la population de référence utilisée pour l'ajustement.

L'ostéoporose et ses complications ont des répercussions économiques notables : le coût a été estimé à 15 milliards d'euros en 2 003 en Europe. L'ostéoporose est un enjeu majeur de santé publique. On estime à 200 millions le nombre de personnes souffrant de cette pathologie dans le monde. 30% des femmes ménopausées en Europe et aux USA ont de l'ostéoporose (IOF, International Osteoporosis Foundation).

Revue de la littérature

L'influence de la génétique sur la variation de la DMO est mesurée à travers l'héritabilité. Une première estimation de cette part de variance attribuable à des facteurs génétiques peut être évaluée à partir de l'étude de jumeaux. Ce type d'étude surestime la proportion de variance génétique, car l'hypothèse que l'environnement contribue de la même manière à la variance du phénotype dans les deux populations de jumeaux est sans doute inexacte. Les études de jumeaux mettent en évidence une héritabilité comprise entre 80 et 90% (Dequeker, Nijs et al. 1987; Pocock, Eisman et al. 1987; Hopper, Green et al. 1998; Videman, Levalahti et al. 2007).

De nombreuses études ont été conduites pour détecter et caractériser les facteurs génétiques de la DMO par des analyses de ségrégation ou par des études utilisant l'information apportée par des marqueurs génétiques. Toutes les analyses de ségrégation démontrent l'existence de corrélations familiales sur la DMO. L'héritabilité est estimée entre 50 à 70% (Gueguen, Jouanny et al. 1995; Cardon, Garner et al. 2000; Deng, Livshits et al. 2002; Peacock, Turner et al. 2002; Nguyen, Livshits et al. 2003; Sigurdsson, Halldorsson et al. 2008). Certains auteurs mettent en évidence l'effet d'un gène majeur (Cardon, Garner et al. 2000; Deng, Livshits et al. 2002; Nguyen, Livshits et al. 2003; Livshits, Deng et al. 2004). D'autres auteurs rapportent seulement l'effet d'une

composante polygénique. D'autres études mettent en évidence des corrélations résiduelles significatives ou des interactions gène-environnement, suggérant ainsi que la plupart de la variation de la DMO reste encore non expliquée (Gueguen, Jouanny et al. 1995; Deng, Livshits et al. 2002; Duncan, Cardon et al. 2003; Livshits, Deng et al. 2004; Pelat, Van Pottelbergh et al. 2007).

Ces résultats divergents peuvent être expliqués, en partie, aux différents types de sélection des individus (aléatoire ou à travers des sujets ayant des valeurs extrêmes de la population générale), ou par le phénotype étudié (ajusté ou non ajusté) ou encore par des corrélations résiduelles significatives. De plus, l'héritabilité de la densité osseuse varie selon le site osseux (Deng, Li et al. 1998; Duncan, Cardon et al. 2003). Par exemple, Duncan (Duncan, Cardon et al. 2003) montrent que dans des familles issues d'un sujet ostéoporotique, l'héritabilité de la densité osseuse est d'environ 60% au rachis lombaire, mais seulement de 40% au col du fémur. Elle varie également selon le sexe, ou encore la population ethnique et son mode de transmission n'est pas encore clairement élucidé (Pocock, Eisman et al. 1987; Peacock, Turner et al. 2002).

Une forte covariance phénotypique existe entre les phénotypes de la DMO à différents sites squelettiques. Plusieurs auteurs ont montré que les corrélations génétiques entre les mesures de la DMO aux rachis lombaire et à la hanche était de l'ordre de 0.60 à 0.80 (Livshits, Deng et al. 2004; Yang, Zhao et al. 2006; Wang, Kammerer et al. 2007).

Ces résultats montrent qu'il existe une composante génétique dans la variabilité de la DMO. On observe cependant peu de consistance entre les différentes analyses. Certains auteurs ont cherché à détecter et à caractériser les facteurs génétiques de la DMO en utilisant l'information apportée par des marqueurs génétiques. Étant donné les forts niveaux de corrélation génétique entre les sites osseux, comme pour l'analyse de ségrégation (Livshits, Deng et al. 2004), certains ont utilisé des approches bivariées pour localiser des régions pouvant potentiellement contenir des gènes à effets pléiotropiques. Deux grands types d'approches existent pour localiser des QTLs : des études de gènes candidats et des études par criblage du génome. Le criblage du génome peut se faire pour la recherche de liaison, mais plus récemment ces approches sont plus souvent utilisées pour la recherche d'association. Les études varient selon les phénotypes étudiés. Nous allons présenter les principaux résultats de ces différentes méthodes pour l'ostéoporose et la DMO.

Études de gènes-candidats

Environ 120 études ont été effectuées entre 2007 et 2009 (Xu, Dong et al.) et une association positive a été rapportée pour plusieurs gènes candidats. Les principaux gènes étudiés pour l'ostéoporose et la variation de la DMO sont brièvement présentés ci-dessous :

- **Le récepteur de la vitamine D** (VDR) fut un des premiers gènes candidats, étudié dans le domaine de l'ostéoporose. En effet, une carence en vitamine D induit une mauvaise minéralisation de la matrice osseuse. Les marqueurs fréquemment étudiés incluent Bsm1, Apa1, Taq1, Cdx2 et Fok1. Par exemple (Mencej-Bedrac, Prezelj et al. 2009) mettent en évidence une association entre le polymorphisme Cdx2 et la DMO au col du fémur chez des femmes pré-ménopausées d'origine caucasienne. Cependant, dans une autre étude également effectuée dans un échantillon de femmes pré-ménopausées d'origine caucasienne, (Horst-Sikorska, Ignaszak-Szczepaniak et al. 2008) ne trouve pas d'influence de ce variant sur la DMO.

- **Le collagène de type 1** (COL1A1) qui est une protéine majeure de l'os. Le polymorphisme Sp1 est un des plus étudié (Grant, Reid et al. 1996; Van Pottelbergh, Goemaere et al. 2001) mais par exemple, (Dincel, Sepici-Dincel et al. 2008) ne trouve pas d'association significative avec la DMO au col du fémur.

- **Le récepteur aux oestrogènes alpha** (ESR1) est un autre gène-candidat potentiel pour la régulation de la masse osseuse. Certains polymorphismes comme XbaI, PvuII ou la répétition TA sont retrouvés associés à la DMO au rachis lombaire dans une population d'hommes (Kastelan, Grubic et al. 2009). Dans une récente méta-analyse d'environ 20 000 participants, dont 75% étaient des femmes, (Richards, Kavvoura et al. 2009) mettent en évidence une association avec le risque de fracture. De façon surprenante, ils ne trouvent pas d'association significative avec la DMO. Pour les SNPs les plus significatifs, les effets des allèles sur la variation de la DMO variaient de 0.04 à 0.18 écart-type par copie d'allèle.

- **Le récepteur lipo-protéine 5** (LRP5). Le gène LRP5 a d'abord été identifié par criblage du génome pour la liaison. Plusieurs études ont reportées des associations significatives pour des polymorphismes de ce gène avec la variation de la DMO (Grundberg, Lau et al. 2008; Sims, Shephard et al. 2008; Richards, Kavvoura et al. 2009) ou pour l'ostéoporose (Ferrari, Deutsch et al. 2005). Dans un grand échantillon de plus de 37 000 individus participant à l'étude GENOMOS, (van Meurs, Trikalinos et al. 2008) mettent en évidence une association avec les polymorphismes Met667 et Val1330 et la variation de la DMO au rachis lombaire. Les effets des allèles sur la variation de la densité osseuse étaient relativement modestes, environ 0.15 écart-type par copie d'allèle soit un effet sur la DMO compris entre 11 et 20 mg/cm^2 au site LS et FN respectivement.

Nous pouvons voir que certains polymorphismes de gènes candidats sont associés dans plusieurs études à la variation de la DMO. Cependant, les effets des polymorphismes sur la variation de la DMO sont faibles et les résultats restent contradictoires.

Au contraire de l'approche gène-candidat qui suppose une hypothèse biologique a priori sur les gènes étudiés, certains auteurs ont cherché à caractériser la composante génétique

de la DMO par criblage du génome en utilisant l'information apportée par les marqueurs au niveau de la liaison et de l'association.

Etudes par criblage du génome

Recherche d'une liaison génétique

Environ une dizaine d'études se sont intéressées à la variation de la DMO au rachis lombaire (LS, pour Lumbar Spine), au col du fémur (FN pour Femoral Neck) ou plus généralement à la hanche. Certaines études ont également cherché à détecter des QTLs à effets pléiotropiques par des approches bivariées. Les résultats majeurs des études publiées entre 2 007 et 2 010 résumés par Xu ((Xu, Dong et al.), tableau 3) sont repris dans le tableau 4. Nous montrons dans la table uniquement les régions de liaison pouvant contenir des gènes potentiellement candidats pour la variation de la DMO.

Certains auteurs retrouvent, par criblage du génome pour la liaison, des régions contenant des gènes potentiellement candidats, comme le gène LRP5 (Kaufman, Ostertag et al. 2008). Parmi ces résultats, la majorité (7 sur 11) des auteurs ont utilisé des approches bivariées pour détecter la liaison à deux traits. Les méthodes de liaison reposent soit sur des analyses de liaisons bivariées effectuées sur les traits originaux ou sur des traits obtenus par une analyse en composantes principales (Xiong, Wang et al. 2007).

On remarque que sur 9 études, 4 utilisent le même échantillon d'individus et trois autres prennent seulement un sous-échantillon constitué de femmes de cette même cohorte. Malgré les mêmes échantillons, les résultats entre études sont très peu concordants. De plus, les régions identifiées par la liaison sont relativement larges, si bien qu'il est difficile de savoir quel est le gène contribuant le plus au signal de liaison.

Recherche d'association

Comme pour les résultats des analyses de liaison par criblage du génome, nous présentons dans le tableau 5 les principaux loci identifiés par association et pouvant contenir des gènes potentiellement candidats pour la variation de la DMO au rachis lombaire (LS), au col du fémur (FN) ou à la hanche (hip). Ces résultats sont issus de la table 4 de Xu (Xu, Dong et al.).

Comme pour les résultats de liaison, neuf études ont été publiées entre 2 007 et 2 010 pour la variation de la DMO au rachis lombaire ou au col du fémur. De nombreux polymorphismes sont identifiés, dont certains sont retrouvés dans des gènes potentiellement candidats, ou en déséquilibre de liaison avec des variants de gènes potentiellement candidats, pour la variation de la DMO.

Par une méta analyse sur un peu moins de 20 000 individus, Rivadeneira (Rivadeneira, Styrkarsdottir et al. 2009) rapportent un certain nombre de polymorphismes localisés dans des gènes potentiellement candidats pour la variation de la DMO. Il faut noter que cette méta-analyse regroupe cinq cohortes (DECODE, Twin UK, Rotterdam study, ERASMUS, Framingham Osteoporosis Study) déjà analysées pour certaines d'entre elles de manière individuelle, comme l'étude deCODE (Styrkarsdottir, Halldorsson et al. 2008; Styrkarsdottir, Halldorsson et al. 2009). Certains des gènes rapportés comme LRP5 ou ESR1 se retrouvent donc également dans les études individuelles.

Pour les études autres que la méta-analyse, on retrouve peu de concordances dans les résultats. Cela tient, en partie, à la faible puissance des études. Les polymorphismes identifiés expliquent une faible proportion de la variance phénotypique. Cela implique que de grands échantillons sont nécessaires pour augmenter la puissance de détection des variants qui ont des effets modestes sur la variation de la DMO. Par exemple, Richards (Richards, Rivadeneira et al. 2008) trouve que les parts de variance de la DMO expliquées par la variation des allèles du polymorphisme sont de l'ordre de 0.6% à 0.2% pour LS et FN au locus du gène-candidat LRP5 et d'environ 0.4% pour LS et FN au locus du gène TNFRSF11B. Les estimations des effets des allèles (coefficient de régression) sont de l'ordre de 0.10 écart-type par copie d'allèle. Ces effets ont été mis en évidence dans un échantillon de 8 557 individus (Tableau 5)**.**

Tableau 4 : Régions chromosomiques de gènes candidats potentiels pour la variation de la DMO identifiées par un LOD score ≥2 par criblage du génome pour la liaison

Cohorte	Nb Fam.	Echant-illon+	DMO++	Locus	LOD*	Gène potentiel	Référence
821 NEMO	103	H	LS	11q12-13	2.64	LRP5, TCIRG1	(Kaufman, Ostertag et al. 2008)
				17q21	3.63	COL1A1, SOST	
				21q22	2.05	COL6, COL6A2	
			FN	13q12-14	2.71	RANKL	
34 Caucasiens	1	Pop.	LS	1q36.3	3.07	WDR8, EGFL3	(Willaert, Van Pottelbergh et al. 2008)
4498 $ Caucasiens (US)	451	Pop	PC1: OF+DMO	14q32	2.61	BMP4	(Xiong, Wang et al. 2007)
		F		14q22	3.53	GLI3	
4498 $ Caucasiens (US)	451	Pop	LS/BFM	6q27	2.42	ESR1	(Tang, Xiao et al. 2007)
				7p22	2.69	IL6, RAC1	
		H		6p25-24	3.15	BMP6	
		F		6q27	2.34	ESR1	
		Pop	Hanche/ BFM	2q32	2.29	GDF8 STAT1	
				6q27	2.30	ESR1	
		H		7q21	2.59	CTR, SERPINE1, PON1	
				13q12	3.23	KL	
4498 $ Caucasiens (US)	451	Pop	LS/ SO	12p11	3.39	LRP6	(Liu, Liu et al. 2008)
				17q21	2.94	COL1A1, SOST, CHAD, HOXB	
			FN/ SO	20q11	3.65	GDF5	
4498 $ Caucasiens (US)	451	Pop	LS/TBLM	7p22	2.53	IL6, TWIST	(Wang, Deng et al. 2008)
		F		7q32	2.67	LEP	
				15q13	4.86	GREM1	
2582 $ Caucasiens (US)	451	F	LS	15q13	3.67	HERC2	(Yan, Liu et al. 2009)
2584 $ Caucasiens (US)	414	F	LS/AAM	8q24	4.59	EXT1	(Zhang, Lei et al. 2009)
2522 $ Caucasiens (US)	414	F	LS/AAM	3p25	2.36	PPARG	(Pan, Xiao et al. 2008)
			FN/AAM	22q13	3.33	EP300	

Tiré de la table 3 (Xu, Dong et al. 2010). **NEMO** : Network on Male Osteoporosis in Europe ; + : H : hommes ; F : Femmes ; Pop : population générale. ++ : Densité Minérale Osseuse ; LS= rachis lombaire ; FN= col du fémur ; TBLM : Poid du corps dégraissé; BFM= Masse graisseuse ; SO : Structure de l'os ; AAM = Age à la ménarche; OF= Risque de fracture ; BS= Taille de l'os ; PC1 : Premier axe de variation d'une ACP ; « / » : analyse de liaison bivariée. * : Statistique du rapport des vraisemblances suivant un mélange pondéré de χ^2 ; $: Ces quatre études utilisent le même échantillon d'individus. $$: Ces trois études utilisent un sous ensemble de femmes de l'échantillon utilisé en $.

Tableau 5 : Principaux loci identifiés pour la variation de la DMO par criblage du génome pour l'association

Puce	Echantillon de « découverte »		Echantillon de réplication		Trait*	Gène	Signification	Référence
Infinium 300K	5 861 (87%)	deCODE	7925 (80%)	deCODE ; DOES ; PERF	LS - hanche	RANKL ; OPG, ESR1 ; MHC; ZBTB40	$1.7x10^{-7}$ à $2.0x10^{-21}$ (n=13 773)	(Styrkarsdottir, Halldorsson et al. 2008)
Illumina 300K	6 865 (86%)	deCODE	8511 (67%)	deCODE ; DOES ; PERF	LS - hanche	SOST, MARK3, SP7, TNFRSF11A	$1.0x10^{-7}$ à $1.8x10^{-9}$ (n=15 375)	(Styrkarsdottir, Halldorsson et al. 2009)
Illumina 550K	2 094 (100%)	TwinUK	6 463 (90%)	Rott. ; Twin UK ; Chingford	LS - FN	TNFRSF11B, LRP5	$6.3x10^{-12}$ à $7.6x10^{-10}$ (n=8 557)	(Richards, Rivadeneira et al. 2008)
Meta-Analyse	19 195	Rot. ; ERASMUS ; Twin UK ; deCODE ; FOS			LS – FN	9 nouveaux locus : GPR177; CTNNB1; MEF2C; STARD3NL; FLJ42280; DCDC5; SOX6; FOXL1; CRHR1. 8 reportés ZBTB40; ESR1; C6orf97; TNFRFS11B; LRP5; SP7; AKAP11; TNFRS11A	$<5.0x10^{-8}$	(Rivadeneira, Styrkarsdottir et al. 2009)
Affy. 500K	1 000 (50%)	Caucas. (US)	1972 (71%) 600 (57%) 2965 (51%) 908 (0%) 2953 (57%)	US fam. (n=593) Ch.HP Ch. BMD Tobago FHS	LS-Hanche	ADAMTS18, TGFBR3	$1.6x10^{-2}$ à $3.5x10^{-3}$ (n=1 972 à 2 953)	(Xiong, Liu et al. 2009)
Affy. 500K	1 000 (50%)	Caucas. (US)	3 355 (59%)	FHS (n=975)	Hanche/IMC	SOX6	$1.5x10^{-6}$ à $6.8x10^{-7}$ (n=1 370)	(Liu, Pei et al. 2009)
Affy. 500K	983 (50%)	Caucas. (US)	2 557 (55%)	FHS (n=750)	FN	IL21R, PTH	$6.7x10^{-3}$ à $6.5x10^{-4}$ (n=2 557)	(Guo, Zhang et al. 2010)
Illumina 610K	800 (100%)	HKSC	720 (60%) 5346 (100%)	1) HKSC 2) HKSC; HKOS; FHS; deCODE, Twin UK	LS - FN	JAG1	$9.1x10^{-5}$ à $5.6x10^{-3}$ (n=720)	(Kung, Xiao et al. 2010)
Illumina 610K	1 524 (100%)	US fam	762 (100%)	AA fam.	FN	CATSPERB	$3.0x10^{-3}$ (n=762)	(Koller, Ichikawa et al. 2010)

Tiré de la table 4 de (Xu, Dong et al. 2010). DeCODE: Population d'origine Islandaise; PERF: Danish Prospective Epidemiological Risk Factor, population de danoises post-ménopausées; DOES: Australian Dubbo Osteoporosis Epidemiology Study; Rotterdam Study : Population d'origine allemande (âge≥ 55 ans); Chingford Study: Population de femmes d'origine anglaise; Twin UK: Famille de jumeaux d'origine britannique; ERASMUS: Rucphen Family study : Familles isolées du sud–ouest de la Hollande; FHS: Framingham Heart Study; ALSPAC: Avon Longitudinal Study of Parents and Children, familles de mères et enfants d'origine britannique; Chin. HP: Population d'origine chinoise avec fracture de la hanche ou non; Ch. BMD: Population d'origine chinoise; Tobaggo cohort: population d'hommes d'origine africaine;AA: Familles de femmes d'origine afro-américaine; HKSC: Hong Kong Southern Chinese; HKOS: Hong Kong Osteoporosis Study. *: LS: rachis lombaire; FN: col du fémur; IMC: Indice de Masse Corporelle.

1.6. Conclusions sur la génétique de la densité osseuse et projet NEMO

La plupart des analyses de ségrégation mettent en évidence une forte composante génétique dans la détermination de la DMO. La part de ressemblance étant estimée entre 50 à 70% pour la variation de la densité osseuse (Cardon, Garner et al. 2000; Deng, Livshits et al. 2002; Livshits, Deng et al. 2004). Les approches gènes candidats ont permis de retrouver plusieurs polymorphismes associés à la variation de la DMO. Compte tenu des avancées rapides dans les technologies de génotypage et la diminution du coût de séquençage, une alternative à l'approche gène-candidat a été la recherche de gènes par criblage du génome. A l'opposé, cette démarche ne nécessite pas d'hypothèses a priori sur les gènes impliqués dans la pathologie. Que ce soit par liaison ou par association, de nombreux loci ont été identifiés dans des régions de gènes candidats potentiels pour la variation de la DMO. Cependant, la plupart des résultats sont peu consistants (Shen, Liu et al. 2005; Xu, Dong et al. 2010). Plusieurs études se sont intéressés à la recherche de QTLs à effets pléiotropiques en utilisant l'information apportée par les marqueurs sur la liaison (Tableau 4) mais peu d'études d'association bivariées ont été réalisées (Liu, Pei et al. 2009). Les problèmes de réplication entre études tiennent en partie au fait que l'ostéoporose, et ses traits associés (la DMO), ont une étiologie complexe (Eisman 1999; Peacock, Turner et al. 2002). Prédire la maladie au regard d'un seul, ou d'un ensemble de gènes, peut s'avérer difficile.

Comme pour les résultats des analyses de liaison univariées, les régions identifiées par des analyses de liaison bivariées sont peu concordantes entre les différentes études malgré les mêmes échantillons utilisés pour certaines d'entre elles (Tableau 4). Les régions localisées par la liaison sont, de plus, relativement grandes, il est donc difficile d'identifier le gène contribuant le plus au signal de liaison.

Toutes les études d'association mettent en évidence des effets modestes des polymorphismes identifiés pour la variation de la DMO. Les parts expliquées sont en général inférieur à 1%. Ces effets nécessitent d'utiliser des échantillons extrêmement grands comme par exemple dans l'étude d'association GENOMOS (van Meurs, Trikalinos et al. 2008) basée sur environ 37 000 individus.

En raison du lien entre la ménopause et la chute de la densité minérale osseuse, la plupart des études utilisent des populations de femmes pré ou post ménopausées. L'avantage de travailler sur ce type de cohorte est que les femmes ont un risque plus élevé que la normale de développer une ostéoporose, ce qui facilite la constitution de l'échantillon. Cependant, après la ménopause, la carence en œstrogènes est un des principaux facteurs induisant une diminution de la densité osseuse. L'utilisation d'une telle cohorte risquerait de faire intervenir des gènes non liés directement à la densité osseuse mais qui affecteraient, par exemple, la production d'oestrogènes.

Malgré l'évidence consistante d'une composante familiale de l'ostéoporose, les nombreuses études conduites sur cette maladie et ses phénotypes associés apportent des réponses divergentes quant à son déterminisme génétique. Ces résultats s'expliqueraient d'une part par l'hétérogénéité de la maladie et les différents modes de recensement utilisés. D'autre part, les données épidémiologiques suggèrent une variation du poids relatif des différents facteurs de risques génétiques et environnementaux avec la ménopause chez la femme. L'étiologie de l'ostéoporose chez l'homme, affranchie de ce facteur hormonal, serait moins hétérogène, en particulier chez les moins de 70 ans. Il est donc probable que dans les données recrutées par des hommes ostéoporotiques, les phénotypes soient plus homogènes. L'étude NEMO est construite sur ce protocole.

Projet NEMO

Le projet NEMO (Network on Male Osteoporosis in Europe) (Cohen-Solal, Baudoin et al. 1998; Baudoin, Cohen-Solal et al. 2002; Van Pottelbergh, Goemaere et al. 2003) est un projet collaboratif qui a débuté en 1995. Ce projet implique les équipes du Dr. Marie Christine de Vernejoul (INSERM, U.606, Hôpital Lariboisière, Paris), du Dr. Jean Marc Kaufman (Département d'Endocrinologie, Université Gent, Belgique) ainsi que du Dr. Maria Martinez (INSERM, U.563, Toulouse).

L'objectif principal du programme de recherche est de caractériser la contribution génétique expliquant la variabilité inter-individuelle de la DMO à LS et FN. La plupart des cohortes existantes pour la DMO et/ou l'ostéoporose sont construites à partir de la population de femmes et le plus souvent post-ménopausées. Dans cette population, les phénotypes DMO sont probablement assez hétérogènes, résultant d'interactions complexes de nombreux facteurs, génétiques ou non (hormones, alimentation, activité physique,..). En revanche, les données collectées dans le cadre du projet collaboratif NEMO offrent la possibilité d'étudier la génétique de la DMO dans une population originale pour laquelle les phénotypes DMO peuvent être potentiellement plus homogènes : hommes relativement jeunes (âge compris entre 19 et 67 ans) et sélectionnés pour des valeurs extrêmes de densité osseuse.

Dans un premier temps, l'équipe a caractérisé la composante génétique de la DMO par analyse de ségrégation (Pelat, Van Pottelbergh et al. 2007). Les analyses ont montré de fortes corrélations familiales pour chacun de ces traits mais ne permettent pas de conclure à l'effet d'un gène majeur. Cependant, la présence d'effets d'interaction gène-majeur/polygénique (G×C) dans le modèle a permis de conclure avec une forte évidence à l'effet d'un gène majeur régulant FN mais pas pour LS. Ces résultats suggèrent une détermination génétique site-spécifique de la densité osseuse

Par la suite, nous avons cherché à identifier les QTLs impliqués en utilisant l'information apportée par les marqueurs génétiques pour la liaison et l'association. Le premier axe de recherche est un criblage du génome pour la liaison dans l'échantillon de familles NEMO, sélectionnées à travers des hommes ayants des valeurs extrêmes (les proposants). Nous avons d'abord recherché des QTLs à effets site spécifique puis des QTLs à effets pléiotropiques. L'objectif du deuxième travail de recherche est l'identification du (ou des) variants causaux impliqués dans la variation de la DMO par criblage du génome pour l'association. L'étude a été conduite dans l'échantillon NEMO d'hommes non apparentés et sélectionnés pour des valeurs basses ou élevées à la DMO. De plus, la recherche d'association a été conduite par analyse jointe de LS et FN. A notre connaissance, aucune autre étude d'association à grande échelle n'a utilisé un tel design : analyse d'association bivariée dans des échantillons de sujets recensés aux extrêmes de la distribution des traits.

Chapitre 2

2. Recherche de QTLs par analyses de liaison univariée et bivariée

Une première approche, pour mettre en évidence l'effet d'un locus sur la variation d'un trait, est de le localiser sur le génome en utilisant l'information apportée par le marqueur sur la liaison. Comme nous l'avons déjà noté, les paramètres génétiques et le mode de transmission des traits complexes ne sont généralement pas bien connus. Pour ces caractères, les méthodes de liaison, les plus appropriées, sont celles qui ne requièrent pas la spécification du modèle génétique (approche « model-free »).

La densité minérale osseuse à différents sites squelettiques est fortement corrélée. Comme nous l'avons vu au chapitre 1.4, une part de cette corrélation pourrait être expliquée par des facteurs génétiques communs. Plusieurs auteurs ont rapporté des corrélations génétiques relativement fortes (entre 0.6 et 0.8) entre les mesures de la densité osseuse à différents sites (Livshits, Deng et al. 2004; Yang, Zhao et al. 2006) et récemment plusieurs études ont utilisé des méthodes de liaison bivariées pour localiser des QTLs à effets pléiotropiques sur différents traits associés à l'os (Tang, Xiao et al. 2007; Wang, Liu et al. 2007; Liu, Liu et al. 2008; Pan, Xiao et al. 2008; Wang, Deng et al. 2008; Zhang, Lei et al. 2009). En effet, l'analyse jointe de traits corrélés peut, théoriquement, s'avérer plus puissante que l'analyse univariée de chaque trait (Jiang and Zeng 1995; de Andrade, Thiel et al. 1997; Allison, Thiel et al. 1998; Mangin, Thoquet et al. 1998; Amos, de Andrade et al. 2001).

Dans le cas d'échantillons de grandes familles (généalogies), l'analyse de liaison « model-free » peut se faire par la méthode des composantes de la variance (VC). Cette méthode a été généralisée à l'analyse multivariée de traits corrélés (Hopper and Mathews 1982; Amos 1994). Cependant, la distribution asymptotique du test de liaison multivarié est complexe, principalement à cause de la non-indépendance des paramètres de liaison et des contraintes imposées sur ces paramètres. Différentes distributions théoriques ont été proposées (Almasy, Dyer et al. 1997; Amos, de Andrade et al. 2001; Wang 2003). Une

solution est d'estimer les niveaux de significations empiriquement, mais les temps de calculs sont prohibitifs : l'estimation empirique requiert de générer et d'analyser un très grand nombre (plusieurs centaines de milliers) de réplicats et les temps de calculs augmentent exponentiellement avec la taille des familles et/ou le nombre de marqueurs simultanément étudiés. D'autres alternatives ont été proposées. Certains auteurs proposent de contraindre le paramètre de corrélation génétique du QTL aux bornes de l'espace des valeurs possibles ($\rho_G \pm 1$) (Almasy, Dyer et al. 1997; de Andrade, Thiel et al. 1997; Amos, de Andrade et al. 2001). On se place donc dans le cas de pléiotropie pure, à savoir le QTL contribue à la variation de chacun des traits (chapitre 1.4). Là encore, comme nous le détaillons plus loin dans ce chapitre, différentes distributions asymptotiques ont été proposées. D'autres approches sont basées sur des techniques de réduction des données. Le principe général est de générer un ensemble de phénotypes indépendants (combinaisons des traits corrélés) pour lesquels on peut conduire des analyses univariées, puis de construire un test global de liaison bivarié basé sur la combinaison additive des analyses univariées (Mangin, Thoquet et al. 1998). Cependant, certaines études de liaison, utilisant ces techniques de réduction des données, basent l'inférence d'un QTL à effets pléiotropiques à partir du meilleur des niveaux de signification associés à chacun des axes de variation. Par ailleurs, souvent elles ignorent aussi le problème des tests multiples qui en résulte (Karasik, Cupples et al. 2004; Tan, Liu et al. 2008).

Ici, nous nous sommes intéressés à ces différentes approches de liaison bivariées, principalement celles basées sur la décomposition de la variance (VC). Nous les avons appliquées à notre criblage du génome pour la liaison de la DMO à LS et FN dans les données familiales NEMO. Dans un premier temps, nous présentons les résultats du criblage du génome pour la recherche de QTLs à effets site-spécifiques (analyse de liaison univariée). Nous présentons ensuite les résultats obtenus par différents tests bivariés lors de notre recherche de QTLs à effets pléiotropiques (paragraphe 2.1). Plusieurs auteurs ont proposé et évalué différentes distributions asymptotiques des tests bivariés basés sur la méthode VC mais les résultats restent contradictoires. Nous avons donc estimé la distribution empirique des tests bivariés dans nos données, puis évalué et comparé l'adéquation des différentes distributions asymptotiques proposées à ce jour dans la littérature (paragraphe 2.2).

2.1. Criblage du génome de la DMO

2.1.1. Matériel et méthodes

2.1.1.1. Les données NEMO

Mesures phénotypiques : La Densité Minérale Osseuse (g/cm^2), aux deux sites squelettiques LS et FN, a été mesurée par la technique couramment utilisée DEXA (Dual Energy Xray Absorptiometry) par la compagnie Hologic (Hologic QDR 200 device ; Software version 7.20, Hologic Inc., Bedford, MA, USA) pour la Belgique et Lunar pour la France (Lunar Corp., Madison, WI). Les machines sont calibrées quotidiennement et les coefficients de variation entre plusieurs mesures étaient inférieurs à 1% pour chacun des centres. Les individus d'une même famille ont été mesurés sur le même appareil.

Au rachis lombaire, le phénotype quantitatif utilisé est la combinaison de la mesure de la densité osseuse entre les vertèbres L2 et L4 pour la compagnie Lunar et entre L1 et L4 pour la compagnie Hologic.

Les mesures phénotypiques de la ***DMO*** (g/cm^2) provenant des deux appareils de mesures ont été ajustées au moyen d'une régression linéaire dans laquelle la covariable *appareil* est binaire. Les mesures de la DMO également recueillies sont les mesures du ***Z-score***.

Les Proposants : Pour être inclus dans la base comme proposant, l'individu devait être un homme ayant une valeur de Z-score $\leq$-2 au rachis lombaire (LS pour Lumbar Spine) ou au col du fémur (FN pour Femoral Neck) (ostéoporose idiopathique sévère) et âgés entre 19 et 67 ans. Les proposants sont donc issus du quantile à 2.5% de la DMO dans la population générale. Ils ont été identifiés à la suite d'une fracture ($\approx$ 49%) ou à cause d'une faible densité osseuse lors d'un examen de routine. Les proposants avec une histoire de causes secondaires d'ostéoporose ont été exclus des analyses.

Les Apparentés : Les informations familiales ont était collectées pour tous les apparentés du premier degré (parents, enfants, germains), pour les conjoints, ainsi que pour les apparentés du deuxième degré et plus. Les apparentés âgés entre 19 et 85 ans acceptant de participer à l'étude ont subit le même examen clinique que les proposants. Ces individus étaient pris en compte dans la base à condition d'être en bonne santé, de ne jamais avoir suivie de traitement médicamenteux pouvant interférer avec le métabolisme osseux.

Les données épidémiologiques pour le sexe, l'âge, le poids et la taille ont été recueillies au moment de l'examen clinique. Le consentement de tous les individus a été obtenu et validé par le comité d'éthique des deux hôpitaux respectifs (L'hôpital de Gent et l'hôpital Lariboisière).

Caractéristiques des Données familiales

Au totale la cohorte NEMO inclue 103 familles, avec au moins deux germains pour lesquels de l'ADN était disponible, d'origine caucasienne recrutées en Belgique (n=72) et en France (n=31) entre 1995 et 2005. Le tableau 6 décrit les principales caractéristiques des familles NEMO selon leurs tailles et le nombre de générations. La taille des familles varie entre 4 et 64 individus par famille avec une taille moyenne de 8.

Tableau 6 : Caractéristiques des familles NEMO

Familles	Taille des familles (S)				Nombre de génération (G)		
	Moyenne (min-max)	S=4	S=5	S=6	Moyenne (min-max)	G=2	G=3
103	8 (4 - 64)	27	19	19	2.5 (2 - 4)	54	45

Parmi les 103 proposants, 84 individus ont un Z-score ≤ -2 au rachis lombaire, 5 au col du fémur et 14 aux deux sites squelettiques exclusivement. Le tableau 7 décrit les principales caractéristiques des données NEMO dans l'échantillon total et par catégorie de sujets : proposants et apparentés hommes/ femmes.

Le tableau 8 donne les niveaux de corrélations entre les traits LS et FN et les covariables. Les données phénotypiques étaient disponibles pour 589 individus dont 103 proposants et 718 apparentés. Les proposants sont en moyenne plus âgés (≈47ans) que leurs apparentés avec un IMC similaire aux groupes des apparentés hommes et femmes (≈24 kg/cm^2). Comme attendu, la valeur moyenne de DMO à LS et FN chez les proposants est plus faible que chez leurs apparentés. L'écart de la DMO chez les proposants à la moyenne de l'échantillon total est plus fort pour LS (0.92 vs 0.78) que pour FN (0.77 vs 0.70). Ceci reflète notre schéma de sélection des individus. En effet, la majorité des proposants sont sélectionnés pour leur valeur phénotypique au site LS. La DMO est similaire entre le groupe des apparentés hommes et femmes. Les corrélations de la DMO avec la covariable IMC sont relativement faibles ou proche de 0. Les niveaux de corrélation de LS et FN sont les plus forts et les plus faibles dans le groupe des apparentés femmes (0.78) et chez les proposants (0.50) respectivement. La figure 6 montre les histogrammes de la distribution de la DMO pour LS et FN dans les différents groupes.

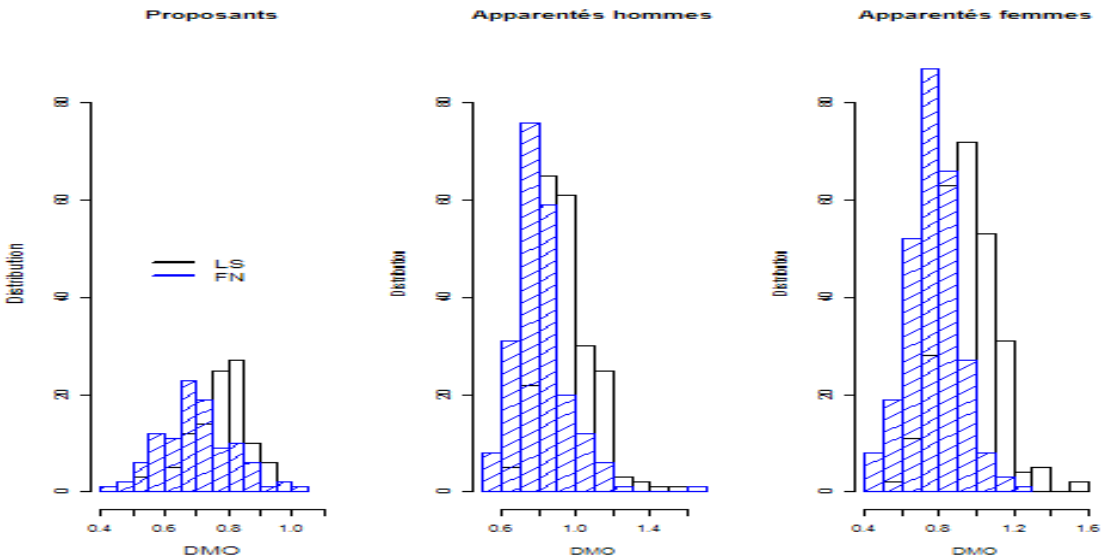
Tableau 7 : Distribution (Moyenne (écart-type) [min; max]) de la DMO et des covariables âge et IMC par groupe d'apparentés

	Total	Proposants	Hommes	Femmes
N	821	103	332	386
Sujets avec covariables et DMO-LS ou FN	589	103	215	271
Age (année) [min;max]	43.1 ± 15.7 [19;82]	46.9 ± 11.7 [19;67]	40.4 ± 15.9 [19;80]	43.8 ± 16.5 [19;82]
IMC (kg/cm^2) [min;max]	24.5 ± 4.1 [14.8;45.2]	24.1 ± 3.4 [14.8;34.5]	24.6 ± 3.7 [16.8;37.0]	24.6 ± 4.7 [16.4 ; 45.2]
DMO (g/cm^2)				
LS [min;max]	0.92 ± 0.15 [0.51;1.60]	0.78 ± 0.09 [0.51;1.04]	0.95 ± 0.14 [0.64;1.53]	0.95 ± 0.16 [0.59;1.60]
FN [min;max]	0.77 ± 0.14 [0.41;1.61]	0.70 ± 0.11 [0.41;1.01]	0.82 ± 0.14 [0.50;1.61]	0.74 ± 0.13 [0.45;1.21]

Tableau 8 : Corrélations de la DMO et des covariables.

	Total	Proposants	Hommes	Femmes
LS, FN	0.72	0.50	0.68	0.78
LS, IMC	0.06	0.06	0.00	0.07
FN, IMC	0.06	0.25	-0.04	0.06

Figure 6 : Histogrammes des distributions de la DMO aux sites LS et FN (hachuré) dans les groupes des proposants et des apparentés hommes et femmes.



Facteurs de risque associés à la DMO

Préalablement aux analyses de liaison, les données phénotypiques de DMO ont été ajustées pour les facteurs de risque trouvés significativement associés ($p<0.05$) à la variation de la DMO. Ces facteurs de risque incluent l'âge, le sexe et l'IMC (Indice de Masse Corporelle ; Poid/taille²). Nous avons également pris en compte le facteur âge^2, afin de tenir compte de possibles relations non linéaires entre la DMO et l'âge

.

Pour supprimer les effets de ces covariables et les possibles interactions entre elles, chaque phénotype de la DMO (LS et FN) a été ajusté au moyen d'une régression linéaire multiple dans chacun des trois groupes séparément, proposants et apparentés hommes et femmes ($DMO = \beta_0 + \beta_1 \times age + \beta_2 \times age^2 + \beta_3 \times IMC$). La part expliquée par ces covariables sur LS et FN ainsi que les coefficients de régression associés sont résumés dans le tableau 9 pour chacun des groupes.

Tableau 9 : Coefficients de régression, part expliquée (R^2) et signification statistique des effets des covariables sur la DMO pour les trois groupes d'individus proposants, hommes et femmes.

DMO	constante	âge	âge^2	IMC	R^2	P (F test)
Proposants						
LS	0.72	0.05	-0.03	0.02	0.19	2.10^{-5}
FN	0.70	-0.14	0.02	0.06	0.29	4.10^{-8}
Hommes						
LS	0.90	-0.03	-	0.02	0.03	0.015
FN	0.82	-0.14	0.02	0.04	0.22	7.10^{-12}
Femmes						
LS	0.90	0.01	-0.02	0.03	0.20	3.10^{-13}
FN	0.77	-0.03	-0.01	0.03	0.35	$<10^{-16}$

Les parts expliquées (R^2) par les facteurs de risque sont relativement importantes (≥20%), sauf dans le groupe des apparentés hommes pour le phénotype LS.

Par la suite, les trois catégories ont été regroupées, puis la DMO a été ajustée pour le facteur sexe conditionnellement au statut des proposants (hommes). Sur 588 individus, la valeur de DMO-FN chez un individu montrait une valeur aberrante (valeur phénotypique au dessus de trois écarts-types). La valeur phénotypique à FN a donc été mise comme manquante.

Les distributions des résidus de la DMO pour chacun des sites sont présentées dans le tableau 10. Après ajustement pour les facteurs de risque, la DMO à LS et FN est significativement plus petite ($p<10^{-15}$) dans le groupe des proposants que chez les apparentés. La corrélation entre LS et FN- DMO est de 0.63 et 0.66 chez les apparentés hommes et femmes respectivement et égale quasiment au double de celle observée chez les proposants (0.35). Ces différences dans les estimations sont expliquées par la sélection sur la valeur basse de DMO chez les proposants.

Tableau 10 : Distribution (moyenne±écart-type) des résidus de la DMO au site LS et FN chez les proposants et les apparentés et corrélation phénotypique

	Total	Proposants	Hommes	Femmes
N (# Sujets phénotypés)	821 (589/588)	103 (103)	332 (215/214)	386 (271)
DMO (g/cm^2)				
LS	0.86 ± 0.14	0.71 ± 0.08	0.89 ± 0.13	0.99 ± 0.13
[min-max]	[0.46-1.39]	[0.46-0.88]	[0.61-1.38]	[0.59-1.39]
FN	0.86 ± 0.11	0.76 ± 0.09	0.88 ± 0.12	0.88 ± 0.10
[min-max]	[0.54-1.48]	[0.54-0.97]	[0.60-1.48]	[0.67-1.30]
Corr. (LS ; FN)	0.69	0.35	0.63	0.66

Afin d'obtenir une distribution normale des traits quantitatif LS et FN, les données phénotypiques ont été transformées dans la base du logarithme népérien. Les significations statistiques obtenues par le test de Shapiro sont 0.09 pour les deux traits LS (Skewness=-0.13 ; Kurtosis =0.54) et FN (Skewness=-0.20 ; Kurtosis =0.33).

Au total, notre échantillon contenait 589 individus ayant des valeurs phénotypiques pour LS et 587 pour FN (1 individu n'avait pas de valeur phénotypique pour FN et 1 montrait une valeur aberrante pour ce même phénotype).

Les analyses de liaison ont été conduites pour les valeurs de la DMO à LS et FN ajustées pour les facteurs de risque associés.

Données moléculaires

Le génotypage des individus a été mené au Centre National de génotypage (CNG, Evry). Les 103 familles ont été génotypées pour un panel de 441 marqueurs autosomiques de type microsatellite. Le set de marqueurs MD10 (Applied Biosystems, Foster City, CA, USA) forme l'ensemble des marqueurs utilisés pour le criblage du génome. Ces

marqueurs ont une densité moyenne de distribution de 7.9 cM et un taux de d'hétérozygotie moyen de 75%.

Les incohérences sur les transmissions mendéliennes ou sur des taux de recombinaisons peu probables entre des marqueurs proches ont été détectées par le logiciel Pedcheck (O'Connell and Weeks 1998). Les fréquences alléliques des 441 marqueurs sont estimées dans nos données familiales à l'aide du programme Vitesse 2.0 (O'Connell and Weeks 1995). L'ordre des marqueurs et les distances inter marqueurs sont obtenus à partir de la carte publique Marshfield (http:/research.marshfieldclinic.org/genetics). Au final, 610 individus avaient de l'ADN dont 219 et 288 étaient des apparentés hommes et femmes respectivement.

2.1.1.2. Méthodes d'analyse de liaison de traits quantitatifs

Les modèles de liaison de type « non-paramétrique » ne nécessitent pas de spécifier la correspondance phénotype – génotype. On conclut à la liaison lorsque la ressemblance au trait entre apparentés et la ressemblance au marqueur étudié sont corrélées. Pour des généalogies de taille plus complexe, une des méthodes les plus largement utilisée est celle des composantes de la variance (Hopper and Mathews 1982; Amos 1994). Ces modèles VC (pour Variance Components) offrent l'avantage de pouvoir analyser les généalogies sans les casser en paires d'apparentés. Elle permet aussi de modéliser l'effet des covariables. Cette approche est en général plus puissante que celle basée sur la méthode de Haseman-Elston (Sham and Purcell 2001). Notre criblage du génome pour la liaison de DMO a été conduit sous le modèle VC.

Modèle VC univarié

Soit y_{ji} la valeur du trait du j^{ème} individu de la i^{ème} famille. Soit g_{ji} son génotype au QTL affectant le trait (composante génétique G, équation 1.5). Au QTL, la relation phénotype-génotype dépend des paramètres génétiques comme les moyennes génotypiques μ_G, la variance résiduelle σ^2, et les effets β de covariables X. Le modèle suppose :

$$y_{ji} = \mu + g_{ji} + c_{ji} + X_{ji}\beta + e_{ji}$$

où μ est la moyenne du trait et C est la composante polygénique.

Les termes g_{ji}, c_{ji} et e_{ji} ne sont pas observables, $E(y_{ji}) = \mu + X_{ji}\beta$ et $y_i \sim N\left(\mu + X_i\beta, \Omega_i\right)$.

Les termes non observables sont supposés non corrélés. Ils sont modélisés au travers de la matrice (Ω_i) de variance-covariance de y_i. Soit M un marqueur génétique lié au QTL et m_{ji}, les génotypes au marqueur observés chez le j^{ème} individu de la i^{ème} famille. Le modèle VC modélise la structure de covariance des données de la famille i conditionnellement au statut IBD observé au marqueur entre les individus et à leur degré d'apparentement. La covariance est décomposée en trois composantes : le QTL, qui est génétiquement lié à M, C, la composante polygénique et E, la composante résiduelle.

$$\Omega_{jl} = \begin{cases} \sigma_A^2 + \sigma_C^2 + \sigma^2 & \text{si j = l} \\ \sigma_{\mathrm{A}}^2 \times \Pi_{\mathrm{jl}} + \sigma_{\mathrm{C}}^2 \times 2\Phi_{\mathrm{jl}} & \text{si j} \neq \text{l} \end{cases}$$

où Π est la matrice des proportions IBD au marqueur entre apparentés de la i^{ème} famille ; σ_A^2 est la variance des effets additifs du QTL ; Φ_i est la matrice des coefficients d'apparentement entre paires d'individus de la famille i (*Kinship* coefficient). Les valeurs de Φ_i sont données dans tableau 2. Par exemple $2\Phi_i = 1/2$ entre un parent et un enfant ou entre des germains ; la composante résiduelle suit une loi normale $N\left(0, \sigma^2 I\right)$ où I est la matrice identité.

La densité de la loi de y_i s'écrit :

$$f(y_i) = (2\pi)^{-\frac{n_i}{2}} \times |\Omega_i|^{-\frac{1}{2}} \times \exp\left(-\frac{1}{2}(y_i - \mu)'\Omega_i^{-1}(y_i - \mu)\right)$$

où n_i est la taille de la $i^{ème}$ famille. La fonction de vraisemblance pour un ensemble de k familles supposées indépendantes est de la forme :

$$L(\alpha \mid y) = \prod_{i=1}^{k} f(y_i \mid \alpha) = \prod_{i=1}^{k} (2\pi)^{-\frac{n_i}{2}} \times |\Omega_i|^{-\frac{1}{2}} \times \exp\left(-\frac{1}{2}(y_i - \mu)'\Omega_i^{-1}(y_i - \mu)\right)$$

où $\alpha = \left(\mu, \sigma_A^2, \sigma_C^2, \sigma^2\right)$ est l'ensemble des paramètres du modèle VC.

Les principales hypothèses qui peuvent être testées sont :

– L'absence d'effets polygénique et d'effets du QTL (modèle sporadique) : σ_C^2 =0, σ_A^2=0. Les valeurs observées du trait pour les individus de la famille i, $i = 1,...,k$ sont modélisées par une régression linéaire standard de la forme :

$$y_i = \mu + X_i\beta + e_i$$

La matrice de variance/covariance phénotypique entre des individus apparentés j et l de la famille i est de la forme :

$$\Omega_{jl} = \begin{cases} \sigma^2 & \text{si j = l} \\ 0 & \text{si j} \neq \text{l} \end{cases}.$$

– Présence d'effets polygéniques et pas de liaison entre le trait et le marqueur (modèle polygénique) : $\sigma_C^2 \neq 0$, $\sigma^2_A=0$. La comparaison des vraisemblances de ces deux modèles emboîtés (sporadique et polygénique) donne une idée de la part de variance phénotypique expliquée par les effets polygénique (l'héritabilité). Le modèle polygénique est l'hypothèse nulle de non liaison trait-marqueur. Lorsque le QTL et le marqueur M, ne sont pas liés (θ=1/2) la proportion IBD est celle attendue à priori entre 2 individus sachant leur lien d'apparentement et donc, $\Pi = \Phi$.

– Liaison trait-marqueur (modèle de liaison) : $\sigma_C^2 \neq 0$, $\sigma_A^2 \neq 0$. Si le QTL et M sont génétiquement liés (θ<0.5) on s'attend à ce que des apparentés qui se ressemblent pour le phénotype se ressemblent aussi au marqueur en partageant des allèles IBD. Donc, $\Pi \neq \Phi$.

Le test de liaison entre le marqueur et le trait repose sur la composante additive des effets du QTL. Sous l'hypothèse nulle H_0 : $\sigma_A^2 = 0$. On considère l'hypothèse alternative suivante : H_1 $\sigma_A^2 > 0$. On utilise un test du rapport des vraisemblances. Comme le paramètre de variance est contraint pour des valeurs positives, le test suit sous l'hypothèse nulle de non liaison un mélange de χ^2 à 0 et 1 degré de liberté : $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ (Self and Liang 1987). On utilise en général, la statistique du **LOD score** qui est équivalente au test du log du rapport des vraisemblances, à une constante près : il faut diviser le logarithme du rapport de vraisemblance par 2xlog(10) $\approx$ 4.6 pour obtenir le LOD score.

Un exemple du modèle VC pour une famille à 2 enfants est donné en annexe (p. 159).

Analyse bivariée

Modèle de décomposition de la variance (VC)

Le modèle bivarié est une extension du modèle VC à un trait dans le cas de données répétées y_1 et y_2 (de Andrade, Thiel et al. 1997; Amos, de Andrade et al. 2001; de Andrade, Gueguen et al. 2002).

En notant σ_1^2 et σ_2^2 les variances des traits y_1 et y_2 respectivement et σ_{12} la covariance entre les traits. La matrice de variance/covariance phénotypique pour la famille i se décompose comme dans le cas univarié en une composante des effets additifs du QTL, une composante polygénique et une composante résiduelle :

$$\Omega_i = \begin{pmatrix} \sigma^2_{A,1} & \sigma_{A,12} \\ \sigma_{A,12} & \sigma^2_{A,2} \end{pmatrix} \otimes \Pi_i + \begin{pmatrix} \sigma^2_{C,1} & \sigma_{C,12} \\ \sigma_{C,12} & \sigma^2_{C,2} \end{pmatrix} \otimes 2\Phi_i + \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \otimes Id_i$$

Où $\otimes$ est le produit de Kronecker

Chaque composante comprend les paramètres de variances et de covariance entre traits, c'est-à-dire : $\left(\sigma^2_{A,1}, \sigma^2_{A,2}, \sigma_{A,12}\right)$; $\left(\sigma^2_{C,1}, \sigma^2_{C,2}, \sigma_{C,12}\right)$; $\left(\sigma_1^2, \sigma_2^2, \sigma_{12}\right)$.

Le test de liaison bivarié repose sur les paramètres de variance et de covariance des effets additifs $\left(\sigma^2_{A,1}, \sigma^2_{A,2}, \sigma_{A,12}\right)$ du QTL pour les traits y_1 et y_2. Les hypothèses sont :

$$\begin{cases} H_0 : \sigma^2_{A,1} = \sigma^2_{A,2} = \sigma_{A,12} = 0 \\ H_1 : \sigma^2_{A,1} > 0 \; ou \; \sigma^2_{A,2} > 0 \; ou \; \sigma_{A,12} > 0 \end{cases}$$

Sous l'hypothèse alternative, au moins un des phénotypes y_1 ou y_2 est lié au marqueur, c'est-à-dire qu'au moins une des variances des effets additifs sur les traits est non nulle.

Comme dans le cas univarié, le test de la liaison génétique pour la variation de y_1 et/ou y_2 se fait par la méthode du rapport des vraisemblances et suit sous l'hypothèse nulle un mélange pondéré de χ^2. Pour des raisons pratiques de maximisation des paramètres, le modèle VC est le plus souvent modélisé en terme de variances et de corrélations : ($\sigma^2_{A,1}$, $\sigma^2_{A,2}$ et ρ_G) ; ($\sigma^2_{C,1}$, $\sigma^2_{C,2}$ et ρ_C) ; (σ_1^2, σ_2^2 et ρ).

Les paramètres de liaison ne sont pas indépendants. Ainsi, si $\rho_G > 0$ alors $\sigma^2_{A,1} > 0$ et $\sigma^2_{A,2} > 0$. Si $\sigma^2_{A,1} = 0$ (ou $\sigma^2_{A,2} = 0$) alors ρ_G n'est plus un paramètre du modèle. La réduction de la dimension de l'espace des paramètres sous l'hypothèse nulle rend la distribution asymptotique du rapport de vraisemblance complexe. Comme nous le

détaillons plus loin (paragraphe 2.1.2), différentes distributions ont été proposées mais, à ce jour la distribution exacte n'est pas connue.

Des alternatives ont été proposées comme par exemple celle de Amos (Amos, de Andrade et al. 2001) ou l'approche de Mangin (Mangin, Thoquet et al. 1998) basée sur une combinaison linéaire des traits d'intérêts.

La première approche (Amos, de Andrade et al. 2001) propose de contraindre la corrélation génétique des effets additifs aux bornes de l'espace $(\rho_G = \pm 1)$. La covariance génétique entre les traits y_1 et y_2 n'est plus un paramètre à estimer ($\sigma_{A,12} = \pm\sqrt{\sigma^2_{A,1}\sigma^2_{A,2}}$.)

Les hypothèses du test sont :

$$\begin{cases} H_0 : \sigma^2_{A,1} = \sigma^2_{A,2} = 0 \\ H_1 : \sigma^2_{A,1} > 0 \; et \; \sigma^2_{A,2} > 0 \end{cases}$$

Sous l'hypothèse nulle de non liaison entre le QTL et les traits y_1 et y_2, les effets du QTL sur les traits sont nuls tandis que sous l'hypothèse de liaison génétique, le QTL exerce des effets pléiotropiques sur les traits y_1 et y_2. Le test de la liaison génétique entre les traits et le QTL est effectué par la méthode du rapport des vraisemblances. Cependant, là encore, la distribution asymptotique du test de liaison est complexe et deux distributions théoriques ont été proposées (paragraphe 2.1.2).

Approche ACP

La deuxième alternative (Mangin, Thoquet et al. 1998) propose de créer une combinaison linéaire des traits d'intérêts y_1 et y_2. Plus spécifiquement, les traits combinés calculés par l'ACP (notés PC_1 et PC_2) sont des combinaisons linéaires des traits d'origines y_1 et y_2. PC_1 et PC_2 sont les axes de variation de l'ACP tel que PC_1 maximise la part de variance expliquée par le modèle. Les deux axes PC_1 et PC_2 sont des vecteurs orthogonaux. Les statistiques du LOD score pour chacun des traits combinés (LOD_{PC1} et LOD_{PC2}) sont calculées par la méthode du rapport des vraisemblances.

La statistique du test de liaison à effets pléiotropiques est égale à la somme des statistiques univariées : $S\_PC = LOD_{PC1} + LOD_{PC2}$ (Mangin, Thoquet et al. 1998).

Sous l'hypothèse nulle de non liaison, les effets additifs gène majeur sur les traits combinés PC_1 et PC_2 sont nuls. Le test de liaison repose sur les hypothèses suivantes :

$$\begin{cases} H_0 : \sigma^2_{A,PC1} = \sigma^2_{A,PC2} = 0 \\ H_1 : \sigma^2_{A,PC1} > 0 \; ou \; \sigma^2_{A,PC2} > 0 \end{cases}$$

Les hypothèses sont testées par la méthode du rapport des vraisemblances. Sous l'hypothèse nulle, le test suit un mélange pondéré de χ^2 à 0, 1 et 2 degrés de liberté :

$$S\_PC \sim \frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$$

2.1.2. Problématique : distributions asympotiques des tests VC bivariés

Dans le cas de deux traits y_1 et y_2, les trois paramètres de liaison ne sont pas indépendants ; la covariance des effets additifs du gène majeur n'est plus définie dès que l'une des deux variances sur les traits y_1 ou y_2 est nulle ($\sigma_{A,12} = 0$ dès que $\sigma^2_{A,1} = 0$ ou $\sigma^2_{A,2} = 0$).

Nous notons *Non-contraint* le test de liaison bivarié des composantes de la variance (Almasy, Dyer et al. 1997; de Andrade, Thiel et al. 1997; Amos, de Andrade et al. 2001). La loi asymptotique du test de liaison bivarié est complexe et pas encore connue à ce jour. Trois distributions ont été proposées (Tableau 11) et évaluées par simulations. La première distribution (de Andrade, Thiel et al. 1997) suppose un mélange de χ^2 à k=1, 2 ou 3 degrés de liberté avec pour poids respectifs ¼, ½ et ¼. Celle proposée par Almasy (Almasy, Dyer et al. 1997) suppose des poids issus d'une loi binomiale de paramètres 2 et ½, $B(2,1/2)$ pour k=0, 1 et 2 degrés de liberté. La troisième (Amos, de Andrade et al. 2001) suppose k=0, 1 et 3 degrés de liberté. Ces distributions reposent sur le raisonnement suivant. L'espace des paramètres de liaison se divisent en 4 quadrants : le premier quadrant est celui où les 2 variances sont strictement positive ; le 2ème est celui où seule une variance ($\sigma^2_{A,1}$) est strictement positive ; le 3ème quadrant est son miroir, c'est-à-dire celui où l'autre variance ($\sigma^2_{A,2}$) est strictement positive ; le 4ème quadrant est celui où les 2 variances sont nulles. La probabilité de l'un ou l'autre de ces quadrants a souvent été considérée comme équiprobable, c'est-à-dire ¼, ce qui explique les poids obtenus à partir de binomiale $B(2,1/2)$.

Nous notons *Contraint* le test de liaison pour la recherche de QTL à effets pléiotropiques sous la contrainte $\rho_G = \pm 1$ (Mangin, Thoquet et al. 1998; Amos, de Andrade et al. 2001; Wang 2003). Là encore, deux distributions du test de liaison ont été proposées (Tableau 11). L'une d'elle suppose que la probabilité que les deux paramètres de variances $\sigma^2_{A,1}$ et $\sigma^2_{A,2}$ soient égales à 0 est nulle. Celle proposée par (Amos, de Andrade et al. 2001) suppose que dans ¼ de l'espace des paramètres, les deux variances sont nulles résultant en une distribution dégénérée de χ^2 à 0 degré de liberté.

Plusieurs de ces distributions ont été évaluées empiriquement par les auteurs qui les ont proposées. La plupart de ces études ne démontrent pas la bonne adéquation des distributions proposées à la distribution empirique du test bivarié, qu'il s'agisse du test *Non contraint* ou *Contraint*. Les distributions C et D sont celles supposées par défaut dans le logiciel SOLAR (Almasy and Blangero 1998).

Nous notons *S_PC*, le test de liaison bivarié basé sur une ACP proposé par Mangin (Mangin, Thoquet et al. 1998).

Tableau 11 : Distributions asymptotiques proposées pour les tests de liaison bivariés basés sur la décomposition de la variance

Test	**Référence**	**Distributions asymptotiques**
Non-contraint	(de Andrade, Thiel et al. 1997) (A)	$\frac{1}{4}\chi^2_1 + \frac{1}{2}\chi^2_2 + \frac{1}{4}\chi^2_3$
	(Almasy, Dyer et al. 1997) (B)	$\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$
	(Amos, de Andrade et al. 2001) (C)	$\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_3$
Contraint	(Amos, de Andrade et al. 2001) (D)	$\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$
	Com. Pers.; (Wang 2003) (E)	$\frac{1}{2}\chi^2_1 + \frac{1}{2}\chi^2_2$
S_PC	(Mangin, Thoquet et al. 1998)	$\frac{1}{4}\chi^2_0 + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2$

2.1.3. Analyses de liaison des données NEMO

Lorsque les données au marqueur sont incomplètes, par manque d'informativité du marqueur et/ou parce que les individus ne sont pas tous génotypés, l'estimation du nombre d'allèles IBD au marqueur est ambiguë. Utiliser l'information apportée par plusieurs marqueurs simultanément permet de mieux spécifier les coefficients IBD mais ceci requiert beaucoup de temps de calcul. Ils sont même parfois impossibles à calculer dans de très grandes familles. Les algorithmes sont souvent limités pour des familles de grande taille (Lander and Green 1987) et/ou lorsqu'on augmente le nombre de marqueurs conjointement analysés (Elston and Stewart 1971). Pour réduire les temps de calcul tout en considérant des familles de taille complexe, certains algorithmes comme celui utilisé dans SOLAR (Almasy and Blangero 1998) utilise alors plutôt des approximations des IBD au QTL. Pour calculer la matrice des IBD dans les familles NEMO, nous avons opté pour des algorithmes de calcul exact. La taille des familles étant relativement grande (de 4 à 64 individus par famille ; 8 en moyenne), nous avons utilisé le logiciel LOKI (Heath 1997) pour calculer la matrice des IBD.

Les analyses de liaison multipoints pour la variation de la DMO à LS et FN ont été faites sur l'ensemble des 22 autosomes par la méthode des composantes de la variance (VC) implémentée dans le logiciel SOLAR (Almasy and Blangero 1998). Les traits LS et FN ajustés pour les covariables significatives ont été analysés séparément puis conjointement.

Pour tous les tests de liaison, nous rapportons les statistiques du LOD score. Les niveaux de signification univariés son calculés pour des mélanges de χ^2 à 0 et 1 degré de liberté.

Pour la recherche de QTL à effets pléiotropiques, nous présentons les résultats des analyses de liaisons bivariées basées sur les phénotypes originaux c'est-à-dire LS et FN (*Non-contraint* et *Contraint*) et les résultats utilisant une analyse en composante principale des traits originaux (*S_PC*). Pour chacune des approches (notée approche VC et approche ACP) nous montrons les meilleurs résultats identifiées par les tests bivariés ou univariés.

Pour les niveaux de signification statistique du test *Non-contraint* et du test *Contraint* nous avons utilisé les mélanges de distributions utilisés par défaut dans SOLAR (Almasy and Blangero 1998), c'est-à-dire les distributions C et D du tableau 11.

Les seuils du test *Non-contraint* sont calculés à partir d'un mélange de χ^2 à 0, 1 et 3 degrés de liberté $\left(\frac{1}{4}\chi_0^2+\frac{1}{2}\chi_1^2+\frac{1}{4}\chi_3^2\right)$. Les seuils du test *Contraint* sont calculés à partir d'un mélange de χ^2 à 0, 1 et 2 degrés de liberté $\left(\frac{1}{4}\chi_0^2+\frac{1}{2}\chi_1^2+\frac{1}{4}\chi_2^2\right)$.

Nous comparons les résultats bivariés et univariés pour la concordance des régions identifiées par l'une ou l'autre de ces méthodes (LOD score ≥1.5). Pour les analyses univariés, nous prenons le meilleur niveau de signification de LS ou FN (LS/FN) et de PC1 ou PC2 (PC1/PC2) calculés pour des seuils asymptotiques de χ^2 à 0 et 1 degré de liberté sans tenir compte du nombre de test.

2.1.4. Résultats

Recherche de QTL à effets site-spécifique

Dans nos données, l'estimation de l'héritabilité de la DMO à LS et FN est fortement significative (p-value<10^{-16}) : h^2=0.63±0.06 pour la DMO-LS et h^2=0.59±0.07 pour FN.

Les résultats des analyses de liaison univariées pour la DMO de LS et FN sur l'ensemble des 22 autosomes sont présentés dans la figure 7. Les régions positivement liées à la DMO-LS ou FN identifiées par un LOD score multipoint supérieur à 1.5 sont données dans le tableau 12.

Figure 7 : LOD score pour l'analyse de liaison de la DMO de LS et FN sur les 22 autosomes (*Manhattan Plot*).

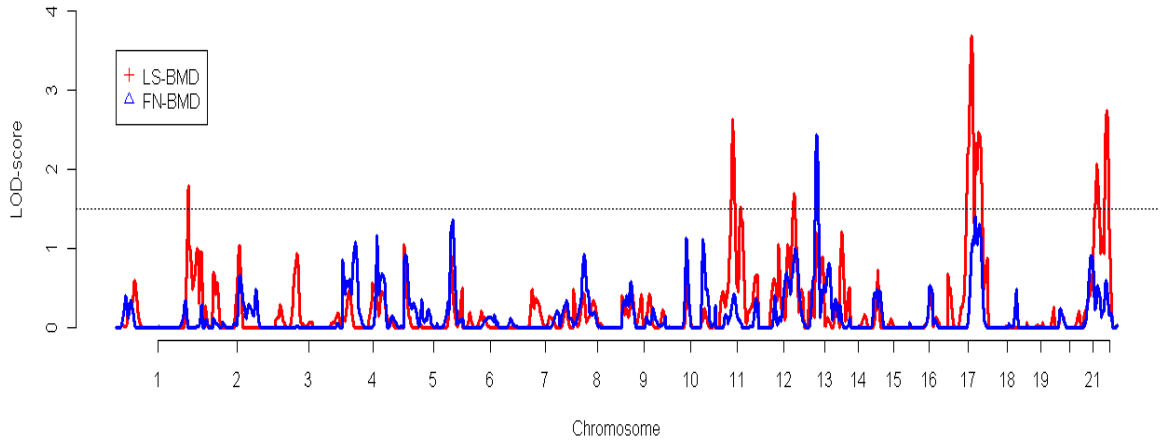


Tableau 12 : Régions chromosomiques identifiées par un LOD score≥ 1.5 pour la DMO à LS ou FN

Chr.	Locus	Marqueur	Pos.* (cM)	LOD	p-value+
LS-DMO					
1	1q42-43	D1S2800	250	1.79	2.0×10^{-3}
11	11q12-13	D11S4191	58	2.63	2.5×10^{-4}
12	12q23-24	D12S78	124	1.69	2.6×10^{-3}
17	17q21-23	D17S787	83	3.68	1.9×10^{-5}
21	21q22	D21S267	44-45	2.06	1.0×10^{-3}
22	22q11	D22S315	17-19	2.74	1.9×10^{-4}
FN-DMO					
13	13q12-14	D13S218	24-27	2.42	4.2×10^{-4}

* : Position en cM selon la carte Marshield.
\+ : Signification statistique asymptotique calculée selon un mélange de χ^2 à 0 et 1 ddl.

Sept régions sont identifiées sur les chromosomes 1, 11, 12, 17, 21, 22 pour DMO-LS et sur le chromosome 13 pour la DMO à FN. Toutes les régions de liaison détectées sont spécifiques à chacun des sites LS ou FN. Ceci est concordant avec des résultats de la littérature (Devoto, Shimoya et al. 1998; Koller, Econs et al. 2000).(Karasik, Myers et al. 2002; Wilson, Reed et al. 2003; Ralston, Galwey et al. 2005). Basé sur la méthode de Lander et Kruglyak (Lander and Kruglyak), une région est significative au niveau du génome (« genome-wide significant ») (17q21-23 : LOD=3.68, p-val=1.94×10^{-5}) et trois autres sont suggestives (11q12-13 ; 22q11 et 13q12-14 ; p-val≤4.3×10^{-4}).

La région sur le 17q21-23 a également été identifiée pour d'autres phénotypes liés à l'os, la taille des os du poignet et la largeur de la tête du fémur (Koller, Liu et al. 2001; Deng, Shen et al. 2003) avec un LOD score multipoint de 3.01 et 3.6 respectivement. La région contient deux gènes potentiellement candidats pour la variation de la DMO : le COL1A1 (Collagène de type 1) et SOST (Sclerosteros/Maladie de van Buchem). Comme nous l'avons vu, le COL1A1 est un des gènes candidats les plus étudiés (paragraphe 1.5). Il a été significativement associé avec une augmentation du risque de fracture lié à l'ostéoporose mais son rôle dans la variation de la DMO reste peu clair (Grant, Reid et al. 1996; Van Pottelbergh, Goemaere et al. 2001; Mann and Ralston 2003; Xu, Dong et al. 2010). Peu d'études ont évalué l'impact du gène SOST sur la variation de la DMO et certains des résultats sont contradictoires (Balemans, Foernzler et al. 2002; Uitterlinden, Arp et al. 2004; Sims, Shephard et al. 2008).

Nous avons localisé deux pics de liaisons suggestifs au niveau du génome pour la variation de la DMO au site LS sur le chromosome 22q11-12 (LOD=2.74) et 11q12-13 (LOD=2.63). Le QTL en 22q11 n'a pour l'instant jamais été retrouvé par d'autres études et ne concorde pas avec des régions de gènes potentiels pour la variation de la densité osseuse. La région génomique du 11q12-13 contient plusieurs gènes candidats potentiels pour la variation de la densité osseuse comme TCIRG1 (T-cell immune regulator 1) et LRP5 (Low-density lipoprotein 5). De nombreuses études ont montré que des polymorphismes de LRP5 étaient associés à la variation de la densité osseuse à LS et FN (Grundberg, Lau et al. 2008; Richards, Rivadeneira et al. 2008; van Meurs, Trikalinos et al. 2008).

Le troisième QTL suggestif pour la liaison est obtenu sur le chromosome 13q12-14 (LOD=2.42) pour la variation au site FN. Des études précédentes ont également identifié des pics de liaison dans la région du 13q14 pour différents phénotypes liés à la densité osseuse (Niu, Chen et al. 1999; Kammerer, Schneider et al. 2003; Hsu, Xu et al. 2007). Les gènes candidats potentiels dans cette région sont le gène TNFSF11 (Tumor Necrosis Factor Lignad superfamily, member 11) qui influence l'expression du gène RANKL. Ces deux gènes sont impliqués dans la régulation de la matrice osseuse. Quelques études ont évalué l'association avec des polymorphismes de RANKL pour la variation de LS et FN (Hsu, Niu et al. 2006; Xiong, Shen et al. 2006; Mencej, Albagha et al. 2008; Mencej-Bedrac, Prezelj et al. 2009).

Tous les autres pics de liaisons identifiés par un LOD score supérieur à 1.5 étaient consistants avec des études précédentes. Le pic dans la région 21q22 (LOD=2.06) était par exemple retrouvé pour la DMO aux lombaires (Streeten, McBride et al. 2006) dans un échantillon d'hommes (LOD=3.36). Les gènes candidats possibles dans cette région sont COL6A1 et COL6A2 (Collagen VI, alpha-1 and alpha-2). La région sur le chromosome 1q42-43 (LOD=1.79) coïncide avec un pic de liaison identifié dans un échantillon de 715 familles d'origine européenne (Ralston, Galwey et al. 2005). De même, le signal de liaison sur le chromosome 12q23-24 (LOD=1.69) coïncide avec une région identifiée pour la variation de FN dans un échantillon de 40 familles multiplex d'origine caucasienne sélectionnées à travers des proposants ostéoporotiques (Devoto, Spotila et al. 2005).

Recherche de QTLs à effets pléiotropiques

La corrélation phénotypique (ρ_P) observée entre LS et FN est égale à 0.65. La corrélation polygénique (ρ_C) et la corrélation résiduelle (ρ_E) sont estimées à 0.77 ± 0.05 et 0.45 ± 0.08, respectivement. Toutes ces corrélations sont positives et la corrélation polygénique est plus importante que la corrélation résiduelle. D'après la formule de

décomposition de la corrélation ρ_P (équation (1.7)), sous le modèle polygénique (c'est-à-dire sans la composante gène majeur), l'effet polygénique commun à LS et FN explique 73% de la corrélation phénotypique totale. Ceci suggère l'existence de facteurs génétiques familiaux communs. Parmi les 821 individus des familles NEMO, 587 avaient des valeurs phénotypiques aux deux traits LS et FN et ont été utilisés pour les analyses.

Approches VC

Les résultats des analyses de liaison des tests bivariés *Non-contraint* et *Contraint* sont reportés dans le tableau 13. Au total neuf régions sont identifiées par un LOD score ≥1.5 par le test de liaison *Non contraint* et huit par le test de liaison sous la contrainte (1q42-43, 5q31-33, 11q12-13, 12q23-24, 13q12-14, 17q21-23, 21q22 et 22q11). On observe une très bonne concordance des régions identifiées par ces deux tests. Une seule région sur le 5p15 n'est pas détectée par le test *Contraint* (LOD=1.38). Pour les deux tests, le meilleur LOD score est trouvé dans la région du 17q21-23 avec un LOD score de 4.01.

Tableau 13 : Statistiques du LOD score pour les régions identifiées par un LOD score multipoint ≥1.5 par le test de liaison *Non-contraint* et *Contraint* pour des seuils asymptotiques

Locus	Marqueur	Pos.* (cM)	Non-contraint	Contraint
1q42-43	D1S2800	250	1.94	1.94
5p15	D5S1981	0	2.38	(1.38)
5q31-33	D5S410	163	1.71	1.61
11q12-13	D11S4191	58	2.77	2.77
12q23-24	D12S78	124	1.80	1.76
13q12-14	D13S218	24-27	2.83	2.71
17q21-23	D17S787	83	4.01	4.01
21q22	D21S267	44-45	2.17	2.17
22q11	D22S315	17-19	2.62	2.62

* : Position en cM selon la carte Marshield.
Un LOD score <1.5 est entre parenthèse.

Approche ACP

Le trait composé PC1 est principalement pondéré par LS $\left(PC1 = 0.80 \times LS + 0.59 \times FN\right)$ et explique 85% de la covariation des traits DMO-LS et FN. L'héritabilité de PC1 est plus forte que celle de PC2. Elles sont estimées à 0.67 ± 0.06 (p-value<10^{-20}) et 0.39 ± 0.07 (p-value<10^{-11}) pour PC1 et PC2 respectivement.

Les régions chromosomiques identifiées par un LOD score multipoints ≥ 1.5 pour le test *S_PC* basé sur la somme des statistiques univariées de PC1 et PC2 (*S_PC*) sont reportés dans le tableau 14. Nous montrons également les résultats des tests de liaisons univariés de PC1 et PC2. Sept régions sont identifiées (5p15, 11q12-13, 12q23-24, 13q12-14, 17q21-23, 21q22, 22q11) par le test *S_PC*. Comme pour l'approche VC, le meilleur LOD score est trouvé sur le 17q21-23 (LOD=3.39).

Les résultats des analyses de liaisons univariés sur les axes de variation PC1 et PC2 calculés par l'ACP sont très différents. Toutes les régions identifiées par PC1/PC2 ont des effets sur PC1 seulement. Aucune région n'est identifiée pour PC2 (LOD<1.5). Le deuxième axe PC2 explique seulement 25% de la variance totale. Une région (5p15) n'est pas détectée par l'analyse univariée de PC1. Autrement pour les autres régions, les LOD scores univariés de PC1 sont similaires à ceux de *S_PC*.

Tableau 14 : Statistique du LOD score pour les régions chromosomiques identifiées par un LOD score multipoint≥1.5 par le test de liaison *S_PC* et résultats de PC1 et PC2

Locus	Marqueur	Pos.* (cM)	S_PC	PC1	PC2
5p15	D5S1981	0	2.09	(0.78)	(1.31)
11q12-13	D11S4191	58	2.00	2.00	(0.01)
12q23-24	D12S78	124	1.77	1.73	(0.04)
13q12-14	D13S218	24-27	2.33	2.02	(0.47)
17q21-23	D17S787	83	3.39	3.39	(0.00)
21q22	D21S267	44-45	2.00	2.00	(0.00)
22q11	D22S315	17-19	2.11	2.03	(0.14)

* : Position en cM selon la carte Marshield.
Un LOD score <1.5 est entre parenthèse

Comparaison des approches univariées et bivariées

Sept régions sont communes aux analyses univariées et bivariées de LS/FN. Les deux loci sur le chromosome 5 n'étaient détectés que par des méthodes bivariées. Pour toutes les méthodes univariées et bivariées, le meilleur LOD score est identifié dans la région du 17q21-22. Quelque soit l'approche utilisée (VC ou ACP), les régions identifiées par les tests de liaison bivariés sont relativement similaires. Sur neuf régions identifiées, sept régions sont communes aux deux approches (5p15, 11q12-13, 12q23-24, 13q12-14, 17q21-23, 21q22 et 22q11). La région 5p15 est détectée par *S_PC* (LOD=2.09) mais pas par le test *Contraint* (LOD<1.5). La tendance s'inverse sur le 5q31-33 (LOD<1.5 pour *S_PC* et LOD=1.61 pour le test *Contraint*). Pour ces deux régions du chromosome 5, les approches *Contraint* et *S_PC* semblent complémentaires. La région du 1q42-43 n'est pas retrouvée par les méthodes utilisant une ACP des traits d'intérêts (PC1/PC2 et *S_PC*).

2.2. Étude empirique des tests de liaison bivariés

Nous avons vu que la distribution asymptotique du test de liaison bivarié basé sur la méthode de décomposition de la variance de deux traits quantitatifs est relativement complexe (paragraphe 2.1.2). Le même problème persiste si la corrélation génétique est contrainte aux bornes de l'espace des paramètres. Les distributions proposées ont été évaluées par simulations mais les résultats restent contradictoires (Tableau 11).

Par exemple, deux études ont généré des données empiriques similaires pour évaluer les distributions asymptotiques des tests de liaison *Non-contraint* et *Contraint* (Amos, de Andrade et al. 2001; Wang 2003). Les distributions supposées par Amos pour chacun de ces tests sont respectivement les mélanges C et D. Wang a évalué les performances de ces deux tests de liaison (*Non-contraint* et *Contraint*) en développant une statistique du score asymptotiquement équivalente au rapport des vraisemblances, mais basée sur une décomposition orthogonale des effets génétiques. Les distributions supposées sont respectivement les mélanges A et E.

Pour le test *Non-contraint*, ces deux auteurs trouvent que les distributions empiriques générées sont proches des distributions asymptotiques différentes qu'ils ont supposé l'un et l'autre, alors que les données sont simulées selon des modèles génétiques similaires. Amos évalue les distributions du test pour des données générées sous le modèle *Contraint* (la corrélation génétique vaut ±1).

Concernant le test *Contraint*, Amos ne trouve pas de convergence claire des valeurs critiques empiriques aux valeurs asymptotiques lorsque la taille de l'échantillon augmente.

Pour cette raison, nous avons estimé les distributions empiriques des tests de liaison bivariés dans nos données NEMO par simulation. Nous comparons les distributions empiriques aux distributions asymptotiques proposées dans la littérature pour les trois tests de liaison bivariés *Non-contraint*, *Contraint* et *S_PC*.

2.2.1. Matériel et méthodes

Modèles simulés

Pour évaluer la distribution empirique des tests de liaisons bivariés, nous gardons la même structure familiale que celle des familles NEMO. Seulement les génotypes au marqueur sont simulés, les phénotypes restent fixes entre réplicats. Nous choisissons 2 marqueurs microsatellites espacés de 8cM, le D17S787 et D17S944 de la région 17q21-23. Le taux d'hétérozygotie est d'environ 75%. Parmi 821 individus, 610 étaient génotypés à ces deux marqueurs.

Nous obtenons une distribution de référence des statistiques de test comme suit :

- Les génotypes aux marqueurs pour un ensemble de familles de même structure que dans nos données NEMO sont simulés avec le programme SIMULATE (Terwilliger, Speer et al. 1993) sous l'hypothèse nulle de non liaison entre le trait et le locus d'intérêt. Ce programme assigne les génotypes des individus en échantillonnant les allèles, supposés en équilibre de Hardy-Weinberg, chez les individus fondateurs des familles en fonction de la fréquence des allèles. Les allèles au marqueur sont ensuite transmis aux enfants avec une probabilité mendélienne égale à ½ (Tableau 2).

- Les probabilités multipoints du nombre d'allèles partagé par descendance (allèles IBD) sont estimés tous les 1 cM par le programme LOKI 2.4.7 (Heath 1997).

- Les analyses de liaisons ont été effectuées avec le programme SOLAR 4.0.6 (Almasy and Blangero 1998) à partir des proportions estimées d'allèles IBD aux marqueurs et des données phénotypiques tel qu'observées dans l'échantillon NEMO.

Nous répétons 12 000 fois ces trois étapes pour générer les distributions empiriques des tests de liaison bivariés (*Non-contraint*, *Contraint* et *S_PC*) et univariés (LS, FN et PC1, PC2). Pour évaluer la distribution du test de liaison sous la contrainte (ρ_G=±1, *Non contraint*) nous calculons deux statistiques de liaison. Une sous la contrainte ou la corrélation génétique ρ_G est contrainte à +1 et une sous la contrainte où ρ_G = -1. La statistique du test *Contraint* est la plus grande des deux. Pour obtenir la statistique du test *S_PC*, nous calculons la somme des statistiques univariées de PC1 et PC2.

Nous montrons également les distributions des tests de liaisons univariés de LS et FN. Toutes les distributions empiriques des tests de liaison sont dérivées au marqueur D17S787, le premier point de la carte.

Analyse des données simulées

Les seuils empiriques de ces distributions correspondent aux quantiles à 90, 95 et 99%. Les erreurs de type 1 sont obtenues par comparaison avec les seuils tabulés aux seuils 10, 5, 1 et 0.5% pour des mélanges de χ^2. Nous montrons les intervalles de confiance des erreurs empiriques de type 1.

Pour comparer les résultats bivariés et univariés nous montrons seulement le meilleur niveau de signification de LS ou de FN (noté LS/FN) et de PC1 ou de PC2 (noté PC1/PC2). Les niveaux de signification empirique de chacune des analyses univariées de LS, FN, PC1 et PC2 sont calculés comme la proportion de réplicats pour lesquels la statistique du LOD score univarié excède la valeur observée dans nos données réelles. Les niveaux de significations empiriques montrés correspondent au meilleur résultat (plus faible niveau de signification) entre LS ou FN et PC1 ou PC2. Nous montrons également les niveaux de significations statistiques empiriques ajustés pour les tests multiples. Pour cela, nous prenons la meilleure des deux statistiques univariées de LS ou FN (t1) et de PC1 ou PC2 (t2) et nous calculons le nombre de réplicats pour lesquels la statistique du LOD score de LS/FN (PC1/PC2) excède t1 (t2).

Pour comparaison, nous montrons aussi les niveaux de signification du test *Non-contraint* et du test *Contraint* utilisées par défaut dans SOLAR (Almasy and Blangero 1998). Les niveaux de signification statistique calculés pour des mélanges asymptotiques de χ^2 à 0, 1 et 3 degrés de liberté $\left(\frac{1}{4}\chi_0^2+\frac{1}{2}\chi_1^2+\frac{1}{4}\chi_3^2\right)$ pour le test *Non-contraint* et des mélanges de χ^2 à 0, 1 et 2 degrés de liberté pour le test *Contraint*. Nous choisissons ces mélanges car ce sont ceux supposés par défaut dans SOLAR. Les distributions sont notées C et D dans le tableau 11. Pour le test *S_PC*, les seuils asymptotiques sont calculés pour des mélanges de χ^2 à 0, 1 et 2 degrés de liberté (tableau 11).

2.2.2. Résultats

Erreurs de type 1 des tests de liaison bivariés

Le tableau 15 montre les valeurs empiriques des statistiques des tests de liaison bivariés et les erreurs de type 1 à 10, 5, 1 et 0.5% du test *Non-contraint*, du test *Contraint* et du test *S_PC*. Similairement, dans la figure 8, nous montrons les distributions asymptotique et empirique de ces tests pour des niveaux de signification inférieure à 10%.

Concernant le test bivarié sans aucune contrainte sur les paramètres de liaison (*Non-contraint*), les quantiles empiriques à 90%, 95% et 99%, calculés dans nos données générées empiriquement, sont respectivement 4.10, 5.59 et 8.34. Les mélanges de distribution A et B montre des tendances inverses. Si l'on suppose un mélange de distribution A, les niveaux de signification empiriques sont inférieurs aux valeurs attendues sous l'hypothèse nulle. Au contraire, le mélange B donne des niveaux de signification plutôt au dessus des valeurs attendues. Utiliser des seuils asymptotiques conduira trop souvent au non rejet de l'hypothèse nulle H_0 pour un mélange A, le test est plutôt conservateur tandis que sous un mélange B, on ne rejettera pas assez l'hypothèse nulle, le test est donc plutôt libéral. Clairement, la distribution empirique est mieux prédite part le mélange C à 0, 1 et 3 degrés de liberté.

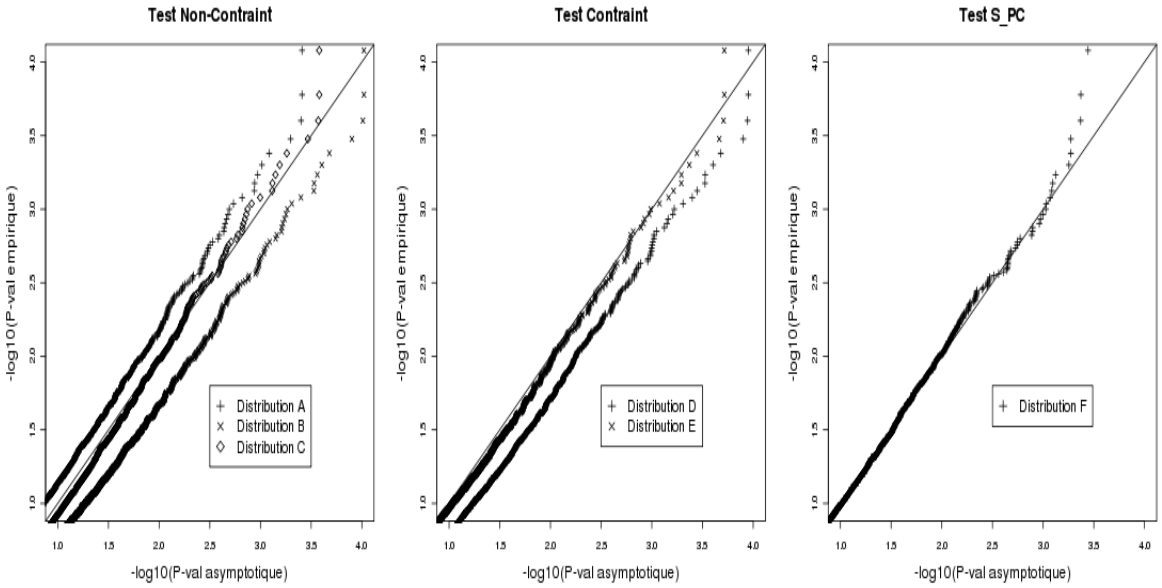
Lorsque l'on impose la corrélation génétique aux bornes de l'espace des paramètres (*Contraint*), les quantiles sont 3.95, 5.45 et 8.40. Les niveaux de signification supposant le mélange D sont toujours supérieurs à ceux calculés empiriquement dans nos données. Utiliser des seuils asymptotiques calculés pour le mélange D est plutôt libéral. Le mélange E prédit le mieux nos résultats de simulations empiriques.

Tableau 15 : Valeurs empiriques à 90, 95 et 99% des tests de liaison bivariés basés sur la méthode des composantes de la variance et erreurs de type 1 à 10, 5, 1 et 0.5%.

		Seuils empiriques			**Erreurs de type 1**			
Test		**90%**	**95%**	**99%**	**10%**	**5%**	**1%**	**0.5%**
Non-contraint	(A)	4.10	5.59	8.34	*0.073*	*0.036*	*0.007*	*0.003*
	(B)				0.171	0.094	0.021	0.010
	(C)				0.119	0.060	0.010	0.004
Contraint	(D)	3.95	5.45	8.40	0.166	0.088	0.019	0.009
	(E)				0.107	0.058	0.011	0.005
S_PC	(D)	3.01	4.24	7.15	0.103	0.051	0.009	0.004

Une inflation ou une baisse significative de l'erreur de type 1 (p-val≤0.05) sont respectivement souligné ou en italique. Un test standard de déviation d'une proportion à une valeur fixe a été utilisé.
Mélange de χ^2 à : (A) 1, 2 et 3 ddl ; (B) 0, 1 et 2 ddl ; (C) 0, 1 et 3 ddl ; (D) 0, 1 et 2 ddl ; (E) 1 et 2 ddl.

Figure 8 : Distributions asymptotiques et empiriques des tests de liaisons bivariés pour des niveaux de signification inférieure à 0.1



Dans les tableaux 16 et 17 nous montrons les niveaux de signification statistiques des meilleurs LOD scores identifiés par les méthodes bivariées ou univariées pour la variation de la DMO. Les niveaux de signification empiriques sont calculés sur 12 000 réplicats.

Comme attendu d'après les résultats de simulations, les niveaux de signification empiriques du test sans la contrainte (*Non-contraint*) sont très proches des niveaux de signification asymptotiques calculés pour des mélanges de χ^2 à 0, 1 et 3 degrés de liberté. Ce n'est plus le cas pour le test sous la contrainte ($\rho_G = \pm 1$, *Contraint*) où les significations sont calculées pour des mélanges de χ^2 à 0, 1 et 2 degrés de liberté (Tableau 16). Les niveaux de signification empiriques sont nettement moins significatifs que les niveaux asymptotiques. Comme attendu d'après nos résultats de simulation, pour le test de liaison reposant sur une ACP, les niveaux de signification empiriques sont similaires aux seuils asymptotiques (Tableau 17).

Pour les analyses univariées, nous calculons également le niveau de signification empirique ajusté pour le nombre de tests. Les niveaux de signification ajustés sont environ le double des niveaux calculés pour des seuils asymptotiques de χ^2 à 0 et 1 degré de liberté. Les niveaux de signification empiriques des tests univariés sont légèrement moins significatifs que les niveaux asymptotiques.

Revenons aux résultats dans notre étude. Sur neuf loci identifiés par les analyses bivariées ou univariées, sept régions sont communes aux deux approches. La force des signaux de liaison obtenus par des méthodes bivariés basées sur les traits LS/FN ou sur les axes principaux de variation (*S_PC*) est globalement similaire, les statistiques empiriques du test *S_PC* ont tendance à être légèrement moins significatives. Les régions localisées par les méthodes de liaison univariée ou bivariée sont peu différentes mais les niveaux de signification sont meilleurs par analyse univariée de LS/FN. Le meilleur niveau de signification est trouvé sur le 17q21-23 (p-value<10^{-4}).

Tableau 16 : Significations statistiques des régions de liaison identifiées par les tests bivariés *Non contraint* et *Contraint* ou par LS/FN

CHR	Marqueur	Test bivarié				Test univarié (LS/FN)*		
		Non-contraint		Contraint				
		Emp.+	Asymp.++	Emp.+	Asymp.++	Emp.[2]	Emp. glob.[3]	Asymp.[1]
1q42-43	D1S2800	$9.4x10^{-3}$	$8.9x10^{-3}$	$7.7x10^{-3}$	$4.3x10^{-3}$	$3.3x10^{-3}$	$4.9x10^{-3}$	$2.0x10^{-3}$
5p15	D5S1981	$3.2x10^{-3}$	$3.5x10^{-3}$	(0.032)	(0.016)	(0.023)	(0.032)	(0.014)
5q31-33	D5S410	0.016	0.015	0.019	$9.4x10^{-3}$	($6.5x10^{-3}$)	(0.020)	($8.2x10^{-3}$)
11q12-13	D11S4191	$1.3x10^{-3}$	$1.5x10^{-3}$	$1.0x10^{-3}$	$6.0x10^{-4}$	$0.2x10^{-3}$	$0.6x10^{-3}$	$0.3x10^{-3}$
12q23-24	D12S78	0.013	0.012	0.012	$6.6x10^{-3}$	$4.0x10^{-3}$	$5.8x10^{-3}$	$2.6x10^{-3}$
13q12-14	D13S218	$0.9x10^{-3}$	$1.3x10^{-3}$	$1.2x10^{-3}$	$6.9x10^{-4}$	$0.3x10^{-3}$	$1.2x10^{-3}$	$0.4x10^{-3}$
17q21-23	D17S787	$<10^{-4}$	$9.7x10^{-5}$	$<10^{-4}$	$3.3x10^{-5}$	$<10^{-4}$	$<10^{-4}$	$0.2x10^{-4}$
21q22	D21S267	$5.3x10^{-3}$	$5.4x10^{-3}$	$4.7x10^{-3}$	$2.5x10^{-3}$	$2.0x10^{-3}$	$2.7x10^{-3}$	$1.0x10^{-3}$
22q11	D22S315	$1.7x10^{-3}$	$2.0x10^{-3}$	$1.3x10^{-3}$	$8.6x10^{-4}$	$0.2x10^{-3}$	$0.5x10^{-3}$	$0.2x10^{-3}$

(.) : LOD score <1.5.

+ : Niveaux de signification calculé empiriquement ; ++ : Signification pour un mélange de χ^2 à 0, 1 et 3 ddl pour le test *Non contraint*, un mélange de χ^2 à 0, 1 et 2 ddl pour le test *Contraint* et un mélange de χ^2 à 0, 1 et 2 ddl pour le test *S_PC*.

* : Meilleur niveau de signification entre LS et FN

1 : Meilleur niveau de signification entre LS et FN calculés pour un mélange asymptotique de χ^2 à 0 et 1 ddl ; 2 : Meilleur niveau de signification empirique entre LS et FN ; 3 : Niveau de signification empirique ajusté pour le nombre de test.

Tableau 17: Significations statistiques des régions de liaison identifiées par le test *Non contraint* pour le test bivarié *S_PC* ou par PC1/PC2

CHR	Marqueur	Test bivarié : S_PC		Test univarié (PC1/PC2)*		
		Emp+	Asymp.$^{++}$	Emp.^{2}	Emp. glob.^{3}	Asymp.^{1}
1q42-43	D1S2800	(0.075)	(0.072)	(0.045)	(0.076)	(0.036)
5p15	D5S1981	$2.8x10^{-3}$	$3.0x10^{-3}$	($5.5x10^{-3}$)	(0.015)	($7.0x10^{-3}$)
5q31-33	D5S410	(0.024)	(0.025)	(0.019)	(0.030)	(0.014)
11q12-13	D11S4191	$3.3x10^{-3}$	$3.7x10^{-3}$	$1.6x10^{-3}$	$2.4x10^{-3}$	$1.2x10^{-3}$
12q23-24	D12S78	$5.5x10^{-3}$	$6.4x10^{-3}$	$2.5x10^{-3}$	$4.2x10^{-3}$	$2.4x10^{-3}$
13q12-14	D13S218	$1.6x10^{-3}$	$1.7x10^{-3}$	$1.6x10^{-3}$	$2.4x10^{-3}$	$1.2x10^{-3}$
17q21-23	D17S787	$<10^{-4}$	$1.4x10^{-4}$	$<10^{-4}$	$<10^{-4}$	$0.4x10^{-4}$
21q22	D21S267	$3.3x10^{-3}$	$3.7x10^{-3}$	$1.6x10^{-3}$	$2.4x10^{-3}$	$1.2x10^{-3}$
22q11	D22S315	$2.8x10^{-3}$	$2.8x10^{-3}$	$1.6x10^{-3}$	$2.4x10^{-3}$	$1.1x10^{-3}$

Nous avons également comparé les distributions asymptotiques des tests de liaison bivariés *Non contraint* et *Contraint* à des distributions de loi Gamma comme suggéré par (Dalmasso, Pickrell et al. 2007). Les valeurs des paramètres de forme (« shape parameter ») et d'intensité (« rate parameter ») sont tirées des moyennes et des variances des distributions des statistiques empiriques. Pour le test *Non contraint*, la moyenne (variance) de la distribution des statistiques est 1.49 (3.85). Le paramètre de forme est 0.58 et le paramètre d'intensité 0.39. De même pour le test *Contraint*, les paramètres de forme et d'intensité sont respectivement 0.58 et 0.40.

Dans la figure 9 nous montrons, en haut, l'histogramme de la statistique empirique et les distributions asymptotiques pour les mélanges de distribution A ($1/4\,\chi_1^2 + 1/2\,\chi_2^2 + 1/4\,\chi_3^2$), B ($1/4\,\chi_0^2 + 1/2\,\chi_1^2 + 1/4\,\chi_2^2$) et C ($1/4\,\chi_0^2 + 1/2\,\chi_1^2 + 1/4\,\chi_3^2$) du test *Non-contraint*. Sur les deux figures du bas, nous montrons les résultats pour le mélange C (c'est-à-dire celui qui correspond le mieux à la distribution générée empiriquement dans nos données) et pour une distribution Gamma de moyenne 1.49 et de variance 3.85. Nous voyons que les distributions asymptotiques de ces deux distributions sont très similaires. De même pour le test *Contraint*, les distributions E ($1/2\,\chi_1^2 + 1/2\,\chi_2^2$) et gamma sont très proches (Figure 10). A partir des moyennes et variances calculées empiriquement dans nos données, les mélanges de distribution des tests de liaison supposant une loi Gamma sont très proche des mélanges de χ^2 C et D.

Figure 9 : Histogrammes de la statistique du test *Non-contraint* sous l'hypothèse nulle pour des distributions asymptotiques pour le mélange A, B et C et pour une distribution Gamma

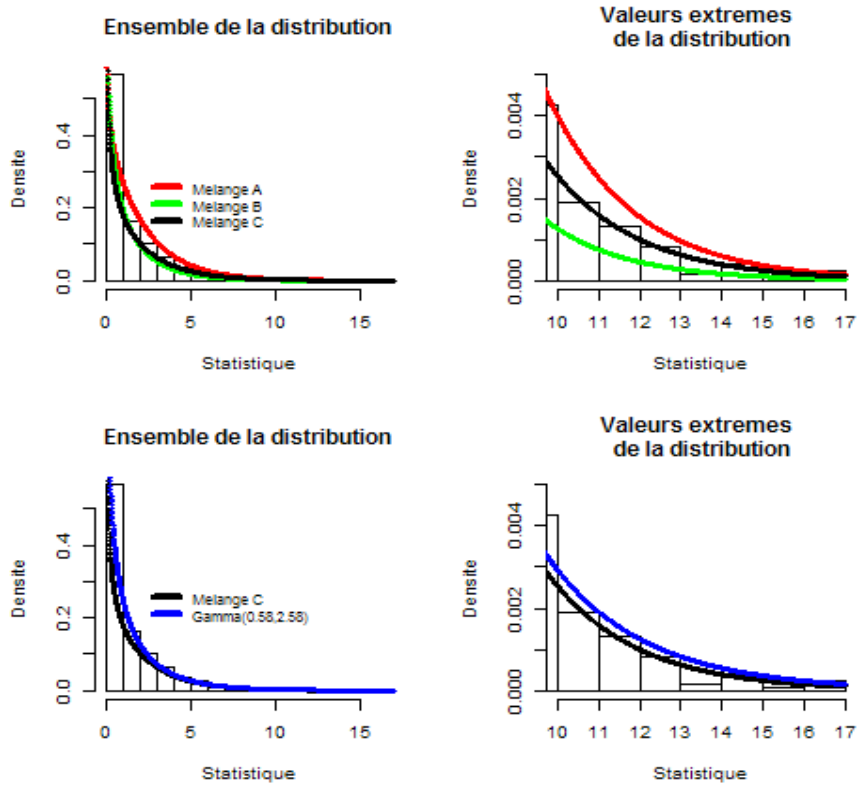
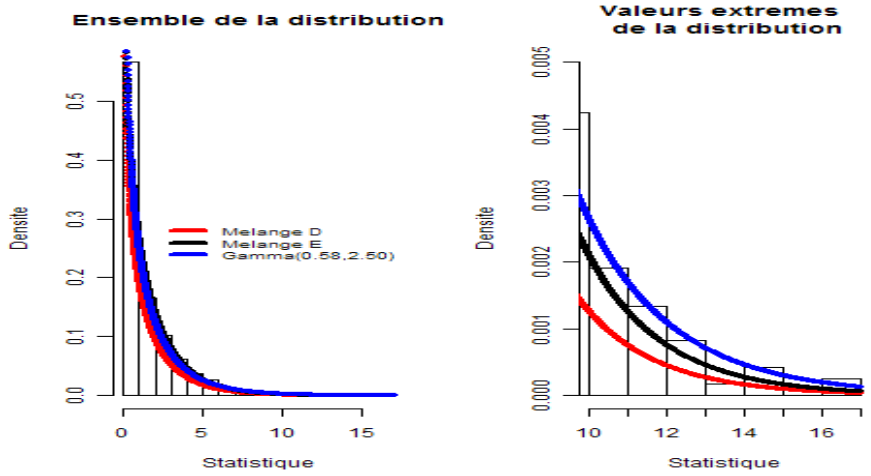


Figure 10 : Histogrammes de la statistique sous l'hypothèse nulle du test *Contraint* et distributions asymptotique pour le mélange D, E et pour une distribution Gamma



2.3. Conclusions de l'étude de liaison dans les données NEMO

L'analyse multivariée peut théoriquement améliorer la puissance de détection des QTLs à effets pléiotropiques. La littérature montre de nombreuses applications de cette approche dans les études de liaison (Tang, Xiao et al. 2007; Liu, Liu et al. 2008; Pan, Xiao et al. 2008; Wang, Deng et al. 2008; Zhang, Lei et al. 2009).

Nos résultats dans les données NEMO, montrent que le modèle général des composantes de la variance et le test basé sur une analyse factorielle (*S_PC*) donnent des résultats globalement similaires. Étant donné qu'il n'y a pas d'ambiguïté sur la distribution asymptotique du test de liaison utilisant des techniques de réduction des données, cela favorise l'application de ce test pour la recherche de QTLs à effets pléiotropiques.

Certains auteurs ont montré que l'analyse jointe de traits corrélés pouvait augmenter la puissance de détecter des QTLs (Jiang and Zeng 1995; Allison, Thiel et al. 1998; Amos, de Andrade et al. 2001), mais que ce gain de puissance est fortement dépendant du degré et de la direction de la corrélation résiduelle entre les traits. En effet, le plus grand gain de puissance est obtenu quand la corrélation induite par les effets génétiques et environnementaux est de sens opposé.

Dans de tels cas, la corrélation phénotypique résultante peut être faible. Concernant l'application aux données NEMO, nous ne sommes pas dans la situation la plus favorable pour observer un gain de puissance de l'analyse jointe vis-à-vis d'une analyse univariée. La corrélation phénotypique des traits est relativement grande (≈ 0.65) et les corrélations du QTL, polygénique et environnementale sont toutes positives. Ceci est confirmé par nos analyses des données réelles, la force des signaux de liaison est en général plus élevée par analyse univariée que par analyse bivariée. Deux loci sur le chromosome 5 sont détectés uniquement par les approches bivariées et pourrait suggérer un QTL à effets pléiotropiques. On peut remarquer que dans ces régions, les résultats des approches utilisant une ACP ou l'approche sous la contrainte donnent des résultats complémentaires, chacune de ces régions étant seulement identifiée par un seul de ces deux tests de liaison.

Ces résultats semblent également être confirmés par d'autres études pour la recherche de liaison dans le cadre de l'ostéoporose et de ses traits associés. En effet, la plupart de ces auteurs ne trouvent pas de nette amélioration de la force du signal de liaison en analyse jointe relativement à des analyses univariées (Devoto, Spotila et al. 2005; Karasik, Dupuis et al. 2007; Wang, Liu et al. 2007). Il faut noter que l'évaluation des niveaux de signification est souvent basée sur des valeurs nominales ce qui rend difficile la comparaison entre études.

En raison de l'augmentation du nombre de degrés de liberté du test de liaison bivarié ainsi que du temps de calcul nécessaire pour évaluer les seuils empiriquement (15min/réplicat dans nos données NEMO), la détection de QTLs à effets pléiotropiques pour la

covariation de la DMO à LS et FN est souvent basée sur la consistance des résultats obtenus entre les régions de liaison identifiées par analyses univariées des traits originaux et du premier axe de variation (expliquant la plus grande part de la variance totale) d'une analyse en composantes principales (Karasik, Cupples et al. 2004; Tan, Liu et al. 2008). Cette démarche ignore le problème des tests multiples et n'est pas un test formel de QTL à effets pléiotropiques comme celui proposé par (Mangin, Thoquet et al. 1998).

Trois distributions asymptotiques différentes des tests de liaisons bivariés ont été proposées. Ceci nous a amené à évaluer par simulations les distributions empiriques de ces tests de liaison sous l'hypothèse nulle. Cette étude de simulation pose le problème de l'interprétation des niveaux de signification statistique des tests joints basés sur la méthode des composantes de la variance et calculés à partir de valeurs nominales. Certaines des distributions ne rejettent pas assez l'hypothèse nulle de non liaison et sont donc conservatives alors que d'autres semblent trop libérales.

Relativement à notre étude, supposant les mélanges C et D pour le test *Non-contraint* et *Contraint*, les résultats obtenus par les deux papiers de Amos et de Wang sont discordants (Amos, de Andrade et al. 2001; Wang 2003) alors que les données sont générées selon des modèles génétiques similaires. Nos résultats de simulations sur le test *Non-contraint* sont consistants avec les résultats de Amos (mélange C, Tableau 11) mais pas avec ceux de Wang (mélange A). Pour le test *Contraint*, nos résultats sont consistants avec les résultats de Wang (mélange E) mais pas avec ceux de Amos (mélange D). A titre d'illustration, nous reportons les résultats dans la figure 28 en annexe.

Amos trouve des niveaux de puissance globalement similaires entre le test *Non-contraint* et le test *Contraint* en utilisant des seuils asymptotiques et empiriques. On peut remarquer que pour le test *Contraint*, utiliser des seuils basés sur des mélanges de χ^2 à 0, 1 et 2 degrés de liberté au lieu d'un mélange à 1 et 2 degrés de liberté augmente également le nombre de faux positifs.

Tout ceci suggère de prendre des précautions quant à l'interprétation des puissances relatives des tests de liaison univariés et bivariés basés sur la méthode VC. Nous n'avons pas évalué la puissance de ces tests et de futures investigations seraient nécessaires pour comparer les performances relatives de ces tests de liaison bivariés dans nos données NEMO.

Chapitre 3

3. Recherche de QTLs par analyse d'association bivariée

Une autre approche pour détecter des loci de traits quantitatifs est d'utiliser l'information apportée par le marqueur sur l'association. Comme on l'a vu, dans une étude de liaison, on cherche à voir si les allèles de loci adjacents se transmettent de façon non indépendante au cours des générations à l'intérieur des familles. Plus deux loci sont génétiquement liés, plus ils sont situés proches l'un de l'autre, ce qui diminue les chances de recombinaison ($\theta \approx 0$). La distance physique entre des loci proches est de l'ordre de quelques mégabases (Mb).

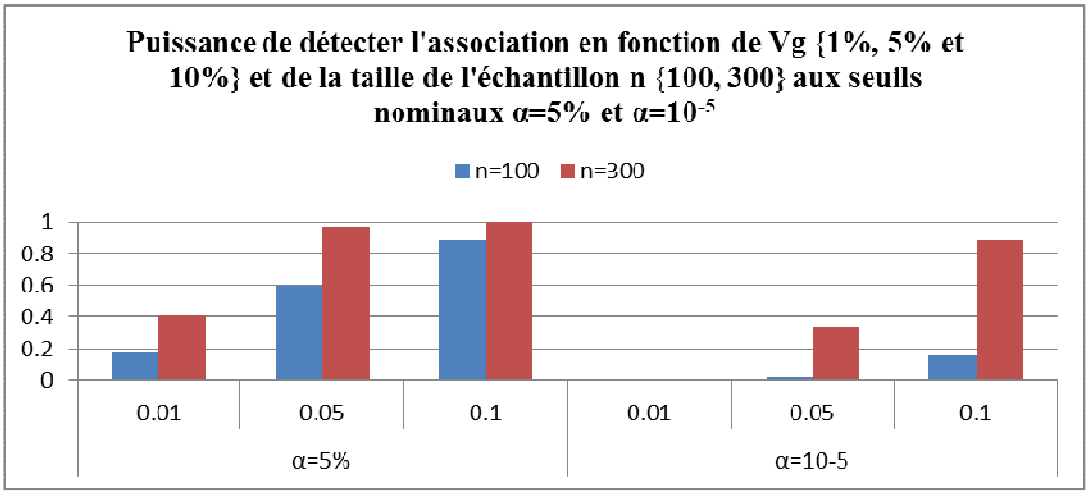
Les études d'association reposent sur l'hypothèse que la mutation a eu lieu depuis déjà plusieurs générations. Au cours des générations successives, la recombinaison va atténuer le déséquilibre de liaison (DL) entre le variant causal et les loci adjacents. Seulement des petites régions du génome, de l'ordre de quelque kilobases (Kb), seront transmises ensemble de génération en génération autour de cette mutation (Cardon and Bell 2001).

Les études d'association peuvent être réalisées dans des échantillons de sujets non apparentés ou apparentés. L'utilisation de chacun de ces échantillons a ses avantages et ses inconvénients. Les études de sujets non apparentés sont plus sensibles à la stratification de population. L'association observée entre les allèles ne résulte alors pas forcément du déséquilibre de liaison mais de la stratification de population. Plusieurs méthodes pour corriger ou contrôler la stratification ont été développées. Ces méthodes sont détaillées au paragraphe 3.1. La plupart des criblages du génome pour l'association, conduits ces dernières années, ont utilisé des échantillons de sujets non apparentés. Avec ce type d'échantillonnage, il est plus facile de recruter un large échantillon d'individus nécessaires pour détecter des variants causaux dont les effets sur la variation du trait sont modestes, mais cela implique de réaliser préalablement à l'analyse d'association des procédures de contrôle qualité des données.

Les approches utilisant des données familiales peuvent se libérer du problème du mélange de population. Les méthodes de type TDT pour des traits quantitatifs (Fulker, Cherny et al. 1999) sont robustes à la stratification car les scores génotypiques sont conditionnés sur les génotypes parentaux. Le principe consiste à décomposer le score génotypique en deux composantes orthogonales, une composante entre-famille et une composante intra-famille. La composante entre-famille prend en compte le phénomène de stratification tandis que la composante intra-famille est significative, seulement s'il existe un déséquilibre de liaison. Nous discutons plus en détails des méthodes d'association dans des données familiales au chapitre 4.

D'une manière générale, la puissance des analyses d'association dépend des effets du QTL, de la taille de l'échantillon et du déséquilibre de liaison entre le variant causal et le marqueur. Nous représentons dans la figure 11 la puissance de détecter le variant causal, en association directe avec le marqueur aux seuils nominaux $\alpha=0.05$ et $\alpha=10^{-5}$. Nous avons supposé un modèle de régression linéaire dans lequel la seule variable explicative est le QTL. Ce QTL explique 1%, 5% et 10% de la variabilité phénotypique totale. Nous faisons varier également la taille de l'échantillon (100 vs 300). Plus la part de variance expliquée par le QTL augmente, plus la puissance augmente.

Figure 11 : Puissance de détecter l'association entre le trait et le variant causal en fonction de la variance génétique (%) et de la taille d'échantillon



Si l'association entre le marqueur et le trait est indirecte, c'est-à-dire si le marqueur n'est pas le polymorphisme causal, la puissance de détecter le variant causal diminue avec la force du déséquilibre de liaison entre le marqueur et le variant causal.

3.1. Introduction

Les études d'association à grande échelles à partir de puces commercialisées, sont devenues abordables, mais il est nécessaire de procéder, préalablement à l'analyse d'association, à différentes étapes de contrôle qualité pour diminuer l'impact des facteurs de confusion. Ces procédures se font au niveau du génotypage et des ADN. Comme nous l'avons mentionné, l'existence d'une association entre le phénotype et le marqueur peut non seulement être due à l'existence d'un déséquilibe de liaison entre le marqueur et le variant génétique causal situé à proximité du marqueur, mais aussi à des mécanismes de stratification de population c'est-à-dire des mélanges de populations ayant des fréquences différentes des allèles. Il est donc aussi nécessaire de vérifier l'homogénéité génétique de la population.

Les données sur les SNPs sont nettoyées en fonction du nombre de données manquantes et pour des variants dont la fréquence est trop faible. Les fréquences alléliques dans la population doivent respecter l'équilibre de Hardy-Weinberg pour éviter de détecter à tort une association avec le QTL. De même, l'ensemble des données d'un individu est supprimé si le taux de données manquantes est trop important. Si deux individus ont des données identiques, un des individus de la paire est enlevé. Pour des études de sujets non apparentés, il est également nécessaire de vérifier l'indépendance des sujets. Les procédures de contrôle qualité vont donc dépendre des critères d'inclusion choisis.

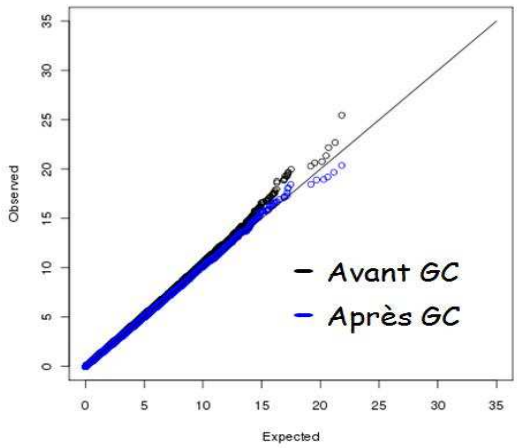
Le dernier point important est de vérifier si les individus sont génétiquement homogènes, c'est-à-dire qu'il n'existe pas des groupes de sujets ayant des différences systématiques dans les fréquences alléliques.

Plusieurs méthodes pour corriger ou pour contrôler la stratification ont été développées. Les plus connues sont la méthode du « genomic control » (GC) (Devlin and Roeder 1999), la méthode de Patterson (Patterson, Price et al. 2006) basée sur l'analyse en composantes principales de la matrice des génotypes ou encore la méthode implémentée dans le logiciel STRUCTURE (Pritchard, Stephens et al. 2000) utilisant des approches bayésiennes pour estimer les proportions du génome, de chaque individu, provenant de populations d'origines différentes. Cette méthode nécessite beaucoup de temps de calcul.

La méthode de correction GC est relativement simple. La méthode repose sur l'hypothèse que seulement une proportion très faible de SNPs sont des « vrais » QTLs par rapport au nombre total de SNPs testés par criblage. La distribution de la statistique obtenue dans les données suite à un criblage est donc proche de celle attendue sous l'hypothèse nulle de non association. Le principe consiste à comparer la distribution observée à la distribution théorique. Devlin (Devlin and Roeder 1999) proposent de corriger uniformément les valeurs des statistiques par un facteur λ, où λ est le rapport de la médiane de la statistique observée sur celle attendue sous l'hypothèse nulle (égale à 0.456 pour un χ^2 à 1 degré de liberté). Plus λ est supérieur à 1, plus la probabilité de détecter à tort une association entre

le trait et le SNP augmente. La procédure graphique consiste à tracer le graphe des quantiles attendus contre les quantiles observés (« QQ plot »). Une illustration en est donnée dans la figure 12 pour un facteur λ=1.14. Les points noirs sont les valeurs des statistiques du criblage du génome et les points bleus sont les statistiques après correction par le facteur λ.

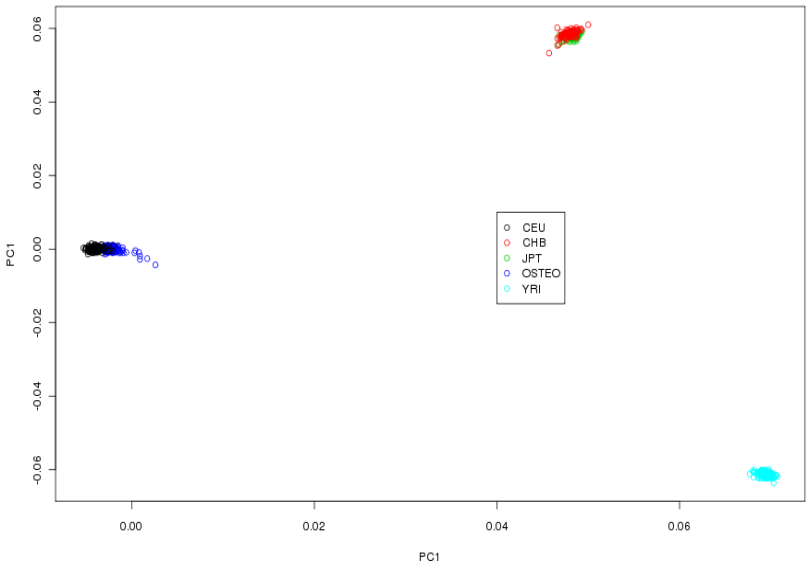
Figure 12 : QQ plot des statistiques de test avant et après correction par λ (GC)



La méthode basée sur une analyse en composantes principales proposée dans EIGENSTRAT (Patterson, Price et al. 2006) permet de contrôler pour la présence de stratification de population. Le principe consiste à décomposer la matrice génotypique en axes de variations orthogonaux telles que la variance du nuage de points autour du premier axe soit maximale, cela revient à choisir cet axe de telle sorte qu'il explique la plus grande part de la variance génétique. Les individus ayant des valeurs génotypiques aberrantes (« outliers ») ne sont pas pris en compte. A partir des axes de variation, il est possible de rechercher les « causes » par exemple en regardant les corrélations entre les covariables et les axes principaux de variation. Afin de contrôler la stratification, les axes de variation peuvent être pris en compte dans le modèle d'association comme covariables.

La détection de stratification de population peut être évaluée au niveau inter-continent par exemple en couplant les données avec l'ensemble des populations de la base Hap Map (http://hapmap.ncbi.nlm.nih.gov/). Il est ainsi possible de regrouper les individus selon le continent d'origine. Nous montrons dans la Figure 13 un exemple dans notre échantillon de caucasiens (données OSTEO issues de la base NEMO). On retrouve notre échantillon dans le groupe des individus de la base Hap Map d'origine européenne (CEU). On distingue clairement deux autres groupes, celui d'origine africaine et celui d'origine asiatique ou chinoise (JPT et CHB).

Figure 13 : Mélange des populations de Hap Map (CEU en bleu et noir, CHB, JPT en rouge et YRI en bleu clair)

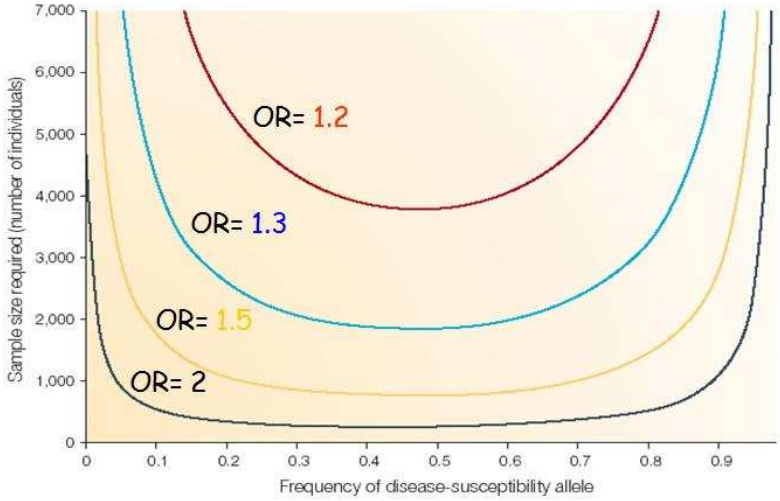


Un des problèmes spécifiques aux analyses à grande échelle est le problème des tests multiples car on répète le même test statistique pour chaque SNP. Par exemple, pour 500 000 SNPs au seuil nominal α de 5% (χ_1^2=3.84), on s'attend à trouver 25 000 faux positifs. Une des approches classiques est d'utiliser une correction de Bonferroni. Le principe consiste à diviser le seuil α de chacun des tests par le nombre total de tests. Sur le même exemple, il faudrait déclarer un SNP significatif au seuil 10^{-7} (χ_1^2=28.36) pour conserver un niveau de signification global α de 5%. La méthode de correction de Bonferroni est conservative car les tests ne sont pas complètements indépendants en raison du déséquilibre de liaison entre les SNPs, en particulier lorsqu'on utilise des cartes de marqueurs denses. D'autres approches de correction des niveaux de signification pour les tests multiples ont été proposées comme le False Discovery Rate (FDR, (Benjamini and Hochberg 1995)). Pour des seuils de signification ajustés pour les tests multiples, seul des variants à effets forts peuvent générer des statistiques elevées dépassant les critères stricts de détection, ce qui est peu probable pour des traits multifactoriels.

Comme nous l'avons montré dans la figure 11 pour un trait quantitatif, nous donnons une illustration dans le cas d'une maladie dans la figure 14. Nous montrons les tailles d'échantillons nécessaires pour détecter l'association entre la maladie et le variant causal avec une puissance de 80% au seuil $\alpha=10^{-7}$ en fonction de la taille de l'effet (OR=1.2 à 2) et de la fréquence de l'allèle (tiré de (Wang, Barratt et al. 2005)). Plus la taille de l'effet (OR) diminue, plus le nombre d'individus nécessaires pour obtenir une puissance de 80%

augmente. De même, plus un des allèles devient rare, plus le nombre d'individus nécessaires pour conserver une puissance de 80% augmente. Par exemple, pour une fréquence de 0.3 et un OR de 1.3, environ 2 000 individus sont nécessaires alors qu'il en faut plus de 4 000 pour une fréquence de 0.1.

Figure 14 : Taille d'échantillon nécessaire pour détecter l'association entre la maladie et le variant causal avec une puissance de 80% au seuil $\alpha=10^{-7}$ en fonction de la taille de l'effet (OR= 1.2, 1.3, 1.5 et 2) et de la fréquence du variant causal (tiré de (Wang, Barratt et al. 2005))



Une variété d'approches existe pour augmenter la puissance statistique de détecter des variants causaux pour des traits quantitatifs complexes.

Plusieurs études ont montré qu'analyser des échantillons d'individus sélectionnés pour des valeurs extrêmes pouvait être plus puissant qu'analyser des échantillons sélectionnés aléatoirement dans la population (Allison 1997; Allison, Thiel et al. 1998; Abecasis, Cookson et al. 2001). Une approche complémentaire pour augmenter la puissance est d'analyser des traits corrélés par des méthodes multivariées. En effet, l'analyse jointe peut théoriquement augmenter la puissance de détection de QTL (Allison, Thiel et al. 1998; Amos, de Andrade et al. 2001). Un autre avantage est qu'il n'est pas nécessaire de corriger pour le nombre de test résultant des analyses effectuées sur chacun des traits séparément. L'analyse jointe peut donc améliorer la capacité de détection de QTL dont les effets sont trop faibles pour être détectés par des analyses univariées (Jiang and Zeng 1995). Plusieurs approches pour l'analyse multivariée de traits corrélés ont été appliquées à des études de liaison (Tableau 4) mais très peu à des études d'association. Diverses méthodes d'analyses d'association ont récemment été proposées pour l'analyse de traits corrélés. Ces méthodes sont le plus souvent basées sur des équations linéaires généralisées (GEE : Generalized Estimating Equations ; (Liang and Zeger 1986)) dans le cadre de données de population (individus non apparentés) ou dans le cadre de données

familiales (Lange, Silverman et al. 2003; Lange, van Steen et al. 2004; Jung, Zhong et al. 2008; Liu, Pei et al. 2009; Pei, Zhang et al. 2009; Yang, Tang et al. 2009; Zhang, Bonham et al. 2009; Zhang, Pei et al. 2009).

Deux papiers (Liu, Pei et al. 2009; Yang, Tang et al. 2009) ont étudié les performances des tests d'association dans des données de population. L'un d'eux a appliqué la méthode GEE qui suppose que les effets du QTL sur les traits sont les mêmes. Une telle contrainte sur les effets du QTL peut surestimer la puissance relative de détecter l'association avec le locus causal vis-à-vis des analyses univariées. La deuxième étude combine des équations linéaires généralisées par le modèle SUR (Seemingly Unrelated Regression, (Zellner 1962)). Ce modèle à l'avantage de pouvoir supposer des effets génétiques différents sur les traits.

Les criblages du génome pour l'association utilisant des traits multivariés sont relativement rares, particulièrement dans des échantillons d'individus sélectionnés aux extrêmes de la distribution phénotypique. Nous avons donc appliqué une analyse d'association jointe basée sur le modèle SUR pour la variation de la DMO aux deux sites squelettiques LS et FN. Les résultats sont montrés dans le paragraphe 3.2. Le gain apporté par l'analyse jointe des traits corrélés vis-à-vis des analyses univariées nous a amené a évaluer les propriétés statistiques des analyses d'association jointes. Nous avons utilisé un modèle d'association bivarié permettant des effets génétiques distincts sur les traits (modèle SUR). Les analyses ont été conduites pour différentes valeurs des effets génétiques et correlation résiduelle entre les traits. De même, nous avons fait varier les schémas d'échantillonnage des individus. Les résultats sont montrés dans le paragraphe 0.

3.2. Criblage du génome de la DMO pour des individus non-apparentés

3.2.1. Matériel et méthodes

3.2.1.1. Les données

Le phénotype utilisé est la valeur du Z-score qui représente la densité osseuse ajustée pour l'âge et le sexe (paragraphe 1.5). L'échantillon NEMO a été étendu au recrutement d'hommes non apparentés sélectionnés pour un Z-score $\leq$ -2. Le reste des individus étaient également des hommes fortement sélectionnés pour leur valeur au site LS (Z-score $\leq$ -2) ou faiblement sélectionnés pour leurs valeurs aux sites LS et FN (Z-score > 0.5). Le schéma de sélection revient à sélectionner les individus pour des valeurs

basses dans les 2.3% de la queue de distribution à gauche et pour des valeurs hautes dans les 30.9% de la queue de distribution à droite.

Caractéristiques de l'échantillon

Les caractéristiques de l'échantillon sont décrites dans le tableau 18. Les corrélations entre les phénotypes et les covariables sont montrées dans le tableau 19. Au total, 330 individus âgés entre 19 et 70 ans étaient disponibles dont 175 et 155 étaient sélectionnés pour des valeurs basses et hautes respectivement. Dans l'échantillon des individus sélectionnés pour des valeurs basses à LS, la valeur de Z-LS est bien plus faible que celle de FN (-2.81 vs -1.20). Au contraire, dans l'échantillon des individus sélectionnés pour des valeurs hautes (Z-score > 0.5), les valeurs aux deux traits sont relativement similaires. La moyenne globale est ainsi plus faible pour Z-LS (-0.80±2.26) que pour FN (- 0.04±1.42). Il en résulte que la corrélation des traits dans l'échantillon total (0.87) est nettement plus forte que celle des deux sous groupes. De même les individus étant seulement sélectionnés sur LS dans le groupe des individus sélectionnés pour des valeurs basses, la corrélation est plus forte dans ce groupe (0.37) relativement au groupe des individus sélectionnés pour des valeurs hautes (0.18). Ces tendances résultent de notre schéma de sélection des individus.

Les valeurs moyennes de l'IMC sont similaires dans les trois groupes. La figure 15 montre les histogrammes des distributions du Z-score chez les individus sélectionnés aux extrêmes gauche (TG) ou droite (TD). Les histogrammes de la distribution des covariables âge et IMC sont montrés en annexe dans la figure 29.

Les corrélations des traits avec la covariable IMC sont faibles et égales à 0.34 et 0.38 pour Z-LS et FN respectivement. La corrélation est plus forte dans le groupe des individus sélectionnés à gauche seulement pour Z-FN et l'IMC (0.29 contre 0.12). Les histogrammes de la distribution de l'IMC et Z-LS et de l'IMC et Z-FN par groupe d'individus sont montrés dans la figure 30 en annexe.

L'effet de l'âge est significatif pour la variation de Z-LS dans le groupe des individus sélectionnés à droite (Z-score > 0.5 ; $p<10^{-4}$). L'effet de l'IMC est significatif pour la variation de Z-FN dans les groupe des individus sélectionnés à gauche (Z-score ≤ -2 ; $p<10^{-3}$).

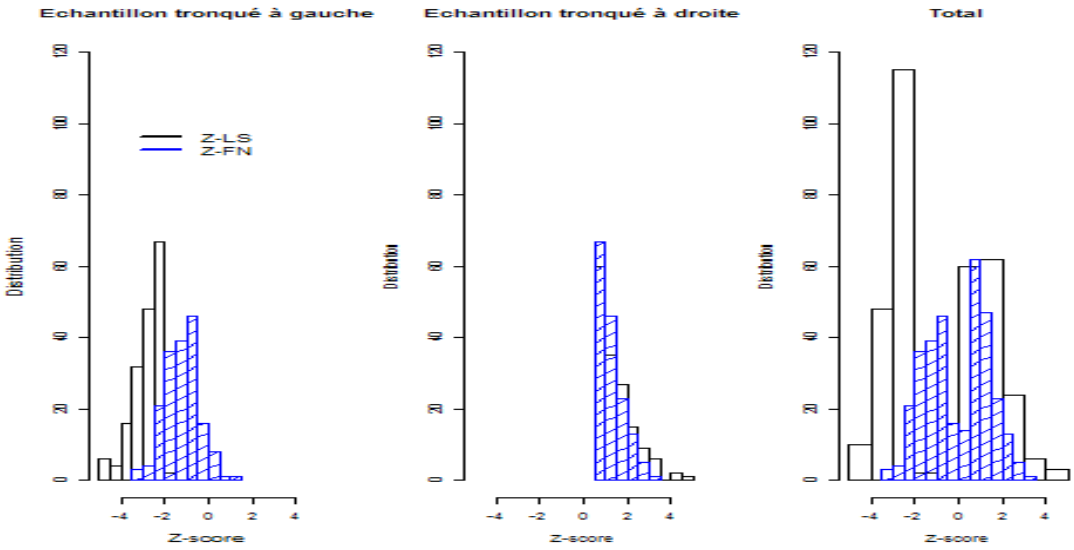
Tableau 18 : Distribution (Moyenne± écart-type [min;max]) du Z-score et des covariables dans les différents groupes d'individus.

	Total	**Z-score ≤ -2**	**Z-score > 0.5**
N	330	175	155
Age (année)	43.5 ± 12.9	46.9 ± 11.8	39.7 ± 13.1
min; max	[19;70]	[19;70]	[20;67]
IMC (kg/cm^2)	25.4 ± 3.6	24.3 ± 3.3	26.6 ± 3.4
min; max	[14.81;38.80]	[14.8;34.5]	[20.3;38.8]
Z-score			
LS	-0.80 ± 2.26	-2.81 ± 0.66	1.48 ± 0.80
min; max	[-5.00;4.99]	[-5.00; -1.87]	[0.55;4.99]
FN	-0.04 ± 1.42	-1.20 ± 0.77	1.27 ± 0.61
min; max	[-3.20;3.48]	[-3.20;1.08]	[0.50;3.48]

Tableau 19 : Corrélations des Z-score et des covariables dans les différents groupes d'individus.

	Total	**Z-score ≤ -2**	**Z-score > 0.5**
Cor. (Z-LS, Z-FN)	0.87	0.37	0.18
Cor. (Z-LS, IMC)	0.34	0.10	0.10
Cor. (Z-FN, IMC)	0.38	0.29	0.12

Figure 15 : Histogrammes des distributions du Z-score dans le groupe des individus sélectionnés pour des valeurs basses ou hautes et dans l'échantillon total.



Analyse de contrôle qualité des données génotypiques

Le criblage du génome pour l'association a été conduit sur la plateforme Illumina 370K. Nous avons utilisé le logiciel PLINK (Purcell, Neale et al. 2007) pour nettoyer les données de génotypage. Les procédures de qualité contrôle des données génotypiques ont été appliquées sur les SNPs et sur les individus.

Pour l'ensemble des individus, le SNP est exclu si :

- son pourcentage de données manquantes est > 2.5% (n=16 629)
- la signification de la statistique du test de Hardy-Weinberg est inférieur à $p<10^{-5}$ (n=413)
- la fréquence de l'allèle mineur est < 5% (n=17 788)

Un total de 30 098 SNPs a été supprimé. Au final, 298 783 SNPs étaient inclus dans les analyses.

Pour l'ensemble des SNPs, un individu est enlevé si :

- son pourcentage de données manquantes est > 4% (n=3)
- l'estimation du taux d'allèles identiques par descendance (IBD) entre paire d'individus dépasse 0.14, alors un des individus de la paire est enlevé (n=5). Un seuil de 0.125 correspond à une relation d'apparenté du troisième degré. Les 5 individus enlevés avaient une proportion au dessus de 0.43.
- ses coordonnées sur les premières composantes principales tel que calculées par EIGENSTRAT (Patterson, Price et al. 2006) dépasse 6 écarts-types de la valeur moyenne (n=9).

Au bilan, sur 330 individus, 313 avaient une valeur de Z-score au rachis lombaire (LS) et au col du fémur (FN) et ont été retenu dans l'étude.

Les principales caractéristiques des données après le contrôle qualité des données sont décrites dans le

tableau 20. L'échantillon d'analyse conprenait 313 individus, dont 167 et 146 étaient sélectionnés pour des valeurs basses ou hautes du Z-score.

Tableau 20 : Caractéristiques (Moyenne ±écart-type [min; max]) du Z-score après le contrôle qualité dans les différents groupes d'individus

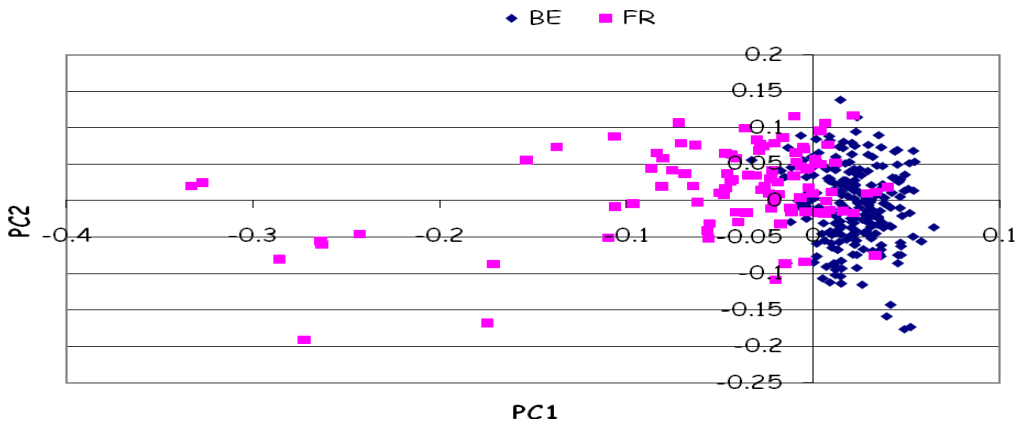
Z-score	**Total** (n=313)	**Z-score ≤ -2** (n=167)	**Z-score > 0.5** (n=146)
LS	-0.80±2.26 [-5.00; 4.99]	-2.80 ±0.64 [-5.00;-1.87]	1.48±0.82 [0.55; 4.99]
FN	-0.04±1.43 [-3.20; 3.48]	-1.19 ±0.78 [-3.20; 1.08]	1.29±0.62 [0.50; 3.48]

Analyse de stratification

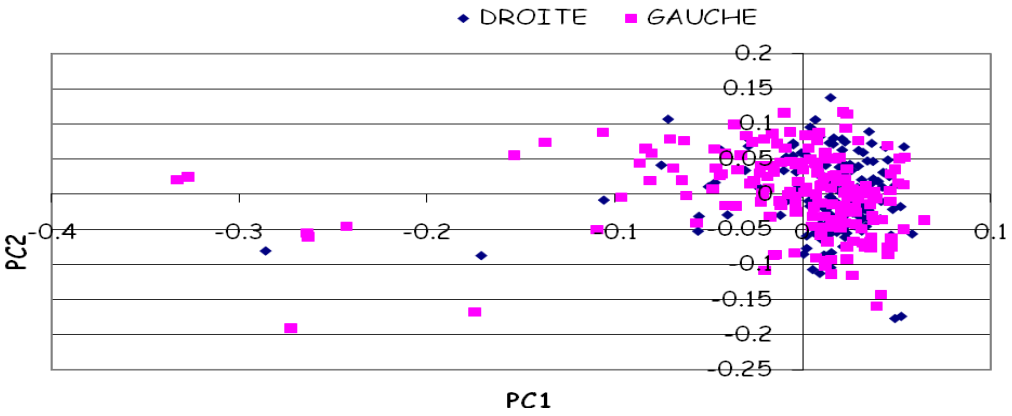
L'analyse en composantes principales a identifié seulement un cluster après suppression des individus extrêmes. Nous montrons les projections des individus sur les deux premiers axes de variation selon le pays d'origine de l'individu (Figure 16.1) et selon leur critère de sélection (Figure 16.2).

Figure 16 : Projection des individus sur les deux premières composantes principales

(1) Projection selon leur pays (Belgique ou France)



(2) Projection selon le critère de sélection des individus pour des valeurs hautes (droite) ou basses (gauche) de la DMO.



3.2.1.2. Méthode d'association bivariée : le modèle SUR

Ce modèle a été proposé par Zellner (Zellner 1962). Pour simplifier, nous nous plaçons dans le cas de deux traits quantitatifs corrélés Y1 et Y2 expliqués par les variables X_1 et X_2. Supposons N individus indépendants, chacun ayant 2 observations, le système d'équations s'écrit :

$$\begin{cases} y_1 = \beta_{10} + X_1 \times \beta_1 + e_1 \\ y_2 = \beta_{20} + X_2 \times \beta_2 + e_2 \end{cases}$$

Il peut être réécrit sous la forme suivante :

$$y = \beta_0 + X \times \beta + e = \beta_0 + \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \times \beta + e$$

où $y = (y_1, y_2)'$ est le vecteur de dimension 2 N x 1 des observations ; X est une matrice diagonale par blocs de dimension 2 N x 2 codant les variables explicatives. $\beta = (\beta_1, \beta_2)'$ est le vecteur de dimension 2x1 des paramètres de régression et $e = (e_1, e_2)'$ est le vecteur de dimension 2 N x 1 des erreurs résiduelles.

Les erreurs résiduelles suivent une distribution normale bivariée de moyenne 0 et de matrice de variance/covariance de la forme suivante :

$$\Omega = \Sigma \otimes Id_N = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \otimes Id_N$$

où σ_1^2 et σ_2^2 sont respectivement les variances résiduelles sur Y1 et Y2, et ρ la corrélation résiduelle entre traits.

La qualité de l'ajustement du système est mesurée par le R^2 de McElroy's. R^2 est la proportion de variance expliquée par la variable X qui tient compte de la matrice de covariance résiduelle Ω (McElroy 1977) :

$$R^2 = 1 - \frac{\hat{e}'\hat{\Omega}^{-1}\hat{e}}{y'\left(\hat{\Sigma}^{-1} \otimes Id_N\right)y}$$

Sous l'hypothèse nulle, il n'y a pas d'association entre les traits et la covariable X (i.e., $\beta_1 = \beta_2 = 0$) tandis que sous l'hypothèse alternative, X a un effet sur au moins un des traits Y1 ou Y2 ($\beta_1 \neq 0$ et/ou $\beta_2 \neq 0$). Les restrictions sur le paramètre β sous H_0 sont écrites sous la forme $R\beta = 0$ où chaque restriction linéairement indépendante est représentée par une ligne de R. Dans notre cas, la matrice R prend la forme suivante :

$$R = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \text{ tel que } R\beta = 0 \leftrightarrow \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_{10} \\ \beta_1 \\ \beta_{20} \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Le test d'association est basé sur un test de Wald et peut être approximé pour des échantillons finis par une distribution de Fisher :

$$F = \frac{\left(R\hat{\beta}\right)'\left(R\,\mathrm{cov}\left(\hat{\beta}\right)R'\right)^{-1}\left(R\hat{\beta}\right)}{2}$$

Sous l'hypothèse nulle, la statistique suit asymptotiquement une loi de Fisher à 2 et 2x(N-2) degrés de liberté.

Le modèle SUR est en général estimé par une variante de la méthode des moindres carrés généralisés (MCG). La matrice de covariance résiduelle entre traits étant inconnue, la première étape consiste à l'estimer à partir de la méthode des moindres carrés ordinaires (MCO). Elle est de la forme $\hat{\Omega} = \hat{\Sigma} \otimes Id_N$. L'estimateur de β est alors calculé par la formule des MCG :

$$\hat{\beta} = \left(X'\left(\hat{\Sigma}^{-1} \otimes Id_N\right)X\right)^{-1} X'\left(\hat{\Sigma}^{-1} \otimes Id_N\right)y$$

Le test univarié est basé sur une simple régression linéaire (paragraphe 1.3). Le modèle d'association entre le trait Y_j et le marqueur est $y_j = \beta_{j0} + X_j \times \beta_j + e_j$ où β_j est le coefficient de régression de l'effet du marqueur.

Sous l'hypothèse nulle de non association entre le trait et le marqueur (β_j=0), le test suit une statistique de Student à N-2 degrés de liberté.

Remarque :

Bien que le modèle SUR permette d'introduire des variables explicatives différentes entre équations, nous nous plaçons dans le cas très simple où ces variables sont identiques pour les deux équations de régression $(X_1 = X_2)$. Il s'agit du génotype au marqueur pour un même individu. Dans cette situation, les paramètres du modèle estimés par la méthode des MCG donnent les mêmes estimations que la méthode des MCO. Dans les modèles univariés ou bivariés, les estimations des paramètres de régression sont donc identiques. Le gain apporté par la prise en compte de la matrice de covariance résiduelle intervient seulement dans le test d'association.

Le criblage du génome pour l'association a été fait avec R sous un simple modèle de régression linéaire dans le cas des analyses univariées de Z-LS et Z-FN et sous le modèle bivarié SUR implémenté dans le package systemfit.

3.2.2. Résultats

La moyenne de la statistique de Fisher du test basé sur le modèle SUR pour le criblage du génome est de 1.018 (écart-type=1.022, médiane=0.70). Les moyennes des statistiques de Student univariées de Z-LS et Z-FN sont respectivement -0.0167 (écart-type=1.011, médiane=-0.0165) et -0.0129 (écart-type=1.006, médiane=0.0104). Ces résultats indiquent qu'il n'y a pas d'inflation significative des statistiques bivariées et univariées. Les graphes quantiles contre quantiles des analyses bivariées et univariées sont montrés dans la Figure 17. Les résultats d'association sur l'ensemble du génome sont montrés dans la Figure 18.

Figure 17 : Graphes quantiles contre quantiles des tests d'association bivariés basé sur le modèle SUR (A) et univariées de Z-LS et FN (B).

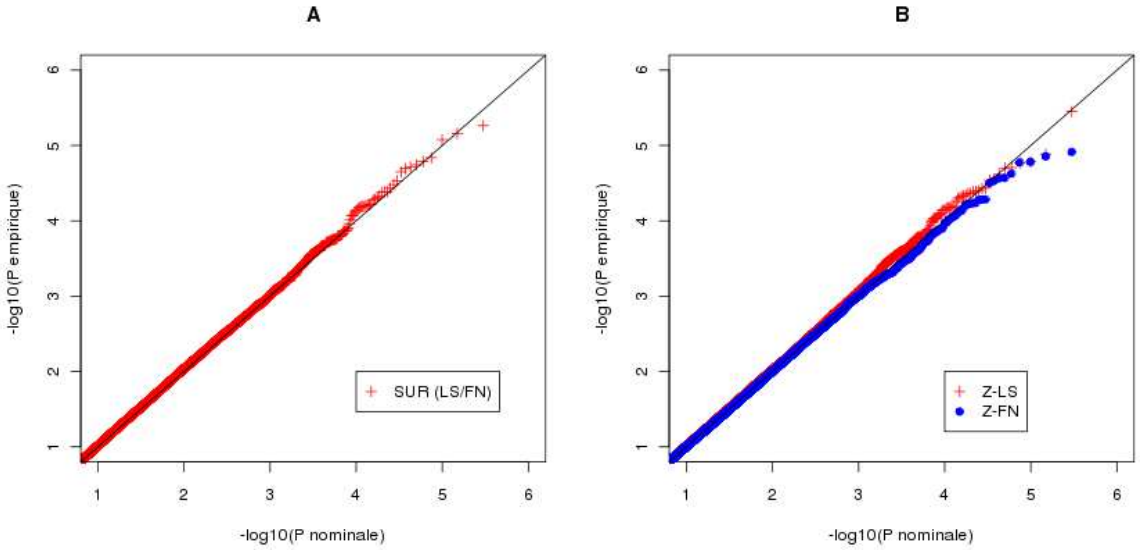
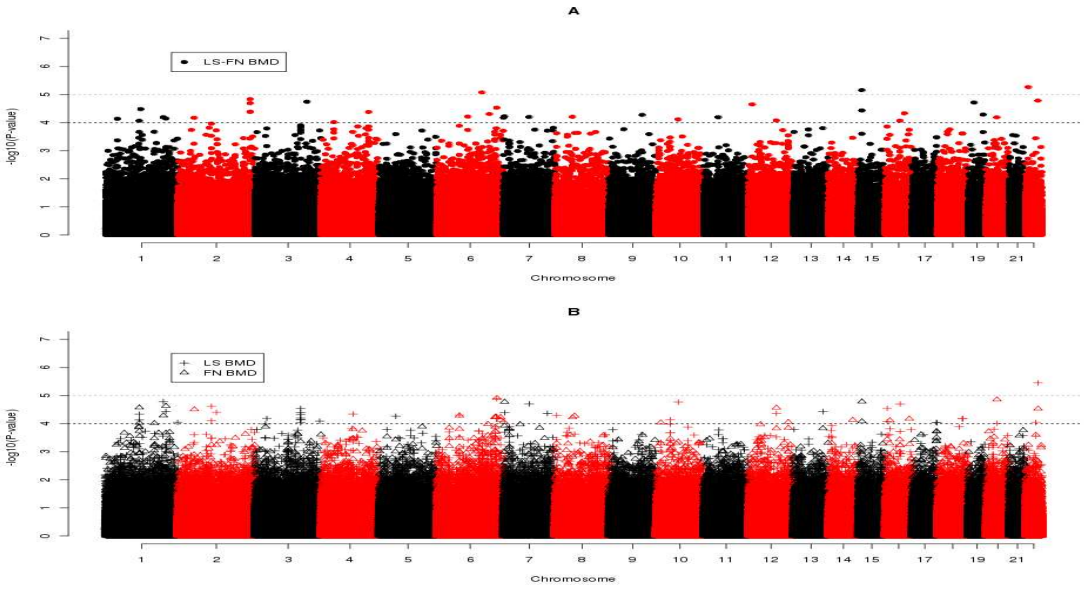


Figure 18 : Graphes Manhattan des résultats d'association sur l'ensemble du génome pour les 298 783 SNPs des analyses bivariées (A) et univariées (B).

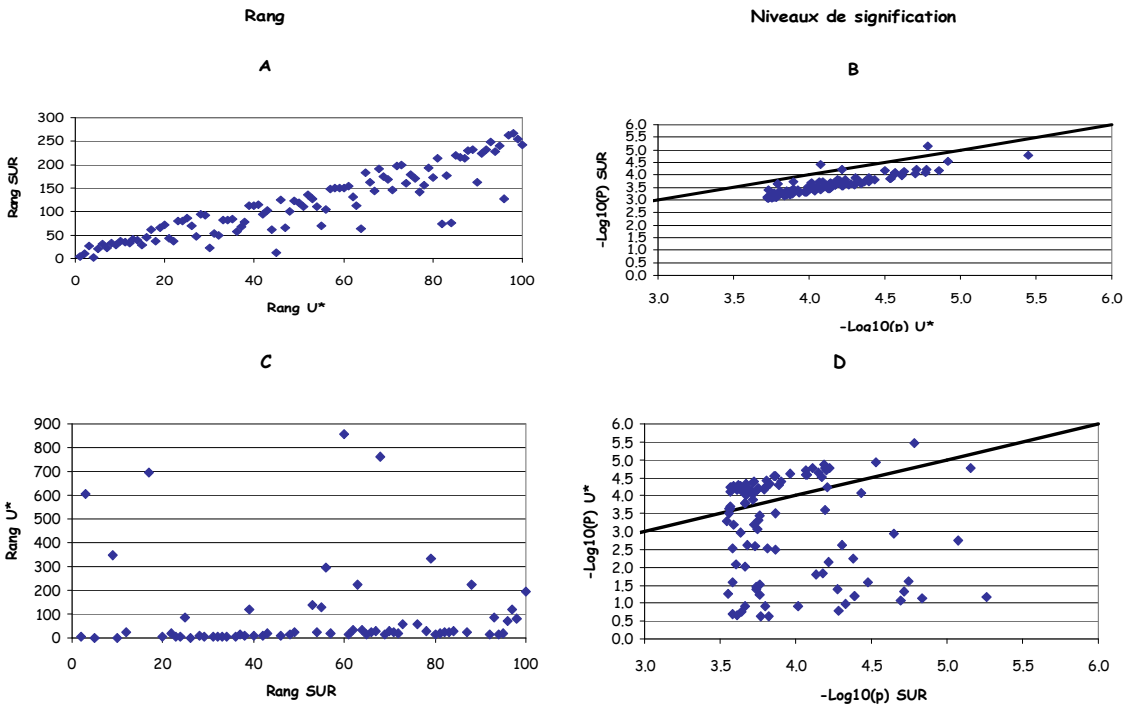


Trente cinq SNPs avec une forte évidence d'association (p-value<10^{-4}) ont été identifiés par analyse bivariée basée sur le modèle SUR. De manière intéressante, plusieurs de ces SNPs n'atteignent pas le seuil nominal de 5% pour l'un ou l'autre des tests univariés de Z-LS ou Z-FN.

Pour comparer les résultats d'association entre ces approches bivariées (SUR) et univariées, nous évaluons leur niveaux de concordance pour un ensemble de SNPs parmi les plus significatifs dans les deux approches (Figure 19).

Pour les analyses univariées, nous prenons le meilleur résultat d'association entre Z-LS et Z-FN c'est-à-dire le niveau de signification le plus petit des deux, non ajusté pour les tests multiples, noté U*. Dans les figures 19 A et C, nous comparons les résultats en termes de rang et de manière similaire, dans B et D, nous comparons les résultats en termes de niveaux de signification. Les figures 19 A et B représentent les résultats des 100 SNPs les plus associés (significatifs) pour U* contre son résultat en analyse bivariée tandis que les figures C et D correspondent au 100 meilleurs résultats identifiés par la méthode bivariée (SUR).

Figure 19 : Rangs et niveaux de signification des analyses univariées (U*) et bivariées (SUR) pour les 100 SNPs les plus associés.



Les 100 meilleurs résultats de U* (Figure 19 A et B) se retrouvent parmi les 268 meilleurs résultats d'association identifiés par la méthode SUR. Les rangs des deux approches sont bien corrélés entre eux (≈ 0.9). De façon similaire, les niveaux de signification de SUR sont très proches de ceux de U* et montrent des bons niveaux de concordance. Les résultats bivariés sont légèrement moins significatifs que ceux des résultats univariés (Figure 19 B) mais sont bien corrélés.

Au contraire, dans l'ensemble des meilleurs SNPs de SUR (Figure 19 C et D), on observe une grande disparité dans les résultats obtenus par les les analyses univariées. Parmi les 100 meilleurs SNPs identifiés par le modèle SUR, 49 sont egalement detectées ($\leq 2.4 \times 10^{-4}$) en analyse univariées. Pour ces SNPs, les significations bivariées étaient seulement légèrement plus faibles (Figure 19 D). Pour l'autre motié des SNPs restant, le meilleur signal est trouvé par analyse jointe. Des nettes différences sont observées pour certains SNPs. Dans environ 30% (17 sur 49) des cas restant, les SNPs ne sont pas detectés (>5%) par analyse univariée, quelque soit le trait LS ou FN.
Les mêmes tendances sont observées si l'on s'intéresse aux 300 meilleurs résultats identifiés par les deux approches. En effet, si l'on prend les 300 meilleurs signaux d'association identifiés par analyses univariées (U*), les rangs des analyses bivariées sont inférieur à 766 (moyenne des rangs=313). Inversement, lorsque que l'on prend les 300 meilleurs résultats identifiés par l'approche SUR, les rangs des analyses univariées sont inférieur à 80 682 (moyenne des rangs=7 062).

Ces résultats montrent que certains signaux d'association détectés par l'approche SUR ne sont pas identifiés par les analyses univariées. Ils suggèrent donc que les analyses bivariées peuvent avoir un intérêt pour l'étude d'association de traits corrélés.

Le résumé des analyses SUR est montré dans le tableau 21 où l'on montre les meilleurs résultats (p-value $\leq 10^{-5}$) obtenus par analyse jointe LS et FN. Ces 17 SNPs identifient 13 régions génomiques distinctes associées avec l'un ou l'autre des traits LS ou FN. Les trois SNPs les plus associés sont localisés dans les régions 6q22 ($p=8.42 \times 10^{-6}$), 15q14 ($p=6.97 \times 10^{-6}$) et 22q11 ($p=5.44 \times 10^{-6}$). Nous montrons également dans le tableau 21 les niveaux de signification des analyses d'association univariées de Z-LS et FN. Si ce niveau de signification est inférieur au seuil nominal de 0.05, nous montrons également les rangs associés. Quatre SNPs situés dans trois régions chromosomiques sont bien classés (Rang <100) dans les analyses univariées de LS ou de FN. Ces SNPs sont situés dans les régions 6q25, 15q14 et 22q13. Les rangs sont égaux à 2 (LS, $p = 1.30 \times 10^{-5}$) et 1 (FN, $p = 1.22 \times 10^{-5}$) sur le 6q25 ; sur 15q14, les plus petit rangs sont observés pour FN : rang 3 ($p=1.65 \times 10^{-5}$) et 22 ($p = 8.38 \times 10^{-5}$) ; sur le 22q13, les rangs sont 1 (LS, $p= 3.54 \times 10^{-6}$) et 8 (FN, $p=2.95 \times 10^{-5}$). L'ensemble des 13 SNPs restant incluait deux des trois meilleurs résultats ($p<10^{-5}$ sur les régions 6p22 et 22q11) détectés par le modèle SUR et montrait un bien meilleur signal d'association par approche bivariée relativement aux approches univariées.
Les contributions génétiques, c'est-à-dire les parts expliquées par la régression (R^2), des SNPs les plus associés sont relativement faibles, ce qui n'est pas très surprenant pour des

traits complexes. Pour le résultat le plus significatif situé dans la région 22q11 ($p=5.44x10^{-6}$), les effets du QTL sur les traits sont en direction opposé. L'allèle mineur A est associé à une diminution et une augmentation du Z-score à LS et FN respectivement. La contribution du QTL dans cette région est relativement forte (3.85%). Il faut noter que cette valeur surestime la contribution génétique que l'on observerait dans des échantillons non sélectionnés.

Tableau 21 : Résultats pour les meilleurs résultats (p-val ≤ 10^{-5}) obtenus par analyse jointe de Z-LS et Z-FN

Information marqueur						Analyses bivariées (SUR)				Analyses univariées					
Chr	**Locus**	**Gène**	**SNP**	**A1 /A2**	**MAF⁺**	**p-val##**	**R^{2*}**	**β_{LS}**	**β_{FN}**	**p-val LS**	**Rang**	**R^2 LS**	**p-val FN**	**Rang**	**R^2 FN**
1	1p13		rs11578748	G/A	0.48	$3.30x10^{-5}$	3.26%	0	0.17	0.989	-	0.00%	$2.67x10^{-2}$	7930	1.57%
2	2q37	SP100	rs1678160	G/A	0.33	$1.45x10^{-5}$	3.52%	-0.05	0.15	0.574	-	0.10%	0.072	-	1.04%
2			rs1649866	A/G	0.33	$2.03x10^{-5}$	3.41%	-0.05	0.14	0.539	-	0.12%	0.086	-	0.95%
2			rs1649883	A/G	0.34	$4.12x10^{-5}$	3.19%	-0.03	0.15	0.686	-	0.05%	0.065	-	1.09%
2			rs1649891	T/C	0.34	$4.12x10^{-5}$	3.19%	-0.03	0.15	0.686	-	0.05%	0.065	-	1.09%
3	3q25	LEKR1	rs6799034	C/T	0.43	$1.81x10^{-5}$	3.45%	0	0.18	0.970	-	0.00%	$2.41x10^{-2}$	7166	1.63%
4	4q32		rs1523558	T/C	0.33	$4.18x10^{-5}$	3.20%	-0.23	-0.06	$5.58x10^{-3}$	1759	2.45%	0.497	-	0.15%
6	6q22		rs2049924	A/G	0.29	**$8.42x10^{-6}$**	3.69%	0.27	0.08	$1.80x10^{-3}$	607	3.09%	0.363	-	0.27%
6	6q23		rs9321496	G/T	0.48	$4.91x10^{-5}$	3.14%	0.08	0.24	0.295	-	0.35%	$2.35x10^{-3}$	695	2.94%
6	6q25	TIAM2	rs998318	G/T	0.31	$2.94x10^{-5}$	3.30%	0.38	0.38	$1.30x10^{-5}$	**2**	5.94%	$1.22x10^{-5}$	**1**	5.98%
12	12p13		rs1017301	C/A	0.34	$2.24x10^{-5}$	3.38%	-0.11	-0.28	0.223	-	0.48%	$1.13x10^{-3}$	348	3.36%
15	15q14	RYR3	rs2437143	C/T	0.38	**$6.97x10^{-6}$**	3.75%	-0.22	-0.35	$8.41x10^{-3}$	2635	2.21%	$1.65x10^{-5}$	**3**	5.80%
15			rs4780133	G/A	0.34	$3.69x10^{-5}$	3.23%	-0.2	-0.33	$1.98x10^{-2}$	6186	1.73%	$8.38x10^{-5}$	**22**	4.86%
16	16q23		rs6564175	C/T	0.42	$4.67x10^{-5}$	3.16%	-0.14	0.06	0.108	-	0.83%	0.521	-	0.13%
19	19p13	FAM125A	rs2303680	G/A	0.41	$1.92x10^{-5}$	3.43%	-0.03	0.17	0.743	-	0.03%	$4.66x10^{-2}$	13886	1.27%
22	22q11	SLC2A11	rs2275979	A/G	0.18	**$5.44x10^{-6}$**	3.85%	-0.07	0.19	0.523	-	0.13%	0.067	-	1.08%
22	22q13	LL22NC03-75B3.6	rs3935378	T/C	0.49	$1.64x10^{-5}$	3.48%	0.35	0.32	$3.54x10^{-6}$	**1**	6.69%	$2.95x10^{-5}$	**8**	5.47%

#: Position en cM (Build 36.3).
+: Fréquence de l'allèle mineur. A1 est l'allèle mineur.
##: p-val est la p-value de la statistique du test de Fisher à 2 et 2x(N-2) ddl.
*: R^2 du système complet prenant en compte la matrice de covariance résiduelle.
§: Les p-values sont rangées dans l'ordre croissant et les rangs sont reportées

3.3. Performances du test d'association bivarié SUR

Dans cette partie nous étudions les propriétés statistiques du test d'association bivarié basé sur le modèle SUR et nous comparons les performances vis-à-vis d'une analyse univariée traditionnelle dans des échantillons d'individus non apparentés. Nous avons conduit des simulations extensives pour différentes valeurs de paramètres et selon plusieurs schémas de sélection des individus.

3.3.1.1. Modèles de simulations

Nous nous intéressons à l'analyse de deux traits quantitatifs corrélés (Y1 et Y2) en fonction de la contribution génétique du QTL. Pour cela, nous faisons varier la taille de l'effet ainsi que la direction des effets induits par le QTL sur chacun des traits.

Nous considérons des modèles génétiques de traits complexes selon des données générées, mimant au mieux nos données réelles concernant l'étude de la DMO. Le locus causal contribuant à une faible proportion des variances phénotypiques, la corrélation résiduelle entre traits approxime la corrélation phénotypique (paragraphe 1.4). En conséquence, nous générons des données pour lesquelles la corrélation résiduelle entre traits est toujours positive.

Afin d'évaluer les performances des tests dans des scénarios ou la direction des effets induits par le QTL et par la composante résiduelle soient en sens opposés, nous faisons varier le signe de la corrélation induite par le QTL à travers les effets du génotype sur les traits. Les scénarios varient en fonction du signe induit par la corrélation due au QTL et selon la force de la corrélation résiduelle.

Nous considérons deux échantillons de taille N=300 et N=1 000 constitués de sujets non apparentés. Nous évaluons deux niveaux de corrélation phénotypique résiduelle : modéré (ρ=0.2) et forte (ρ=0.6). Nous supposons un QTL bi-allélique et notons q et p la fréquence de l'allèle mineur et majeur respectivement. Nous supposons que la valeur du trait Y_j diminue avec le nombre d'allèles rares.

Pour chaque individu i, son génotype g_i (g_i = 0, 1 ou 2 sous un modèle additif ; i=1,…, N) au marqueur est simulé à partir d'une distribution binomiale de paramètre q. Le vecteur phénotypique Y_i pour chaque individu i, $i = 1,...,N$ est obtenu en générant des échantillons à partir d'une distribution normale bivariée, c'est-à-dire
$Y_i \sim \mathrm{N}\left(\left(E\left(a_1 g_i\right), E\left(a_2 g_i\right)\right), \Sigma\right)$ où a_j, $j = 1,2$ est l'effet additif génotypique sur le trait Y_j avec $E\left(a_j g_i\right), j = 1,2$ les moyennes génotypiques sur le $j^{ème}$ phénotype et $\Sigma = \left(\sigma_{ij}^2\right)$

est la matrice (2x2) de variance phénotypique. Les moyennes génotypiques sur le trait Y_j sont données par :

$$E\left(a_j g_i\right)=\begin{cases}-2pa_j & si\ g_i=0\\ (q-p)a_j & si\ g_i=1\\ 2qa_j & si\ g_i=2\end{cases}$$

La force de l'effet additif relativement à la variance phénotypique est exprimée par l'héritabilité du QTL h_j^2 et elle est définie par la proportion de variance phénotypique expliquée par la variation du QTL sur le trait Y_j, c'est-à-dire $h_j^2=\mathrm{var}\left(a_j g\right)\big/\mathrm{var}\left(Y_j\right)$. On peut en déduire l'expression analytique de h_j^2 de la forme $h_j^2=2q(1-q)\times a_j^2$.

Différentes combinaisons des fréquences de l'allèle mineur (q=0.1 vs 0.4) et des effets additifs sont choisies pour obtenir des tailles d'effets (h^2) de 0%, 0.5%, 1% et 3%. Nous faisons varier le signe de a_j (positif ou négatif). Les effets induits par le QTL sur Y1 et Y2 sont soit de même signe, soit de signe opposé, induisant une corrélation génétique $\left(\rho_G\right)$ égale à +1 ou – 1 respectivement. Si le QTL n'affecte qu'un seul des traits, alors $\rho_G=0$. Les trois principaux modèles génétiques étudiés sont fonction de la contribution du QTL sur les traits Y_j :

- Le SNP n'a pas d'effet sur Y1 et Y2 $\left(h_1^2=h_2^2=0\right)$.
- Le QTL exerce des effets sur Y1 seulement $\left(h_1^2>0;\ h_2^2=0\right)$ et donc $\rho_G=0$.
- Le QTL exerce des effets pléiotropiques sur les deux traits $\left(h_1^2>0;\ h_2^2>0\right)$. Pour ce scénario, deux cas sont considérés :

  (1) Le QTL affecte similairement Y1 et Y2 $\left(h_1^2=h_2^2\ et\ \rho_G=1\right)$ (2) le QTL affecte différemment Y1 et Y2 $\left(h_1^2\neq h_2^2\ et\ \rho_G=\pm1\ ou\ h_1^2=h_2^2\ et\ \rho_G=-1\right)$.

Pour une combinaison donnée des valeurs des paramètres (ρ, h_1^2, h_2^2 et ρ_G = ±1), nous générons des échantillons de sujets non apparentés sélectionnés en fonction de leurs valeurs aux traits. Nous utilisons un seuil négatif comme critère de sélection à gauche (TG) et une valeur positive pour le critère de sélection à droite (TD). Les individus sont retenus si leur valeur de Z-score est inférieure à TG ou supérieur à TD. Deux ensembles de valeurs sont considérés pour TG et TD :

- TG=TD=0 i.e. les individus sont sélectionnés aléatoirement dans la population, noté S0.

- TG=-2 et TD=0.5 i.e. les individus sont fortement sélectionnés pour des valeurs extrêmes de Z-score au trait Y1 mais faiblement sélectionnés pour des valeurs élevées de Z-score au trait Y1, noté S1.

Afin de se rapprocher au mieux de notre schéma de sélection dans nos données réelles NEMO, nous avons étendu l'étude pour une situation dans laquelle la valeur au trait Y2 est également sélectionnée pour des valeurs élevées. Ce schéma de sélection est noté S2.

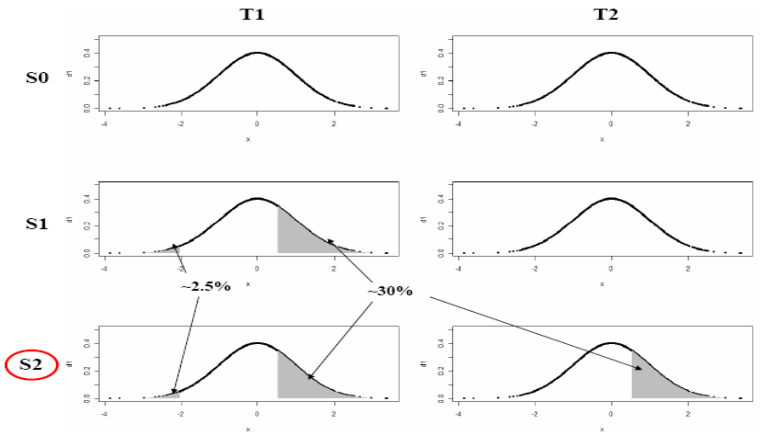
Les valeurs des critères de sélection choisies (TG=-2 et TD=0.5) signifient que les individus sont sélectionnés dans les 2.3% à gauche et 30.9% à droite des queues de distribution d'une loi normale. Ces régions apparaissent en grisées dans la Figure 20. Les individus satisfaisant aux critères de sélection sont gardés, les autres sont retirés de l'échantillon. Dans toutes les simulations, le nombre total de sujets est N et le nombre de sujets sélectionnés à gauche et à droite est égal à N/2, respectivement.

Chaque réplicat simulé est soumis à une analyse d'association bivariée (noté SUR) et à deux analyses d'association univariées notées U(Y1) et U(Y2). Les moyennes et les écarts-types de chaque statistique d'association aussi bien que les moyennes et écarts-types des paramètres d'association (coefficients de régression) sont calculés sur K réplicats.

La puissance et l'erreur de type 1 sont calculées comme les proportions de réplicats pour lesquels la statistique de test dépasse une valeur nominale (T_α). Cette valeur est dérivée d'une statistique de Fisher à 2 et 2x(N-2) degrés de liberté pour l'analyse bivariée et d'une statistique de Student à N-2 degrés de liberté pour l'analyse univariée. Les valeurs sont évaluées pour des seuils nominaux de 5%, 1%, 0.1% et 10^{-5}. L'erreur de type 1 est estimée sur 20 000 réplicats sous le modèle $h_1^2 = h_2^2 = 0$. Nous montrons également la probabilité de détecter l'association pour au moins un des traits Y1 ou Y2 noté U(Y1 ou Y2), c'est-à-dire, nous calculons le nombre de réplicats pour lesquels la statistique de Y1 ou de Y2 dépasse Tα. Sous l'hypothèse alternative, le SNP étudié est le variant causal et $h_1^2 > 0$, les puissances sont estimées pour 1 000 réplicats.

Pour comparer les résultats des analyses bivariées et univariées, les puissances et les erreurs empiriques de type 1 des analyses univariées ont été ajustées par une correction de Bonferroni (noté U_b). Nous calculons pour cela la proportion de réplicats pour lequel au moins une des deux statistiques de Y1 ou de Y2 dépasse la valeur seuil corrigée $T_\alpha/2$.

Figure 20 : Schémas de sélection des individus pour des traits Y1 et Y2 chacun issu d'une loi normale.



S0 : Les individus sont sélectionnés aléatoirement dans la population générale.
S1 : Les individus sont sélectionnés pour Y1 dans la queue de la distribution à gauche (2.3%) et à droite (30.9%) mais ne sont pas sélectionnés pour Y2.
S2 : Les individus sont sélectionnés pour Y1 dans la queue de la distribution à gauche (2.3%) et faiblement sélectionnés à droite (30.9%) pour Y1 et Y2.

3.3.1.2. Résultats

Erreurs de type 1

Les tableaux 22 et 23 listent les résultats des estimations sous l'hypothèse nulle. Le tableau 22 montre les statistiques pour N=1 000 et le tableau 23 donnent les erreurs de type 1 pour des seuils nominaux de 5% et 1% des tests bivariés et univariés en fonction de la fréquence de l'allèle (q=0.1 vs 0.4), de la corrélation résiduelle (ρ =0.2 vs 0.6) et de la taille de l'échantillon (N=300 vs 1 000).

Quand le SNP n'a pas d'effet sur les traits Y1 et Y2, les valeurs moyennes et écarts-types des statistiques des tests d'association bivariés et univariés sont proches des valeurs théoriques quelque soit la force de la corrélation résiduelle ρ, la fréquence de l'allèle q ou le schéma de sélection des individus (S0, S1 ou S2). De même, les erreurs de type 1 des tests d'association sont proches des valeurs nominales. Cependant, comme attendu, la probabilité de détecter au moins un des traits Y1 ou Y2 par une analyse univariée est approximativement deux fois plus grande que celle observée pour le test bivarié. Appliquer une correction de Bonferroni donne des niveaux de significations légèrement conservatifs, particulièrement lorsque la force de la corrélation résiduelle augmente. Dans la suite, les performances relatives des tests sont déduites à partir des valeurs nominales. Afin de comparer les niveaux de puissance pour des seuils de signification similaires,

nous utilisons pour les analyses univariées des niveaux de signification ajustés par une correction de Bonferroni (U_b).

Tableau 22 : Statistiques des tests d'association bivariés et univariés sous l'hypothèse nulle pour N=1 000 individus et selon le schéma de sélection S0, S1 ou S2

ρ	q	S0		S1		S2	
Analyse bivariée : μ-F (sd)							
0.2	0.1	1.00 (1.00)		1.00 (1.01)		1.00 (1.01)	
	0.4	1.00 (1.00)		1.01 (1.02)		1.02 (1.02)	
0.6	0.1	1.00 (1.01)		1.00 (0.99)		1.00 (1.01)	
	0.4	1.01 (1.01)		1.00 (1.00)		1.01 (1.00)	
Analyse univariée : μ-T (sd)							
		μ-Y1	**μ-Y2**	**μ-Y1**	**μ-Y2**	**μ-Y1**	**μ-Y2**
0.2	0.1	0.01 (1.00)	-0.01 (1.00)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)
	0.4	0.00 (0.99)	0.00 (1.00)	0.00 (1.00)	-0.01 (1.01)	0.00 (1.01)	0.00 (1.00)
0.6	0.1	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)	0.00 (1.00)	0.01 (1.00)	0.01 (1.00)
	0.4	0.00 (1.01)	-0.01 (0.99)	-0.01 (1.00)	-0.01 (1.00)	-0.01 (1.00)	0.00 (1.00)

*S0 : Les individus sont sélectionnés aléatoirement ; S1 : Les individus sont sélectionnés pour leur valeur au trait Y1 ; S2 : Les individus sont sélectionnés pour leurs valeurs aux traits Y1 et Y2.

Tableau 23 : Estimations des erreurs de type 1 pour un seuil nominal de 5% et 1%, des tests bivariés et univariés

		n=1 000*						n=300*					
		S0		S1		S2		S0		S1		S2	
		5%	1%	5%	1%	5%	1%	5%	1%	5%	1%	5%	1%
ρ	q	Analyses bivariées						Analyses bivariées					
0.2	0.1	5.05	1.02	4.99	0.97	4.96	1.05	5.30	0.98	4.81	0.99	5.17	1.07
	0.4	4.87	0.96	5.35	1.04	5.40	1.12	5.18	1.10	5.42	1.21	5.37	1.18
0.6	0.1	4.99	1.12	4.93	0.95	5.11	1.08	4.99	1.03	4.99	1.03	5.52	1.18
	0.4	5.10	1.07	5.06	1.00	5.30	1.00	5.24	1.17	4.98	1.05	5.24	1.24
		Analyses univariées						Analyses univariées					
		U(Y1)						U(Y1)					
0.2	0.1	5.14	1.01	4.94	1.08	4.92	0.96	5.14	0.99	4.78	0.97	5.02	1.09
	0.4	4.74	0.90	5.05	0.95	5.12	1.11	5.17	1.03	5.25	1.13	5.00	1.20
0.6	0.1	4.96	1.03	5.17	0.99	4.77	1.06	4.93	0.94	4.99	1.03	5.17	1.16
	0.4	5.15	1.00	4.97	0.95	5.13	0.96	5.09	1.05	4.78	0.91	4.88	0.96
		U(Y2)						U(Y2)					
0.2	0.1	4.81	0.98	5.01	1.11	4.94	1.04	5.23	1.03	4.85	0.99	4.99	0.96
	0.4	5.11	1.01	5.17	1.08	5.27	0.96	4.95	0.97	5.18	1.04	5.40	1.03
0.6	0.1	4.88	0.99	5.12	0.96	5.06	1.01	5.04	0.98	5.02	1.03	5.24	1.15
	0.4	4.95	1.09	5.05	0.98	5.02	0.99	4.97	0.99	4.83	0.97	4.98	1.03
		U(Y1 ou Y2)						U(Y1 ou Y2)					
0.2	0.1	9.58	1.93	9.27	2.10	8.01	1.70	10.07	2.04	9.06	1.91	8.32	1.74
	0.4	9.48	1.85	9.39	1.92	8.52	1.77	9.70	1.99	9.87	2.12	8.56	1.91
0.6	0.1	8.55	1.76	7.82	1.55	7.03	1.48	8.70	1.81	7.79	1.65	7.33	1.78
	0.4	8.82	1.88	7.68	1.55	7.15	1.47	8.75	1.90	7.44	1.51	7.17	1.54
		U(Y1 ou Y2)_b						U(Y1 ou Y2)_b					
0.2	0.1	4.91	0.99	4.92	0.98	4.13	0.97	4.93	1.01	4.44	0.95	4.27	0.80
	0.4	4.70	0.89	4.99	0.98	4.47	0.95	4.94	0.96	5.09	1.11	4.44	1.01
0.6	0.1	4.23	1.03	3.86	0.77	3.66	0.78	4.47	0.83	3.95	0.87	4.07	0.95
	0.4	4.44	0.86	3.83	0.79	3.68	0.74	4.60	1.02	3.96	0.82	3.75	0.81

*S0 : Les individus sont sélectionnés aléatoirement ; S1 : Les individus sont sélectionnés pour leur valeur au trait Y1 ; S2 : Les individus sont sélectionnés pour leurs valeurs aux traits Y1 et Y2.
$U(Y_j)$: Probabilité de détecter l'association avec le trait Y_j.
U(Y1 ou Y2) : Probabilité de détecter l'association avec au moins un des traits Y1 ou Y2 non ajustée
U(Y1 ou Y2)_b : Probabilité de détecter l'association avec au moins un des traits Y1 ou Y2 corrigée par une correction de Bonferroni.

Puissance

Les tableaux 24 et 25 (pages 121 et 122) montrent les moyennes (écarts-types) des statistiques d'association bivariées et univariées respectivement, en fonction des effets du QTL (h^2=0, 0.5%, 1% ou 3%), avec ou sans effets pléiotropiques ($\rho_G = 0$ ou $\rho_G = \pm 1$), de la force de la corrélation résiduelle ρ et du critère de sélection des individus sur un échantillon de 1 000 individus. La figure 21 montre les principales tendances des statistiques des tests bivariées et univariées pour différents échantillonnages des individus.

Statistiques d'association bivariées (Tableau 24)

Nous retrouvons des tendances déjà bien connues dans le cadre de l'analyse bivariée de traits quantitatifs corrélés. Quelque soit le modèle génétique, la moyenne μ-F augmente lorsque l'effet du QTL (h_1^2 / h_2^2) augmente indépendamment de ρ_G et de ρ.

De plus, lorsque les individus sont sélectionnés aléatoirement dans la population, la moyenne de la statistique est plus grande en présence d'un QTL à effets pléiotropiques ($\rho_G = \pm 1$) qu'en absence d'effets pléiotropiques ($\rho_G = 0$) et ceux d'autant plus si $\rho_G = -1$. Néanmoins, pour ρ=0.6, la statistique bivariée augmente relativement à la situation où $\rho_G = 0$ seulement si le QTL induit des effets pléiotropiques de sens opposés ($\rho_G = -1$). Les mêmes tendances sont observées pour le schéma de sélection S1.

Quelque soit le schéma de sélection des individus (S0, S1 ou S2), ces résultats confirment que la puissance dépend de la force de la corrélation résiduelle. Si les effets induits par le QTL sont de signes opposés ($\rho_G = -1$) ou lorsque le QTL a un effet sur Y1 seulement ($\rho_G = 0$), les moyennes des statistiques augmentent avec la corrélation résiduelle ρ. La tendance s'inverse dans la situation où les effets sont de même sens ($\rho_G = +1$), la moyenne μ-F diminue si la corrélation résiduelle augmente. Il en résulte que si la force de la corrélation résiduelle est élevée (ρ=0.6), les moyennes des statistiques bivariées μ-F sont plus élevée dans le scénario $\rho_G = 0$ que dans la situation $\rho_G = +1$ pour les schémas de sélection S0 et S1.

De manière générale, la sélection augmente la moyenne de la statistique mais le mode de sélection optimale des individus dépend du modèle génétique simulé et de la force de la corrélation résiduelle. Dans les modèles $\rho_G = 0$ et $\rho_G = -1$, les plus grandes moyennes des statistiques sont observées dans la situation ou les individus sont sélectionnés sur un

seul trait (S1) alors que dans le modèle $\rho_G = +1$, les plus grandes valeurs sont observées dans la situation où les individus sont sélectionnés sur les deux traits (S2).

Quelque soit la force de ρ, les moyennes des statistiques μ-F sont maximales pour le modèle $\rho_G = -1$ dans les scénarios S0 et S1. Ceci reste vrai seulement si ρ est forte (ρ=0.6) dans la situation S2. Si ρ est faible, les plus grandes statistiques sont observées sous le modèle $\rho_G = +1$ pour la sélection S2. Ces tendances sont illustrées dans la figure 21.a.

Statistiques d'association univariées (Tableau 25)

En général, les moyennes augmentent si l'effet du QTL (h_1^2 / h_2^2) augmente et varient peu en fonction de la force de la corrélation résiduelle ρ.

Pour le trait Y1 et des niveaux fixés de h_1^2, les valeurs moyennes des statistiques univariées sont très similaires, quelque soit ρ_G, dans des échantillons sélectionnés aléatoirement. Les niveaux de puissances des analyses univariées de Y1 seront donc comparables, quelque soit la présence ou non d'effets pléiotropiques. Appliquer des sélections extrêmes augmente les valeurs des statistiques (Figure 21.b). Pour la sélection S1, les valeurs des statistiques T de Student sont augmentées selon un facteur constant ($\approx$1.9) quelque soit le modèle. De même, les niveaux de puissances restent similaires quelque soit la force de la corrélation résiduelle (ρ) et le signe de la corrélation génétique. Au contraire, sous une sélection S2, les statistiques de Y1 ne varient pas de manière uniforme selon le modèle génétique. Les valeurs moyennes des statistiques sont parmi les plus élevées et les plus faibles pour les modèles à effets pléiotropiques $\rho_G = +1$ et $\rho_G = -1$ respectivement. Comme noté dans le cas bivarié, le schéma de sélection optimal dépend de ρ_G. Lorsque $\rho_G = 0$ ou $\rho_G = -1$, la puissance de détecter l'association avec Y1 est meilleure sous une sélection S1 que sous une sélection S2, alors que si $\rho_G = +1$, la puissance est meilleure si les individus sont sélectionnés sur les deux traits.

Pour le trait Y2, la statistique dépend clairement du modèle génétique simulé et de l'échantillonnage (Figure 21.c). Si les individus sont sélectionnés aléatoirement, les valeurs moyennes des statistiques univariées de Y1 et Y2 sont très similaires sous les modèles à effets pléiotropiques ($\rho_G = \pm 1$), quelque soit la force de la corrélation ρ. Appliquer une sélection sur les individus augmente la statistique de Y2 pour le modèle $\rho_G = +1$. Cependant, la tendance inverse est observée dans le modèle $\rho_G = -1$. Dans ce

scénario ($\rho_G = -1$), la moyenne de la statistique est la plus forte dans la situation S0. Finalement, pour le modèle $\rho_G = 0$, c'est-à-dire lorsque le QTL n'a pas d'effet sur Y2, les moyennes des statistiques sous S1 et S2 sont biaisées et artificiellement augmentées en raison de la corrélation résiduelle entre les traits, et ceci d'autant plus que cette corrélation est forte.

Appliquer une sélection extrême des individus sur un ou deux traits (S1 ou S2) est optimal pour les analyses univariées quand le QTL induit des effets pléiotropiques de même sens sur les traits ($\rho_G = +1$). Les moyennes des statistiques sont plus élevées que dans le cas d'une sélection aléatoire des individus.

Lorsque le QTL induit des effets en sens opposés ($\rho_G = -1$), la sélection augmente la valeur de la statistique pour Y1 seulement. Au contraire, la statistique de Y2 diminue relativement à S0 quelque soit le nombre de traits sélectionnés (S1 ou S2). Ces tendances s'amplifient dans la situation S1 si la corrélation résiduelle entre traits est forte (ρ=0.6). Au contraire, si la corrélation est modérée (ρ=0.2), ces tendances s'amplifient dans la situation S2. Finalement, dans la situation ou seulement Y1 est associé au QTL ($\rho_G = 0$), les plus grandes statistiques sont observées pour des échantillons sélectionnés. Cependant, la moyenne de la statistique d'association du trait Y2 augmente aussi artificiellement, en particulier si les individus sont sélectionnés sur les deux traits (S2).

Tableau 24 : Moyenne et écart-type (μ-F) des statistiques bivariées estimés sur 1 000 individus lorsque le variant génétique est associé aux traits (h_1^2 >0%) en fonction de l'effet du QTL sur les traits (h_1^2 / h_2^2), du signe de la corrélation génétique ρ_G et de la corrélation résiduelle ρ.

ρ	ρ_G	h_1^2 / h_2^2	**μ-F (sd)**		
			S0	**S1**	**S2**
0.2	**0**	**0.005/0**	3.69 (2.61)	10.10 (4.34)	9.23 (4.18)
		0.01/0	6.17 (3.32)	18.94 (5.86)	17.97 (5.74)
		0.03/0	17.02 (6.08)	59.02 (10.56)	55.30 (10.32)
	1	**0.005/0.005**	5.08 (2.99)	11.42 (4.86)	15.17 (5.41)
		0.005/0.01	7.45 (3.66)	14.07 (5.04)	19.08 (6.58)
		0.01/0.01	9.50 (4.42)	22.89 (6.86)	29.84 (7.89)
		0.03/0.03	26.90 (7.34)	72.04 (12.89)	92.88 (14.41)
	-1	**0.005/0.005**	7.26 (3.92)	13.91 (5.30)	9.57 (4.48)
		0.005/0.01	10.57 (4.36)	17.22 (5.65)	11.05 (4.88)
		0.01/0.01	13.69 (5.42)	27.72 (7.40)	19.29 (6.56)
		0.03/0.03	39.95 (9.40)	89.83 (13.86)	68.63 (13.23)
0.6	**0**	**0.005/0**	4.88 (3.04)	11.26 (4.51)	9.96 (4.47)
		0.01/0	8.79 (4.04)	22.54 (6.75)	19.78 (6.16)
		0.03/0	25.04 (7.34)	69.67 (11.85)	62.92 (11.30)
	1	**0.005/0.005**	4.09 (2.69)	10.41 (4.49)	12.22 (4.78)
		0.005/0.01	6.36 (3.60)	12.67 (5.06)	15.60 (5.52)
		0.01/0.01	7.33 (3.85)	20.35 (6.34)	23.81 (6.76)
		0.03/0.03	20.42 (6.53)	63.56 (11.19)	73.11 (11.61)
	-1	**0.005/0.005**	13.70 (5.32)	20.94 (6.59)	16.02 (5.84)
		0.005/0.01	19.71 (6.52)	27.35 (7.55)	21.20 (6.92)
		0.01/0.01	26.06 (7.40)	42.83 (9.58)	34.26 (8.87)
		0.03/0.03	78.65 (14.17)	143.61 (19.15)	124.18 (17.77)

*S0 : Les individus sont sélectionnés aléatoirement ; S1 : Les individus sont sélectionnés pour leur valeur au trait Y1 ; S2 : Les individus sont sélectionnés pour leurs valeurs aux traits Y1 et Y2.

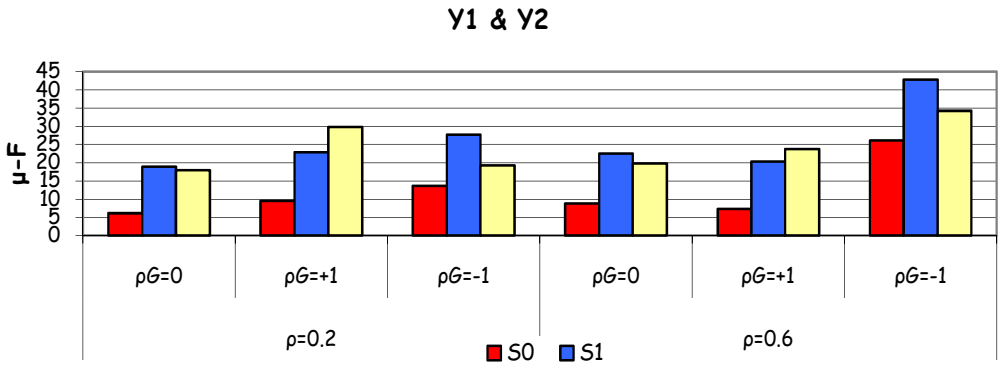
Tableau 25 : Moyenne et écart-type (μ-T) des statistiques univariées estimés sur 1 000 individus pour une MAF de 0.1 lorsque le variant génétique est associé aux traits (h_1^2>0%).

ρ	ρ_G	h_1^2 / h_2^2	S0		S1		S2	
			μ-Y1	**μ-Y2**	**μ-Y1**	**μ-Y2**	**μ-Y1**	**μ-Y2**
0.2	**0**	**0.005/0**	2.26 (1.01)	-0.03 (1.00)	4.23 (0.99)	0.98 (1.02)	4.02 (1.00)	2.59 (1.05)
		0.01/0	3.15 (1.01)	-0.02 (0.97)	5.95 (0.98)	1.41 (1.00)	5.78 (0.96)	3.60 (1.00)
		0.03/0	5.53 (1.06)	0.00 (1.00)	10.69 (0.96)	2.33 (0.96)	10.34 (0.98)	6.34 (1.01)
	1	**0.005/0.005**	2.23 (1.00)	2.20 (0.98)	4.15 (0.97)	3.21 (1.07)	4.99 (0.97)	4.83 (1.03)
		0.005/0.01	2.19 (0.97)	3.23 (1.02)	4.19 (0.97)	4.21 (0.99)	5.36 (1.02)	5.69 (1.07)
		0.01/0.01	3.18 (1.00)	3.19 (1.04)	5.96 (1.01)	4.72 (1.04)	7.06 (0.96)	6.93 (1.09)
		0.03/0.03	5.57 (1.01)	5.57 (0.98)	10.64 (1.01)	8.60 (1.07)	12.28 (0.96)	12.57 (1.13)
	-1	**0.005/0.005**	2.20 (1.01)	-2.26 (1.01)	4.18 (1.00)	-1.26 (0.99)	3.15 (0.99)	0.36 (1.02)
		0.005/0.01	2.26 (0.99)	-3.22 (0.97)	4.21 (0.92)	-2.16 (0.99)	2.69 (1.03)	-0.60 (0.99)
		0.01/0.01	3.18 (1.04)	-3.17 (1.04)	5.95 (0.98)	-1.95 (1.01)	4.56 (0.96)	0.44 (0.97)
		0.03/0.03	5.57 (1.01)	-5.58 (1.04)	10.68 (0.97)	-3.91 (1.01)	8.51 (0.96)	0.29 (0.95)
0.6	**0**	**0.005/0**	2.23 (1.00)	0.00 (0.96)	4.19 (0.97)	2.35 (0.98)	3.89 (0.99)	2.69 (0.97)
		0.01/0	3.13 (0.99)	-0.05 (0.98)	6.01 (1.00)	3.26 (0.97)	5.64 (0.97)	3.87 (0.98)
		0.03/0	5.55 (1.02)	0.01 (0.98)	10.62 (0.98)	5.44 (0.96)	10.08 (0.96)	6.63 (0.96)
	1	**0.005/0.005**	2.22 (1.00)	2.24 (0.99)	4.17 (1.00)	4.06 (1.02)	4.58 (0.99)	4.59 (1.01)
		0.005/0.01	2.25 (1.01)	3.24 (1.03)	4.17 (1.00)	4.79 (1.03)	4.90 (0.94)	5.40 (0.99)
		0.01/0.01	3.17 (1.03)	3.18 (1.00)	5.97 (1.01)	5.83 (1.01)	6.52 (0.96)	6.58 (1.00)
		0.03/0.03	5.55 (1.01)	5.60 (1.03)	10.70 (0.98)	10.47 (0.99)	11.48 (0.93)	11.75 (0.99)
	-1	**0.005/0.005**	2.20 (1.02)	-2.30 (1.02)	4.13 (0.97)	0.48 (0.95)	3.23 (0.97)	0.81 (0.92)
		0.005/0.01	2.23 (1.00)	-3.21 (1.00)	4.15 (1.00)	-0.20 (1.00)	2.97 (0.99)	0.03 (0.97)
		0.01/0.01	3.10 (1.02)	-3.22 (1.01)	5.99 (0.96)	0.70 (0.97)	4.75 (0.97)	1.11 (0.95)
		0.03/0.03	5.52 (1.04)	-5.60 (1.05)	10.68 (0.95)	0.57 (0.93)	8.89 (0.92)	1.69 (0.88)

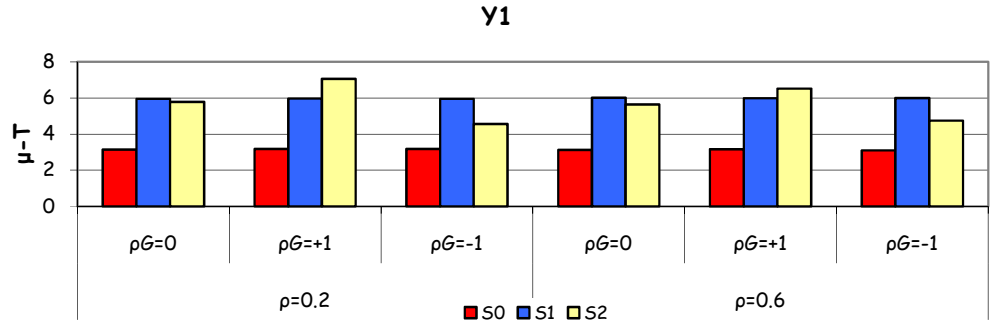
*S0 : Les individus sont sélectionnés aléatoirement ; S1 : Les individus sont sélectionnés pour leur valeur au trait Y1 ; S2 : Les individus sont sélectionnés pour leurs valeurs aux traits Y1 et Y2.

Figure 21 : Histogrammes des moyennes des statistiques des tests bivariées basé sur le modèle SUR (μ-F) et univariées (μ-T) de Y1 et Y2 par schéma de sélection : aléatoire (S0), sélection sur Y1 (S1) ou sur Y1 et Y2 (S2) lorsque N=1 000 individus, des héritabilités $h_1^2 = h_2^2 = 1\%$ en fonction de ρ_G et de ρ.

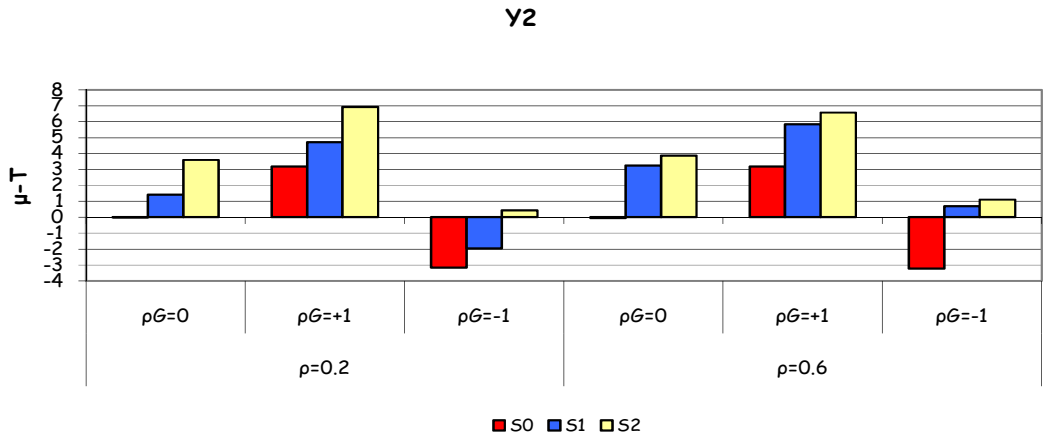
a) Analyses bivariées de Y1 & Y2



b) Analyses univariées de Y1



c) Analyses univariées de Y2

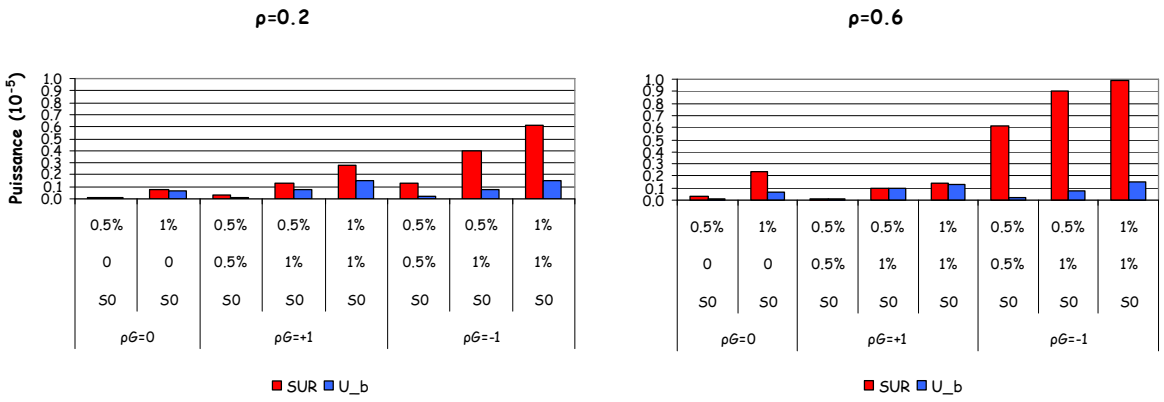


Puissances relatives des tests bivariés et univariés

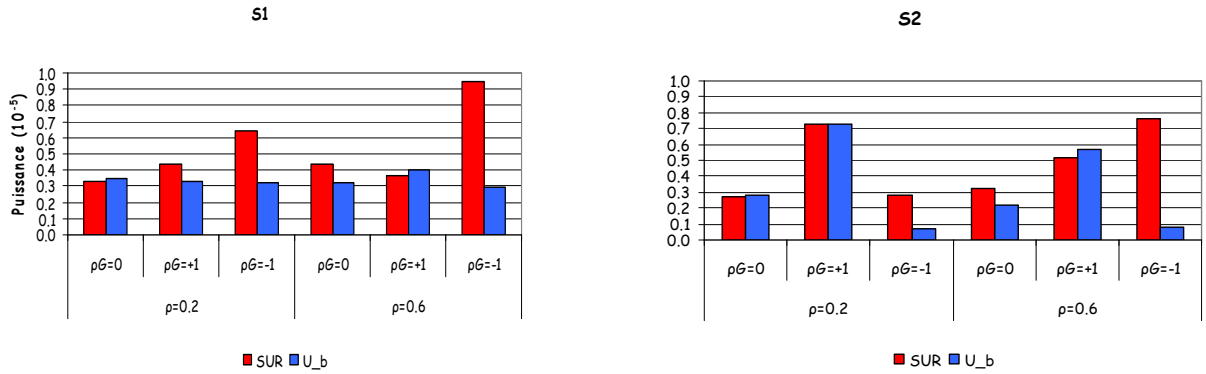
Les estimations de la puissance des tests de détecter l'association avec au moins un des traits Y1 ou Y2 en utilisant l'approche bivariée SUR sont comparées aux niveaux de puissance de détecter l'association par analyses univariées de Y1 ou Y2 ajustée par une correction de Bonferroni (U_b) au seuil nominal de 10^{-5} (Tableau 26). La figure 22.a montre les histogrammes des puissances dans le cas d'une sélection aléatoire des individus au seuil nominal de 10^{-5} en fonction des variances expliquées par le QTL sur les traits Y1 et Y2 $\left(h_1^2, h_2^2\right)$ pour N=1 000 individus et une corrélation résiduelle modérée ou forte (ρ=0.2 ou ρ=0.6). Les histogrammes des puissances en fonction de la force de la corrélation résiduelle ou en fonction du schéma de sélection (S0, S1, S2) sont montrés dans les figures 22.b et c lorsque $h_1^2 = 0.5\%$, $\rho_G = 0$ et $h_1^2 = h_2^2 = 0.5\%$ $\rho_G = \pm 1$, N=1 000 individus et ρ=0.2 ou 0.6.

Figure 22 : Estimations des niveaux de puissances (au seuil 10^{-5}) des analyses bivariées (SUR) et univariées ajustés par une correction de Bonferroni (U_b) pour différent jeux de valeurs des paramètres et 1 000 individus.

a) Estimations des puissances en fonction de l'héritabilité au QTL pour des traits modérément corrélés (ρ =0.2) ou fortement corrélés (ρ =0.6) dans des échantillons sélectionnés aléatoirement (S0).



c) Estimations des puissances pour les sélections S1 et S2 selon la force de la corrélation résiduelle (ρ =0.2 ou ρ =0.6) et une héritabilité $h_1^2 = h_2^2 = 0.5\%$.



Comme pour les statistiques de test, les performances du test bivarié dépendent du degré de corrélation résiduelle entre les traits et de la direction de la corrélation résiduelle (ρ) et de la corrélation génétique (ρ_G). D'une manière générale, quelque soit la force de la corrélation résiduelle et le schéma de sélection, le plus grand gain de puissance est obtenu dans la situation ou la corrélation génétique induite par le QTL est négative ($\rho_G = -1$) et ce gain est d'autant plus important que la corrélation résiduelle est forte (ρ =0.6).

Si les individus sont sélectionnés aléatoirement (Figure 22.a), l'analyse bivariée est toujours au moins plus puissante que l'analyse univariée avec un gain maximal si $\rho_G = -1$. Pour les autres situations, le plus grand gain du bivarié vis-à-vis de l'univarié est trouvé si $\rho_G = +1$ si la corrélation résiduelle est modérée (ρ =0.2) et pour la situation $\rho_G = 0$ si les traits sont fortement corrélés (ρ =0.6) (Figure 22.a).

Sous la sélection S1 (Figure 22.c), l'analyse bivariée est plus puissante que l'analyse univariée dans la plupart des cas. Sa puissance est légèrement plus faible que celle trouvée en analyse univariée dans les situations où $\rho_G = 0$ et ρ =0.2 et dans la situation où $\rho_G = +1$ et ρ =0.6. Pour des traits modérément corrélés (ρ =0.2), les taux de puissance sont égaux à 64.6% (SUR) vs 31.7% (U_b) si $\rho_G = -1$; 43.7% (SUR) vs 32.6% (U_b) si $\rho_G = +1$; 32.9% (SUR) vs 34.9% (U_b) si $\rho_G = 0$ (N=1 000 individus).

Sous la sélection S2 (Figure 22.c), l'analyse bivariée est plus puissante que l'analyse univariée quand $\rho_G = -1$ et quand $\rho_G = 0$ et ρ = 0.6. Pour tous les scénarios restant, l'analyse bivariée basée sur le modèle SUR montre des niveaux de puissance identiques ou légèrement inférieurs à ceux des analyses univariées. Contrairement à ce que l'on a pu observer jusqu'à présent, si les traits sont modérément corrélés (ρ =0.2), la puissance de l'analyse bivariée est meilleure si $\rho_G = +1$ que si $\rho_G = -1$.

Les meilleurs niveaux de puissance des analyses bivariées et univariées dépendent du nombre de traits sélectionnés. Si ρ=0.2, les puissances maximales des analyses bivariées et univariées sont trouvées dans la même situation c'est-à-dire lorsque $\rho_G = +1$ et sous une sélection S2 (72.5% pour SUR vs 72.9% pour U_b). Si les traits sont fortement corrélés (ρ=0.6), la puissance maximale de l'analyse bivariée est trouvée dans la situation où $\rho_G = -1$ et sous la sélection S1 (94.5%) tandis que la puissance maximale de l'analyse univariée est trouvée si $\rho_G = +1$ et une sélection S2 (56.8%). Comme montré dans le tableau 26, toutes ces tendances sont observées pour différentes valeurs des paramètres et pour des tailles d'échantillons de N=300 individus.

A travers ces résultats, on a pu voir que le plus fort niveau de puissance dépendait de la corrélation résiduelle, des effets du QTL et du schéma de sélection des individus. En effet, elle est la plus forte sous un modèle génétique pour lequel le signe des corrélations induites par le QTL et par la résiduelle est opposé, et ce d'autant plus que la corrélation augmente. Ceci reste vrai quelque soit le schéma de sélection quand la corrélation résiduelle est élevée. Au contraire, le meilleur taux de puissance bivarié est obtenu sous le modèle à effets pléiotropiques induisant des effets de même sens lorsque les traits sont modérément corrélés et sélectionnés sur les deux traits.

Tableau 26 : Taux de puissances ($p=10^{-5}$) des analyses bivariées et univariées sous différents jeux de paramètres.

N	ρ	ρ_G	h_1^2 / h_2^2	S0		S1		S2	
				SUR	U_b	SUR	U_b	SUR	U_b
1 000	**0.2**	**0**	**0.005/0**	0.008	0.008	0.329	0.349	0.273	0.277
			0.01/0	0.072	0.068	0.901	0.913	0.872	0.888
			0.03/0	0.785	0.789	1.000	1.000	1.000	1.000
		1	**0.005/0.005**	0.032	0.013	0.437	0.326	0.725	0.729
			0.005/0.01	0.127	0.072	0.675	0.510	0.889	0.891
			0.01/0.01	0.277	0.153	0.966	0.932	0.999	0.998
			0.03/0.03	0.995	0.949	1.000	1.000	1.000	1.000
		-1	**0.005/0.005**	0.132	0.023	0.646	0.317	0.280	0.068
			0.005/0.01	0.393	0.077	0.844	0.316	0.406	0.033
			0.01/0.01	0.610	0.150	0.994	0.918	0.893	0.468
			0.03/0.03	1.000	0.976	1.000	1.000	1.000	1.000
	0.6	**0**	**0.005/0**	0.034	0.012	0.440	0.323	0.325	0.217
			0.01/0	0.234	0.060	0.967	0.925	0.926	0.845
			0.03/0	0.977	0.816	1.000	1.000	1.000	1.000
		1	**0.005/0.005**	0.016	0.015	0.368	0.399	0.514	0.568
			0.005/0.01	0.094	0.096	0.564	0.602	0.760	0.797
			0.01/0.01	0.140	0.130	0.926	0.930	0.982	0.988
			0.03/0.03	0.930	0.904	1.000	1.000	1.000	1.000
		-1	**0.005/0.005**	0.613	0.022	0.945	0.293	0.763	0.079
			0.005/0.01	0.899	0.079	0.990	0.310	0.934	0.042
			0.01/0.01	0.989	0.148	1.000	0.909	1.000	0.550
			0.03/0.03	1.000	0.997	1.000	1.000	1.000	1.000
300	**0.2**	**0**	**0.005/0**	0.001	0.001	0.074	0.074	0.008	0.009
			0.01/0	0.002	0.003	0.013	0.013	0.075	0.080
			0.03/0	0.051	0.048	0.880	0.895	0.850	0.864
		1	**0.005/0.005**	0.001	0.000	0.011	0.007	0.046	0.053
			0.005/0.01	0.005	0.002	0.024	0.014	0.079	0.089
			0.01/0.01	0.010	0.006	0.124	0.089	0.286	0.306
			0.03/0.03	0.211	0.117	0.963	0.920	0.995	0.994
		-1	**0.005/0.005**	0.003	0.002	0.027	0.009	0.006	0.000
			0.005/0.01	0.015	0.001	0.062	0.006	0.017	0.001
			0.01/0.01	0.024	0.001	0.222	0.086	0.088	0.011
			0.03/0.03	0.551	0.106	0.991	0.908	0.932	0.530
	0.6	**0**	**0.005/0**	0.000	0.001	0.120	0.086	0.006	0.003
			0.01/0	0.009	0.004	0.014	0.008	0.085	0.056
			0.03/0	0.204	0.069	0.955	0.900	0.913	0.832
		1	**0.005/0.005**	0.001	0.001	0.015	0.016	0.021	0.029
			0.005/0.01	0.002	0.002	0.017	0.027	0.051	0.061
			0.01/0.01	0.005	0.002	0.102	0.124	0.150	0.187
			0.03/0.03	0.112	0.104	0.925	0.943	0.968	0.973
		-1	**0.005/0.005**	0.026	0.000	0.107	0.010	0.042	0.000
			0.005/0.01	0.103	0.004	0.229	0.012	0.119	0.000
			0.01/0.01	0.185	0.005	0.583	0.077	0.395	0.018
			0.03/0.03	0.960	0.120	1.000	0.880	1.000	0.575

*S0 : Les individus sont sélectionnés aléatoirement ; S1 : Les individus sont sélectionnés pour leur valeur au trait Y1 ; S2 : Les individus sont sélectionnés pour leurs valeurs aux traits Y1 et Y2.

3.4. Conclusions de l'étude d'association pour des individus non apparentés

Le criblage du génome pour l'association avec le Z-score utilise un échantillon d'hommes non apparentés et sélectionnés pour des valeurs basses (Z-score ≤ -2) ou élevées (Z-score ≥ 0.5) de la densité osseuse dans le but d'augmenter la puissance de détecter l'association.
Cette approche de sélection des individus a souvent été utilisée pour des études de liaisons (Devoto, Spotila et al. 2005; Kaufman, Ostertag et al. 2008; Sims, Shephard et al. 2008; Willaert, Van Pottelbergh et al. 2008; Zhang, Sol-Church et al. 2009) mais rarement dans les études d'association (Sims, Shephard et al. 2008; Kung, Xiao et al. 2010) malgré le gain de puissance apporté par ce type d'échantillonnage (Allison 1997; Allison, Heo et al. 1998; Abecasis, Cookson et al. 2001). En raison de la taille relativement petite de notre echantillon NEMO, aucun SNP ne montre d'évidence pour l'association à l'un ou l'autre des phénotypes du Z-score à LS ou FN au seuil conservatif pour un criblage du génome de $1.7x10^{-7}$ (0.05/298 783).

Cependant, le gain de puissance apporté par la sélection des individus pour des valeurs extrêmes nous a permis d'identifier trois régions (6q22; 15q14; 22q11) avec un seuil de significativité inférieur à 10^{-5} par analyse jointe de Z-LS et FN. A notre connaissance, aucune de ces régions n'a été rapportée associée à la variation de la DMO dans la littérature. Il est intéressant de noter que deux de ces régions, sur le 15q14-15 et le 22q11, contiennent des gènes potentiellement intéressants qui sont exprimés dans le muscle squelettique.

Les résultats d'association bivariés et univariés ont été comparés en termes des niveaux de signification statistique et des rangs des SNPs identifiés par l'une ou l'autre des approches. Tous les meilleurs SNPs identifiés en analyse univariée étaient également retrouvés en analyse bivariée. Inversement, de nouveaux SNPs sont retrouvés fortements associés par les analyses bivariées mais n'atteignent pas le seuil nominal de 5% par les approches univariées, sachant que ces comparaisons sont faites sans aucun ajustement des niveaux de signification pour les tests multiples.

Nous avons évalué les performances du test d'association basé sur le modèle SUR (Seemingly Unrelated Regression, (Zellner 1962)) et comparé les niveaux de puissances relativement aux tests univariés basés sur les modèles classiques de régression linéaire. L'intérêt du modèle SUR est de permettre des effets génétiques différents sur chacun des traits. Nous avons conduits des simulations extensives pour le cas de deux traits quantitatifs corrélés. L'effet du QTL avait, ou pas, des contributions égales en termes de la force des effets et de leurs directions sur les traits. De plus, nous avons simulé des échantillons d'individus sélectionnés ou non pour des valeurs extrêmes d'une distribution normale. Jusqu'ici, les performances des tests bivariés ont été étudiées dans des échantillons d'individus non apparentés ou pour des familles sélectionnées aléatoirement

dans la population. La plupart du temps, les modèles supposaient des effets génétiques identiques sur les traits (Yang, Tang et al. 2009; Zhang, Pei et al. 2009).

La puissance dépend des effets du QTL sur les traits, de la force de la corrélation residuelle phénotypique et des signes des corrélations induites par le QTL et la corrélation résiduelle. Ces résultats coïncident avec des études précédentes de la littérature pour l'analyse de liaison (Almasy, Dyer et al. 1997; Allison, Thiel et al. 1998; Amos, de Andrade et al. 2001).

Dans des échantillons sélectionnés aléatoirement, nous retrouvons des tendances de puissance bien connues. L'analyse bivariée est plus puissante que l'analyse univariée lorsque le QTL exerce des effets pléiotropiques sur les traits. Cette augmentation relative de puissance est la plus grande si le signe des corrélations induites par le QTL et par la résiduelle est en sens opposé.

Pour tous les modèles génétiques évalués, la puissance des analyses bivariées et univariées est meilleure dans des échantillons d'individus sélectionnés pour des valeurs extrêmes de la distribution que sélectionnés aléatoirement dans la population, mais aucun schéma de sélection évalué n'est optimal sur l'ensemble des modèles. Appliquer une sélection extrême sur un seul des traits (S1) donne de meilleures performances relativement à une sélection sur les deux traits (S2) si le QTL affecte un seul des traits ou si le QTL exerce des effets pléiotropiques de sens opposés. Inversement, appliquer une sélection sur les deux traits (S2) montre des niveaux de puissance supérieure vis-à-vis d'une sélection sur un seul des traits (S1) si le QTL affecte les deux traits, avec des effets de même sens.

D'une manière générale, dans des échantillons sélectionnés, l'approche bivariée basée sur le modèle SUR montre de bien meilleure performance que l'approche univariée lorsque les effets induits par le QTL sont en sens opposés. Autrement, pour les autres situations, les puissances sont relativement similaires. De manière intéressante et contrairement à ce que l'on peut observer dans des échantillons aléatoires, la puissance du test bivarié de détecter l'association dans des échantillons sélectionnés peut être la plus grande lorsque les effets induits par le QTL sur les traits sont de même sens plutôt qu'en sens opposé.

Deux études ont évalué la puissance des tests d'association bivarié par simulations dans une population de sujets non apparentés pour la variation de la DMO (Liu, Pei et al. 2009; Yang, Tang et al. 2009). Ces deux publications utilisent des modèles basés sur des modèles GEE (Generalized Estimating Equations ; Liang and Zeger 1986). Liu (Liu, Pei et al. 2009) développe une approche pour l'analyse d'association bivariée combinant deux modèles linéaires généralisés pour un trait qualitatif et quantitatif. Ces deux modèles ont des termes d'erreurs corrélés et sont combinés dans un système d'équation unique grâce au modèle SUR. Ce modèle assume des effets génétiques différents sur chacun des traits. La seconde étude (Yang, Tang et al. 2009) étudie la puissance du test d'association basé sur le modèle classique des GEE. Ce modèle impose le même effet génétique sur les traits. Le test est donc basé sur un seul degré de liberté comme le test univarié.

Ainsi, lorsque le QTL affecte avec la même contribution chacun des traits, la puissance du test bivarié basé sur des GEE sera meilleure que celle évaluée dans notre étude, basée sur le modèle SUR. Inversement, si le QTL n'affecte pas les traits de façon similaire, la puissance sous le modèle GEE sera moins bonne que celle évaluée sous le modèle SUR. Suposer les mêmes effets génétiques réduit la puissance du test bivarié par rapport au test univariée lorsque le QTL affecte un seul des traits. Le modèle GEE est donc plus approprié lorsque le QTL exerce des effets pléiotropiques similaires sur chacun des traits. Cependant, les effets génétiques induits par le QTL sont rarement connus a priori. Le modèle SUR, quant à lui, permet de considérer des effets génétiques différents sur chacun des traits tout en tenant compte de la structure de corrélation entre les traits.

Ainsi, comme montré dans cette étude, que ce soit dans des échantillons sélectionnés aléatoirement ou non, le modèle SUR est trouvé au moins aussi puissant et dans certaines situations bien meilleur que les approches univariées basées sur des modèles de régression linéaire, même quand le QTL n'exerce pas d'effets pléiotropiques sur les traits. Au vue de nos résultats de simulations, le modèle bivarié basé sur le modèle SUR semble performant pour détecter l'association de traits corrélés et montre au moins une puissance équivalente sans augmenter l'erreur de type 1, relativement aux analyses univariées. Néanmoins, distinguer des effets pléiotropiques d'effets site spécifiques est un véritable challenge, même par des approches bivariées.

Chapitre 4

4. Méthodes d'association dans des données familiales

Dans le cadre d'un atelier international Genetic Analysis Workshop 16 dont l'objectif est de comparer des méthodes d'analyses à partir de jeux de données réelles issues de la population de Framingham, nous avons utilisé des échantillons de familles dont les valeurs quantitatives simulées mimaient au mieux des phénotypes de maladies cardiaques.

Grâce à ce large échantillon de données familiales, nous avons évalué trois méthodes d'association pour des traits quantitatifs. Pour des individus apparentés, une approche intéressante est celle proposée par Fulker (Fulker, Cherny et al. 1999), permettant de tester l'association dans la méthode d'analyse de liaison basée sur la décomposition de la variance (paragraphe 2.1.1). L'association du QTL avec un marqueur est modélisée par son effet sur la moyenne phénotypique. Cette méthode a été généralisée par la suite par Abecasis (Abecasis, Cardon et al. 2000) pour des généalogies de taille plus complexe avec le test du QTDT (Quantitative Trait Disequilibrium Test). Le test d'association est, dans ce contexte, robuste à la stratification de population. En effet, les scores des génotypes aux marqueurs sont décomposés en deux composantes orthogonales, une composante entre-famille qui prend en compte le phénomène de stratification et une composante intra-famille qui n'est significative qu'en présence d'association.

Pour augmenter la taille effective des individus pris en compte dans l'estimation du paramètre d'association, Havill (Havill, Dyer et al. 2005) proposent d'utiliser également l'information apportée par les fondateurs. Cette variante est le test QTLD (Quantitative Trait Linkage Disequilibrium). Contrairement au test QTDT, le test QTLD est sensible à la stratification de population.

Une autre approche est de supposer que les composantes inter et intra-familles sont identiques comme le modèle du Measured Genotype (MG), proposé par Hopper (Hopper and Mathews 1982; Boerwinkle, Chakraborty et al. 1986). Ce modèle exploite à la fois les variations entre les familles et intra-familiales. Le modèle d'association MG, basé sur un modèle linéaire mixte, a été proposé comme une alternative puissante pour la détection

de QTL (Havill, Dyer et al. 2005; Aulchenko, de Koning et al. 2007). Comme pour les tests QTDT et QTLD, ce modèle nécessite l'estimation de la composante polygénique ce qui peut être long en temps de calcul. Aulchenko et al (2007) ont proposé une extension du test MG adaptée aux données de criblage du génome. Afin de sélectionner seulement un sous ensemble de marqueurs pour tester l'association avec le test MG, les phénotypes sont d'abord ajustés pour la composante polygénique. Le modèle d'association repose alors sur une régression linéaire classique pour des données de sujets non apparentés. Seulement les SNPs les plus significatifs sont testés pour l'association par le test MG. Cependant, comme pour le test QTLD, le test MG est sensible à la stratification de population.

Toutes ces méthodes introduisent un test d'association dans la méthode de décomposition de la variance permettant de tenir compte des corrélations familiales. Ces trois tests d'association QTDT, QTLD et MG sont applicables dans des données familiales, mais diffèrent dans la quantité d'information utilisée pour tester l'association. MG tient compte de tous les individus génotypés et phénotypés alors que les méthodes de décomposition orthogonale (QTLD et QTDT) n'utilisent qu'un sous effectif de cet échantillon. Pour QTDT, l'échantillon est encore plus réduit, car ce test n'utilise pas l'information apportée par les fondateurs. En conséquence, QTDT pourrait manquer de puissance relativement à QTLD ou MG. Cependant, à la fois QTLD et MG peuvent être affectés par une association allélique due à une stratification de population.

La puissance relative des ces trois méthodes n'a été que peu étudiée (Havill, Dyer et al. 2005; Aulchenko, de Koning et al. 2007). L'objectif de l'étude GAW16 été d'explorer l'erreur de type 1 et de comparer la puissance relative de ces trois méthodes (QTDT, QTLD et MG) pour l'analyse d'association de traits quantitatifs dans des données familiales de grande taille.

4.1. Matériel et méthode

4.1.1. Les données GAW16

La population de Framingham est suivie depuis 1948, l'objectif étant d'identifier les facteurs communs ou les caractéristiques contribuant aux maladies cardiovasculaires en suivant son développement sur une longue période de temps et dans un large groupe de participants qui n'ont pas encore développés de symptômes pour les maladies cardiaques ou qui ont déjà souffert d'une attaque cardiaque. L'étude s'est orientée sur les facteurs génétiques associés à l'élaboration d'une maladie cardio-vasculaire. Les prélèvements ADN ont commencé dans les années 1990 et en 2007, une phase de génotypage a été effectuée sur environ 550 000 SNPs répartis sur tout le génome. Le jeu de marqueurs est

issu d'une puce Affymetrix 500K (Gene chip Human Mapping 500K) ainsi que de SNPs additionnels de gènes candidats (50K Human Gene Focused Panel).

Les **données simulées** dans le cadre du congrès sont issues du problème 3. Au total, 200 réplicats ont été générés. Seulement les phénotypes des individus sont simulés, les génotypes étant fixés.

Les traits quantitatifs simulés sont :

- le taux de Triglycérides (*TG*)
- le taux de « bon » cholestérol (*HDL* : High Density Lipoprotein)
- le taux de « mauvais » cholestérol (LDL : Low Density Lipoprotein)

Ainsi que les covariables sexe, âge et alimentation (*Diet*).

La moyenne d'âge simulée, basée sur la population de Framingham, est de 43 ans (min=19 ans et max=80 ans).

La covariable âge est décomposée en 13 classes, chacune d'intervalle égale à 5 ans et la covariable alimentation (*Diet*) est ordinale (de 4 à 16).

Modèles génétiques : Pour chaque réplicat, les phénotypes HDL, LDL et TG ont été simulés sur la base des modèles génétiques suivant :

$$HDL = \hat{\mu}_{HDL|\text{âge stratifié, sexe}} \times \hat{\sigma} + h_{\alpha_1} \times a_{\alpha_1} + h_{\alpha_2} \times a_{\alpha_2} + h_{\alpha_3} \times a_{\alpha_3} + h_{\alpha_4} \times a_{\alpha_4} + h_{\delta_1} \times a_{\delta_1} + [apoly] \times h_{apoly} + h_{\varepsilon} \times a_{\varepsilon}$$

$$LDL = \hat{\mu}_{LDL|\text{âge stratifié, sexe}} \times \hat{\sigma} + h_{\beta_1} \times b_{\beta_1} + h_{\beta_2} \times b_{\beta_2} + h_{\beta_3} \times b_{\beta_3} + h_{\beta_4} \times b_{\beta_4} + h_{\beta_5} \times I(\alpha_4 \times \beta_5) + h_{\delta_2} \times a_{\delta_2} + [bpoly] \times h_{bpoly} + h_{\varepsilon} \times b_{\varepsilon}$$

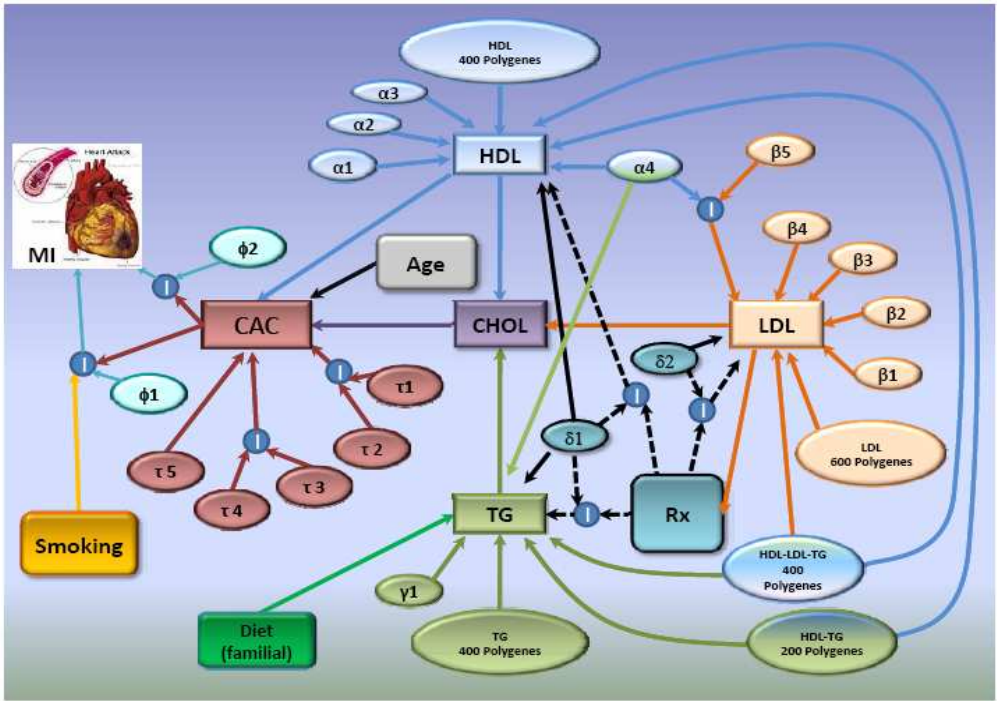
$$TG = \hat{\mu}_{TG|\text{âge stratifié, sexe}} \times \hat{\sigma} + h_{\gamma_1} \times g_{\gamma_1} + h_{\alpha_4} \times g_{\alpha_4} + h_{\delta_1} \times g_{\delta_1} + h_{Diet} \times g_{Diet} + [gpoly] \times h_{gpoly} + h_{\varepsilon} \times g_{\varepsilon}$$

où h est la valeur des variables explicatives des effets gènes majeurs et des effets polygéniques notés respectivement :

- (α_1, α_2, α_3, α_4, δ_1) et [apoly] pour HDL
- (β_1, β_2, β_3, β_4, β_5, δ_2) et [bpoly] pour LDL
- (γ_1, α_4, δ_1) et [gpoly] pour TG
- I(α_4x β_5) est l'effet d'interactions entre α_4 et β_5.

Le **modèle de simulation** est montré dans la figure 23.

Figure 23 : Modèle de simulation



Les corrélations phénotypiques entre les traits et les héritabilités (en % sur la diagonale) de chacun des traits sont indiquées dans le tableau suivant :

	HDL	LDL	TG
HDL	60 %	-0.21	-0.28
LDL		55 %	0.24
TG			40 %

4.1.2. Méthodes d'association pour données familiales

Nous reprenons les mêmes notations utilisées tout au long de ce manuscrit. Les phénotypes des individus de la famille i sont représentés par le modèle linéaire suivant :

$$y_i = \mu + \beta \times g_i + c_i + e_i$$

Sous un modèle additif, on définit le score génotypique g_{ik} au marqueur d'un individu k dans une famille i par :

$$g_{ik} = \begin{cases} 1 & si\ A1A1 \\ 0 & si\ A1A2 \\ -1 & si\ A2A2 \end{cases}$$

Le vecteur y_i est le vecteur des phénotypes de la famille i ; $y_i \sim N(\mu_i, \Omega_i)$. β est le coefficient de l'effet du marqueur ; c_i et e_i sont les effets aléatoires polygéniques et résiduelles de la famille i, de moyennes nulles et de variances respectives σ_C^2 et σ^2.

L'effet aléatoire c_i tient compte des liens d'apparentés entre les individus.

Les coefficients de la matrice de variance/covariance phénotypique entre des individus k et l de la famille i sont données par : $\Omega_{kl} = \begin{cases} \sigma_C^2 + \sigma_E^2 & \text{si k = l} \\ 2\Phi_{kl} \times \sigma_C^2 & \text{si k} \neq \text{l} \end{cases}$

où Φ_{kl} est la matrice des coefficients d'apparentement entre des apparentés k et l de la famille i (paragraphe 1.1).

Modèle du Measured Genotype (MG)

Le modèle d'association du **Measured Genotype** (Boerwinkle, Chakraborty et al. 1986) est basé sur une régression linéaire mixte à effets aléatoires C. La moyenne phénotypique μ_{ik} de l'individu k de la famille i est donnée par :

$$\mu_{ik} = \mu + \beta \times g_{ik}$$

et la variance du modèle est $\Omega_{kl} = \begin{cases} \sigma_C^2 + \sigma_E^2 & \text{si k = l} \\ 2\Phi_{kl} \times \sigma_C^2 & \text{si k} \neq \text{l} \end{cases}$

Le test d'association MG entre le locus marqueur et le trait porte sur le paramètre de régression β (β=0 vs β≠0). Sous l'hypothèse nulle H_0 (β=0), il n'y a pas d'association entre le locus marqueur et le phénotype et $\mu_{ik} \equiv \mu$. Pour faire ce type de test, on utilise le test du rapport des vraisemblances. Sous H_0 la statistique du rapport des vraisemblances suit un χ^2 à 1 degré de liberté (χ_1^2).

On remarque que dans le calcul du score génotypique g_{ik}, tous les individus k de la famille i sont pris en compte.

Modèle de décomposition orthogonale des scores génotypiques

On peut décomposer le score génotypique g_{ik} en deux composantes orthogonales (Fulker, Cherny et al. 1999) sous la forme : $g_{ik} = E(g_{ik}) + (g_{ik} - E(g_{ik}))$

où

- $E(g_{ik})$ représente la moyenne du score génotypique de l'individu k de la famille i sachant sa place dans la fratrie (variation entre les familles).
- $g_{ik} - E(g_{ik})$ représente la déviation du score génotypique observée par rapport à l'attendu pour l'individu k (variation à l'intérieur des familles).

Posons :

$$\begin{cases} b_{ik} = E(g_{ik}) \\ w_{ik} = g_{ik} - E(g_{ik}) \end{cases} \quad \textbf{(1.8)}$$

L'association avec le marqueur est modélisée par :

$$\mu_{ik} = \mu + \beta_b \times b_{ik} + \beta_w \times w_{ik}$$

où b_{ik} et w_{ik} sont les composantes génotypiques inter et intra-familles ; β_b et β_w sont respectivement l'effet génotypique moyen au locus entre les familles et l'effet de la déviation génotypique par rapport la moyenne (génotypique) attendue pour l'individu k de la famille i.

Le test d'association repose sur le paramètre β_w qui représente l'effet de l'association à l'intérieur des familles. Sous l'hypothèse nulle, $\beta_w = 0$, on ne peut mettre en évidence une association entre le marqueur et la variation du trait à l'intérieur des familles. Le paramètre β_b est estimé. Sous l'hypothèse alternative ($\beta_w \neq 0$), il existe une association entre la variation du trait à l'intérieur des familles et la variation des génotypes au marqueur. La statistique du rapport des vraisemblances suit sous l'hypothèse nulle un χ_1^2 à 1 degré de liberté.

***- Quantitative Trait Disequilibrium Test* (QTDT)**

- Si l'individu k est un fondateur, le score génotypique b_{ik} est égale à g_{ik} et $w_{ik} = g_{ik} - b_{ik} = 0$.
- Si l'individu k est un apparenté, notons P_k et M_k le père et la mère de l'individu k de la famille i.

Pour une famille nucléaire i où k est l'individu de la fratrie l, avec n_l individus dans la fratrie, les **scores génotypiques** inter (b_{ik}) et intra (w_{ik}) familles sont définis selon la place de l'individu dans la famille par :

$$\begin{cases} b_{ik} = b_i = \begin{cases} \dfrac{g_{iP} + g_{iM}}{2} & \textit{si les parents sont typés} \\ \displaystyle\sum_{k}^{n_l} \dfrac{g_{ik}}{n_l} & \textit{si au moins un des parents n'est pas typé} \end{cases} \\ w_{ik} = g_{ik} - b_{ik} \end{cases}$$

où g_{iP}, g_{iM} et g_{ik} sont les scores génotypiques du père, de la mère et de l'individu k dans la famille i et n_l est le nombre de germains dans la fratrie, $n_l \leq n_i$ avec n_i le nombre d'individus dans la famille i.

Pour une famille plus complexe et seulement si les parents sont typés, b_{ik} s'écrit plus généralement sous la forme :

$$b_{ik} = \begin{cases} \dfrac{b_{iP_k} + b_{iM_k}}{2} & \textit{si aucun des parents n'est fondateur} \\ \dfrac{g_{iP_k} + b_{iM_k}}{2} & \textit{si seulement le père est un fondateur} \end{cases}$$

On remarque que les b_{ik} sont tous identiques à l'intérieur d'une même fratrie l de taille n_l.

- Test du *Quantitative Trait Linkage Disequilibrium* (QTLD)

- Si l'individu k est un fondateur : b_{ik} est égale à 0 et $w_{ik} = g_{ik}$.
- Si l'individu k est un apparenté, les scores génotypiques sont identiques à ceux définis pour le test QTDT.

Pour ne pas porter à confusion nous notons (b'_{ik}, w'_{ik}) les scores génotypiques inter et intra-familles calculés pour le test QTLD. Le modèle d'association est :

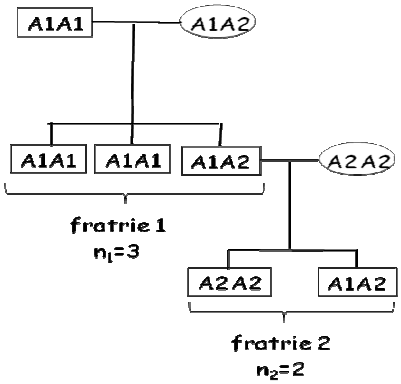
$$\mu_{ij} = \mu + \beta'_b \times b'_{ik} + \beta'_w \times w'_{ik}$$

Sous l'hypothèse nulle ($\beta'_w = 0$ et β'_b est estimé), on ne peut mettre en évidence une association entre la variation du trait à l'intérieur des familles et la variation des

génotypes au marqueur. La statistique du rapport des vraisemblances suit un χ_1^2 à 1 degré de liberté.

Exemple :

Nous donnons ci-dessous un exemple du calcul des composantes orthogonales des scores génotypiques b_{ik} et w_{ik} dans une famille avec 2 fratries (l=2) pour laquelle tous les individus sont génotypés. Les scores génotypiques pour les trois tests d'association MG, QTDT et QTLD sont données dans le tableau. Pour QTLD, les scores génotypiques des composantes inter et intra-familles pour des individus apparentés sont identiques pour le test QTDT et QTLD, nous ne les montrons pas dans la table.



		MG	QTDT		QTLD	
		g_{ik}	b_{ik}	w_{ik}	b'_{ik}	w'_{ik}
Fondateur	A1A1	1	1	0	0	1
	A1A2	0	0	0	0	0
Fratrie 1 n_l=3	A1A1	1	1/2	1/2		
	A1A1	1	1/2	1/2		
	A1A2	0	1/2	-1/2		
Fondateur	A2A2	-1	-1	0	0	-1
Fratrie 2 n_l=2	A2A2	-1	-1/4	-3/4		
	A1A2	0	-1/4	1/4		

Les deux tests d'association (QTDT et QTLD), basés sur une décomposition orthogonale des scores génotypiques, n'utilisent pas le même nombre d'individus pour le calcul des composantes inter et intra-familles. Pour le test QTDT (Abecasis, Cardon et al. 2000), le génotype des fondateurs est inclus dans la composante inter-familles, alors que pour le test QTLD (Havill, Dyer et al. 2005), le génotype des fondateurs est inclus dans la composante intra-familles.

Familles informatives et taille effective des échantillons utilisés

Pour certaines configurations des individus et selon les données génotypiques disponibles, les fratries ne sont pas informatives. Les fratries non informatives sont celles

pour lesquelles on ne peut mettre en évidence une variation à l'intérieur de la fratrie. La fratrie de l'individu k est non informative dans les situations suivantes :

– Si les deux parents sont des fondateurs typés et tous les deux homozygotes alors quelque soit le germain k dans la fratrie, tous les génotypes des enfants sont identiques, on ne peut mettre en évidence de variation à l'intérieur de la fratrie et $g_{ik} = E(g_{ik})$.

  En effet, si les parents sont homozygotes A1A1 ($g_P = g_M$ =1) ou A2A2 ($g_P = g_M$ =-1), alors tous les enfants sont soit A1A1 (g_{ik} =1 et b_i =1) soit A2A2 (g_{ik} =-1 et b_i =-1). Si les parents sont pour l'un A1A1 (par exemple g_P =1) et pour l'autre A2A2 (g_M =-1) alors tous les enfants sont forcément A1A2 (g_{ik} =0 et b_i =0).

– Si au moins un des parents n'est pas typé ($b_{ik} = \sum_{k}^{n_l} \frac{g_{ik}}{n_l}$) et s'il n'y a qu'un seul individu k dans la fratrie l ($n_l = 1$) ou, si tous les enfants de la fratrie l sont de même génotype homozygote ou hétérozygote alors on ne peut mettre en évidence de variation à l'intérieur de la fratrie et $g_{ik} = E(g_{ik})$.

D'une manière générale, les fratries informatives de ces deux tests d'association (QTDT et QTLD) sont celles pour lesquelles :

– Si les parents sont des fondateurs typés : au moins un des parents doit être hétérozygote.

– Ou : si au moins un des parents n'est pas typé, il faut qu'il existe au moins deux enfants dans la fratrie de génotypes différents.

Ces trois méthodes d'association QTDT, QTLD et MG différent dans la quantité d'information, utilisée au marqueur, pour tester l'association. En effet, la méthode MG utilise tous les individus génotypés alors que QTDT et QTLD n'utilise qu'une partie de cet échantillon. De plus, QTDT utilise encore moins d'information relativement à QTLD. Pour QTDT, si l'individu k est un fondateur alors il ne contribue pas dans l'estimation de l'effet d'association intra-famille w_{ik}. Le nombre effectif d'individus est le nombre d'individus apparentés tel que la composante intra-famille soit différente de 0 ($w_{ik} \neq 0$). Tandis que pour QTLD, le nombre d'individus pris en compte est l'ensemble des individus apparentés et fondateurs tel que w'_{ik} soit différent de 0.

4.2. Stratégies d'analyse

Choix des traits quantitatifs étudiés pour les analyses d'association

Les traits quantitatifs utilisés pour cette étude sont le taux de Triglycérides (TG) et le taux de « bon » cholestérol (HDL), ainsi que les informations sur les covariables mesurées à la première visite pour les trois générations. Dans chaque réplicat, les traits ont été ajustés pour l'âge puis pour le sexe à l'aide d'une simple régression où l'âge est regroupé en 13 classes. Les valeurs résiduelles de HDL et TG ont été utilisées comme les deux traits d'intérêts.

TG ne suivant pas une distribution normale, nous l'avons transformé en utilisant la méthode de Yalcin et al (2004) basé sur les rangs. Le trait obtenu est nommé : TG_R.

Les analyses ont été conduites avec le logiciel statistique R.

Nettoyage des données génotypiques

Les données génotypiques non filtrées ont été recueillies à partir de la puce Affymetrix 500K (Gene chip Human Mapping 500K). Elles ont été filtrées sur la base du score de confidence BRLMM (Bayesian Robust Linear Model with Mahalanobis distance) développé par la société Affymetrix. Le seuil de rejet est fixé au niveau standard ½. Nous avons nettoyé ces données par :

- Exclusion des SNPs avec un taux de données manquantes supérieur à 5%, une position sur la carte inconnue ou une fréquence de l'allèle mineur (MAF : Minor Allele Frequency) inférieure à 1%.
- Tous les génotypes d'un individu sont mis à zéro si son taux de données manquantes est supérieur à 5%.
- Exclusion des SNPs significatifs à l'hypothèse nulle de l'équilibre d'Hardy-Weinberg (p-value$\leq 10^{-6}$).
- Les génotypes sont mis à zéro pour tous les individus de la famille si le SNP montre une erreur de transmission mendélienne.

Choix des vrais gènes majeurs

Trois gènes majeurs ont été sélectionnés sur les chromosomes 8 et 19. Le gène α_4 situé sur le chromosome 8 avait un effet pléiotropique sur TG et HDL. Deux gènes co-localisés α_2 et δ_1 sur le chromosome 19, avaient un effet sur HDL et un effet sur HDL et TG

respectivement. Pour étudier l'erreur de type 1, un « faux » gène majeur a été choisi aléatoirement sur le chromosome 7.

A partir de ces gènes majeurs, nous avons identifiés tous les SNPs dans une région de 100kb de part et d'autre du vrai gène majeur. Ces régions définissent les régions candidates pour étudier la puissance et l'erreur de type 1 des tests d'association.

Pour chacune de ces régions candidates, le coefficient de corrélation r^2 du déséquilibre de liaison entre le variant causal et le SNP a été évalué avec Haploview 4.1 (Barrett, Fry et al. 2005) sur 737 fondateurs.

Le tableau 27 donne les principales caractéristiques des SNPs testés : la position en paires de base (bp), le nom du SNP, la fréquence de l'allèle mineur (MAF), le symbole pour les gènes majeurs, la part de variance attribuable à l'effet gène majeur (σ_G^2) pour HDL et TG et le coefficient de corrélation r^2 du déséquilibre de liaison entre les allèles du gène majeur et chaque SNP non associé ($r^2 \leq 0.003$) mais génétiquement liés.

Tableau 27 : Caractéristiques des SNPs testés (causaux et non causaux)

Chr	Pos (bp)	SNP	MAF (%)	Gène	$\sigma_G^2(HDL)$	$\sigma_G^2(TG)$	r^2 (symbole)
7	24 734 008	rs2521760	12.7	aucun			
	24 822 557	rs10215692	11.0				0.002 (aucun)
8	19 794 163	rs17091651	10.0				0.002 (α_4)
	19 868 351	rs3200218	21.7	α_4	0.3% (DOM)	0.4% (ADD)	
	19 943 326	rs4244457	32.9				0.003 (α_4)
19	46 010 146	rs11083567	18.2				0 (α_2) 0.001 (δ_1)
	46 089 501	rs8103444	24.4	α_2	0.2% (ADD)	-	0 (δ_1)
	46 210 613	rs8192719	24.9	δ_1	0.3% (ADD +10%)	0.3% (ADD -15%)	0 (α_2)
	46 335 684	rs1631931	13.5				0 (α_2) 0 (δ_1)

On observe que 2% des individus mesurés à la première visite prennent des traitements pour diminuer le taux de « mauvais » cholestérol (LDL). En δ_1, les homozygotes pour l'allèle majeur et les hétérozygotes répondent à ce traitement médicamenteux. Pour ces individus, le taux de HDL augmente de 10% et le taux de TG diminue de 15%.

Echantillon de familles

Tous les individus non apparentés (187 singletons au total) ont été exclus. Sur un total de 940 pedigrees, 704 ont ensuite étés sélectionnés car ils contenaient aux moins 2 individus non fondateurs avec des données génotypiques et phénotypiques. Sur un total de 12 407 individus, après reconstitution des familles complètes à partir des données réelles de Framingham, 6 009 avaient des données phénotypiques.

Selon le SNP, entre 5 826 et 5 995 avaient une valeur phénotypique et génotypique et parmi eux, environ 10% étaient des fondateurs. La taille moyenne des familles variait de 4 à 639 individus avec en moyenne 8 individus par famille.

Analyses

Toutes les analyses d'association ainsi que le test de stratification ont été effectués en utilisant la commande *qtld* de SOLAR 4.0.7 (Almasy and Blangero 1998).

Le test de stratification proposé par (Fulker, Cherny et al. 1999) et utilisé par SOLAR 4.0.7 repose sur les hypothèses suivantes:

$$\begin{cases} H_0 : \beta_b = \beta_w \\ H_1 : \beta_b \neq \beta_w \end{cases}$$

Sous H_0, la statistique du rapport des vraisemblances suit un χ^2 à 1 degré de liberté.

Nous avons évalué les trois méthodes d'association (QTDT, QTLD et MG) dans des données familiales sous plusieurs conditions :

- Distribution du trait TG. En effet, les tests d'association sont utilisés dans un cadre de décomposition de la variance et utilisent la méthode du maximum de vraisemblance. Ces méthodes sont sensibles à la normalité des traits.

  Notation : TG_R.

- Avec/sans l'inclusion de la covariable *Diet*. En effet, cette covariable affecte le niveau de triglycérides et est corrélée entre membre d'une même famille.

  Notation : HDL_Diet, TG_Diet.

- Les tests d'association QTLD et MG étant sensibles à la stratification, nous avons comparé les résultats obtenus selon que l'on tienne compte ou non des SNPs significatifs au seuil nominal de 5% pour le test de stratification. Dans ce

cas les valeurs du χ_1^2 des tests d'association QTLD et MG ont été mise à zéro dans les réplicats pour lesquels le SNP était significatif pour le test de stratification.

Notation : QTLD|S, MG|S.

Pour chacun des traits (HDL, HDL_Diet, TG, TG_Diet et TG_R) et chacun des SNPs sélectionnés dans la région, nous avons effectué les trois tests d'association (et le test de stratification) pour chaque réplicat. Nous calculons les moyennes et écarts-types des statistiques des tests d'association sur 200 réplicats. Nous montrons aussi les moyennes et écarts-types des paramètres d'association (coefficients de régression) en annexe. Le taux d'erreur et la puissance sont définis comme la proportion de réplicats dont la statistique du χ_1^2 dépasse la valeur nominale.

Nous avons évalué l'erreur de type 1 sous deux hypothèses :

- Absence d'association et absence de liaison

  Les résultats sont montrés aux SNPs rs2521760 et rs10215692. Le premier a été choisi aléatoirement sur le chromosome 7 et le second a été sélectionné dans une région de 100kb autour du rs2521760 tel que la valeur du déséquilibre de liaison r^2 soit minimum (Tableau 27).

- Absence d'association en présence de liaison

  Deux SNPs sur le chromosome 8 (rs17091651 et rs4244457) et deux sur le chromosome 19 (rs11083567 et rs1631931) ont été sélectionnés dans les régions du chromosome 8 et 19 tel que le déséquilibre de liaison avec les gènes majeurs soit minimum ($r^2 \leq 0.003$) (Tableau 27).

La puissance des tests est estimée pour les vrais gènes majeurs α_4 sur le chromosome 8 et α_2, δ_1 sur le chromosome 19. Ces deux SNPs sont liés ($\approx$120kb) mais non associés ($r^2 = 0$).

Les trois tests d'association, ainsi que le test de stratification, suivent une loi de χ_1^2 à 1 degré de liberté sous l'hypothèse nulle. On s'attend à une moyenne de 1 et une variance de 2 (écart-type ≈ 1.41) sur les 200 réplicats.

4.3. Résultats

Caractéristiques des distributions des traits

Le tableau 28 montre les principales caractéristiques des distributions des traits (HDL, TG et TG_R) calculées pour 200 réplicats : moyenne ($\pm$ écart-type) des valeurs des traits, du kurtosis, du skewness et l'estimation de la composante polygénique (h^2).

Tableau 28 : Caractéristiques des distributions des traits : moyenne (écart-type)

Trait	moyenne (écart-type)	Kurtosis	Skewness	h^2
HDL	57.48 (13.45)	0.59 (0.18)	-0.008 (0.03)	0.53 (0.02)
TG	131.34 (68.88)	16.09 (14.50)	2.47 (0.57)	0.30 (0.03)
TG_R	249.30 (68.80)	-0.02 (0.01)	0.003 (0.01)	0.33 (0.02)

La composante polygénique est plus forte pour le trait HDL que pour TG (0.53 ± 0.03 vs 0.30 ± 0.03). Contrairement à TG, HDL ne s'écarte pas des caractéristiques d'une distribution d'une loi normale. Transformer le trait TG (TG_R) augmente légèrement la valeur de la composante polygénique.

Caractéristiques des distributions des tests

Hypothèses nulles

Les tableaux 29 et 30 montrent les estimations empiriques des moyennes ($\pm$ écarts-types) des statistiques du χ_1^2 sous l'hypothèse nulle. Pour les trois tests QTDT, QTLD et MG, nous montrons les estimations sous l'hypothèse d'absence d'association et absence de liaison (Tableau 29) et sous l'hypothèse d'absence d'association en présence de liaison (Tableau 30) avec et sans les SNPs significatifs pour le test de stratification (QTLD|S, MG|S). Nous montrons également les erreurs de type 1 au niveau nominal de 5%. Le taux d'erreur pour le test de stratification est donné dans la dernière colonne.

Tableau 29 : Moyenne des estimations des statistiques des tests d'association (± écart-type moyen) et des erreurs de type 1 (p-nom=5%) sous l'hypothèse d'absence d'association et absence de liaison

Trait	Chr	SNP	QTDT	QTLD	QTLD|S*	MG	MG|S*	p-nominal=5%					
								QTDT	QTLD	QTLD|S*	MG	MG|S*	Strat**
HDL	7	rs2521760	**0.48** (0.62)	0.82 (1.00)	0.82 (1.00)	0.73 (0.85)	0.72 (0.86)	**0%**	**1%**	**1%**	**0%**	**0%**	**1%**
		rs10215692	0.99 (1.16)	**0.51** (0.61)	**0.46** (0.56)	**0.44** (0.53)	**0.42** (0.53)	4%	**0%**	**0%**	**0%**	**0%**	7%
HDL Diet	7	rs2521760	**0.48** (0.62)	0.82 (1.00)	0.82 (1.00)	0.73 (0.86)	0.73 (0.86)	**0%**	**1%**	**1%**	**0%**	**0%**	**1%**
		rs10215692	0.99 (1.17)	**0.51** (0.61)	**0.46** (0.56)	**0.44** (0.53)	**0.42** (0.53)	4%	**0%**	**0%**	**0%**	**0%**	7%
TG	7	rs2521760	0.86 (1.10)	1.00 (1.15)	0.97 (1.15)	**0.64** (0.70)	**0.60** (0.69)	3%	4%	4%	**1%**	**1%**	**8%**
		rs10215692	**1.44** (1.89)	**1.98** (2.49)	**1.77** (2.22)	**1.32** (1.70)	1.23 (1.61)	**11%**	**18%**	**15%**	**8%**	7%	3%
TG_Diet	7	rs2521760	0.87 (1.16)	1.00 (1.16)	0.94 (1.16)	**0.62** (0.64)	**0.56** (0.63)	4%	4%	4%	**0%**	**0%**	11%
		rs10215692	**1.43** (1.78)	**2.03** (2.52)	**1.86** (2.28)	**1.35** (1.74)	1.25 (1.62)	**10%**	**17%**	**16%**	**8%**	7%	**2%**
TG_R	7	rs2521760	0.99 (1.28)	0.92 (1.24)	0.76 (1.03)	**0.48** (0.63)	**0.44** (0.62)	5%	4%	**2%**	**1%**	**1%**	7%
		rs10215692	**1.32** (1.68)	0.92 (1.22)	0.92 (1.23)	**1.30** (1.65)	**1.29** (1.66)	9%	3%	3%	8%	8%	**1%**

* : Test d'association après avoir retiré les SNPs significatifs pour le test de stratification.
** : Proportion de SNPs significatifs pour le test de stratification avec p-nominal=5%

Tableau 30 : Moyenne des estimations des statistiques des tests d'association (± écart-type moyen) et des erreurs de type 1 (p-nom=5%) sous l'hypothèse d'absence d'association en présence de liaison

Trait	Chr	SNP	QTDT	QTLD	QTLD|S *	MG	MG|S *	p-nominal=5%					
								QTDT	QTLD	QTLD|S*	MG	MG|S*	Strat**
HDL	8	rs17091651	1.03 (1.12)	**1.36** (1.51)	**1.36** (1.51)	0.93 (1.01)	0.93 (1.01)	**2%**	7%	7%	3%	3%	**0%**
		rs4244457	**1.50** (1.79)	1.03 (1.36)	**0.60** (0.92)	**0.63** (0.79)	**0.51** (0.75)	6%	6%	**1%**	**1%**	**1%**	**21%**
	19	rs11083567	0.85 (0.97)	**0.48** (0.63)	**0.38** (0.53)	**0.40** (0.51)	**0.37** (0.52)	**2%**	**0%**	**0%**	**0%**	**0%**	7%
		rs1631931	**1.86** (2.02)	0.72 (1.09)	**0.61** (0.85)	1.08 (1.24)	0.99 (1.19)	**12%**	**2%**	**1%**	**2%**	**1%**	8%
HDL Diet	8	rs17091651	1.03 (1.12)	**1.36** (1.50)	**1.36** (1.50)	0.93 (1.01)	0.93 (1.01)	**2%**	7%	7%	3%	3%	**0%**
		rs4244457	**1.50** (1.79)	1.03 (1.37)	**0.60** (0.93)	**0.63** (0.79)	**0.51** (0.74)	6%	6%	**1%**	**1%**	**1%**	**21%**
	19	rs11083567	0.86 (0.98)	**0.48** (0.64)	**0.38** (0.53)	**0.40** (0.51)	**0.37** (0.52)	3%	**0%**	**0%**	**0%**	**0%**	7%
		rs1631931	1.86 (2.02)	0.72 (1.09)	**0.62** (0.85)	1.08 (1.24)	1.00 (1.19)	**12%**	**2%**	**1%**	**2%**	**1%**	7%
TG	8	rs17091651	**0.70** (0.88)	0.93 (1.31)	0.88 (1.32)	0.89 (1.10)	0.83 (1.12)	**1%**	4%	4%	5%	5%	**11%**
		rs4244457	**1.76** (1.87)	1.27 (1.47)	0.99 (1.33)	**0.65** (1.13)	**0.61** (1.12)	**14%**	7%	5%	3%	3%	**13%**
	19	rs11083567	0.75 (0.95)	1.11 (1.22)	1.03 (1.18)	**0.55** (0.77)	**0.53** (0.77)	**1%**	4%	3%	**2%**	**2%**	5%
		rs1631931	0.93 (1.35)	1.19 (1.41)	1.10 (1.33)	**0.62** (0.79)	**0.61** (0.79)	6%	**9%**	7%	**1%**	**1%**	**2%**
TG Diet	8	rs17091651	0.73 (0.90)	0.92 (1.31)	0.85 (1.31)	0.89 (1.13)	0.84 (1.14)	**1%**	4%	4%	5%	5%	**11%**
		rs4244457	**1.82** (1.97)	**1.29** (1.51)	0.97 (1.34)	**0.66** (1.19)	**0.60** (1.18)	**14%**	7%	5%	3%	3%	13%
	19	rs11083567	0.76 (1.00)	1.10 (1.26)	1.01 (1.19)	**0.54** (0.78)	**0.53** (0.78)	**1%**	4%	3%	**1%**	**1%**	5%
		rs1631931	0.94 (1.27)	1.18 (1.37)	1.09 (1.30)	**0.62** (0.73)	**0.60** (0.73)	4%	7%	5%	**0%**	**0%**	**2%**
TG_R	8	rs17091651	0.75 (0.99)	0.89 (1.22)	0.77 (1.16)	0.80 (0.89)	0.74 (0.89)	**2%**	4%	3%	**1%**	**1%**	**8%**
		rs4244457	**1.52** (1.74)	0.75 (1.20)	**0.66** (1.20)	**0.52** (0.98)	**0.49** (0.99)	**9%**	4%	4%	3%	3%	**14%**
	19	rs11083567	0.73 (1.03)	**0.66** (1.15)	**0.66** (1.16)	**0.50** (0.69)	**0.50** (0.69)	**2%**	**1%**	**1%**	**0%**	**0%**	**2%**
		rs1631931	0.97 (1.29)	0.85 (1.04)	0.79 (0.98)	0.75 (0.93)	0.74 (0.93)	**2%**	3%	**2%**	**1%**	**1%**	**2%**

* : Test d'association après avoir retiré les SNPs significatifs pour le test de stratification ; **: Proportion de SNPs significatifs pour le test de stratification avec p-nominal=5% ;

Les estimations empiriques des trois tests d'association donnent des résultats presque similaires que le SNP étudié soit lié ou non au polymorphisme fonctionnel. En général, les taux d'erreur sont plus petits ou proches du seuil nominal 5% pour les trois tests d'association excepté au SNP rs10215692 (chromosome 7) pour TG. De même, sur le chromosome 8, on observe un excès de résultats significatifs pour le test d'association QTDT au rs1631931 pour HDL (p-nom=12%) et rs4244457 pour TG (p-nom=14%). Les taux d'erreur restent inchangés selon que l'on tienne compte ou non de la covariable « Diet » (HDL *vs* HDL_Diet et TG *vs* TG_Diet). Comme attendu, enlever les SNPs significatifs pour le test de stratification diminue la moyenne de la statistique pour QTLD et MG. Après correction pour la normalité (TG_R), les erreurs de type 1 sont légèrement diminuées, en particulier pour QTDT et QTLD. De manière intéressante, on n'observe pas d'inflation de l'erreur de type 1 pour le trait non normalisé TG.

Les mêmes tendances sont observées pour les moyennes des statistiques de test. Elles sont relativement proches des valeurs attendues et restent stables selon que l'on ajuste ou non les résidus pour la covariable Diet.

Pour chacun de ces tests, les erreurs de type 1 sont proches des valeurs attendues. Les tableaux 34 et 35 en annexe (pages 164 et 165) montrent les estimations des effets alléliques (coefficients de régression β_w, β_w', β et les effets moyens inter-famille β_b, β_b' standardisés par la variance du trait) sous les deux hypothèses nulles de non association. Ces estimations restent similaires selon que l'on tienne compte ou non de la covariable Diet (HDL *vs* HDL_Diet et TG *vs* TG_Diet) et ne varient pas en fonction de la normalité du trait TG excepté pour QTLD. Les moyennes des effets alléliques du SNP considéré sont toutes proches de 0.

Hypothèse alternative

Le tableau 31 montre les estimations empiriques des moyennes ($\pm$ écarts-types) des statistiques du χ_1^2 lorsque le marqueur étudié est en association directe avec le polymorphisme fonctionnel. Nous montrons dans le tableau 32 les proportions de résultats significatifs pour des seuils nominaux de 1% et 0.1%. Le tableau 36 en annexe montre les estimations empiriques des effets alléliques sous l'hypothèse d'association. Les résultats sont reportés avec et sans les SNPs significatifs pour le test de stratification.

Tableau 31 : Moyenne des estimations des statistiques des tests d'association (± écart-type moyen) et des erreurs de type 1 (p-nom=5%) sous l'hypothèse d'association

Trait	Chr	SNP	QTDT	QTLD	QTLD|S *	MG	MG|S *	p-nominal=5%					
								QTDT	QTLD	QTLD|S *	MG	MG|S *	Strat **
HDL	8	α_4	17.88 (6.28)	30.65 (8.51)	27.18 (11.51)	30.96 (8.24)	27.88 (11.55)	100%	100%	91%	100%	91%	9%
	19	α_2	1.38 (1.35)	3.79 (2.42)	3.56 (2.59)	9.56 (4.29)	8.62 (5.05)	7%	49%	46%	92%	83%	11%
		δ_1	7.13 (3.80)	9.80 (4.42)	9.77 (4.48)	17.00 (5.79)	16.90 (6.00)	80%	95%	95%	100%	99%	1%
HDL Diet	8	α_4	17.87 (6.28)	30.64 (8.51)	27.17 (11.51)	30.96 (8.22)	27.88 (11.54)	100%	100%	91%	100%	91%	9%
	19	α_2	1.38 (1.35)	3.79 (2.41)	3.56 (2.59)	9.57 (4.28)	8.62 (5.05)	7%	49%	46%	92%	83%	11%
		δ_1	7.13 (3.80)	9.80 (4.42)	9.76 (4.48)	17.00 (5.79)	16.89 (6.00)	80%	95%	95%	100%	99%	1%
TG	8	α_4	2.21 (2.46)	6.18 (4.44)	5.88 (4.67)	10.93 (5.54)	9.92 (6.31)	16%	61%	59%	91%	83%	11%
	19	δ_1	3.11 (2.87)	8.28 (5.03)	7.97 (5.29)	12.91 (5.28)	12.13 (6.16)	28%	81%	78%	99%	92%	7%
TG Diet	8	α_4	2.21 (2.35)	6.29 (4.39)	6.01 (4.60)	11.11 (5.34)	10.19 (6.07)	16%	61%	59%	91%	83%	11%
	19	δ_1	3.15 (3.00)	8.35 (5.15)	8.04 (5.42)	13.05 (5.32)	12.28 (6.22)	32%	81%	77%	99%	92%	7%
TG_R	8	α_4	3.35 (3.16)	4.75 (4.05)	4.67 (4.11)	13.04 (5.85)	12.67 (6.34)	33%	47%	46%	98%	94%	4%
	19	δ_1	5.15 (3.58)	7.72 (4.56)	7.46 (4.64)	18.21 (5.89)	17.46 (7.04)	57%	77%	75%	100%	95%	5%

* : Test d'association après avoir retiré les SNPs significatifs pour le test de stratification.

** : Proportion de SNPs significatifs pour le test de stratification avec p-nominal=5%

Tableau 32 : Puissances pour des seuils nominaux de 5%, 1% et 0.1%.

Trait	Gène	p-nom =1%			p-nom=0.1%		
		QTDT	**QTLD|S***	**MG|S***	**QTDT**	**QTLD|S***	**MG|S***
HDL	α_4	100%	91%	91%	89%	91%	91%
	α_2	0%	13%	63%	0%	1%	34%
	δ_1	49%	73%	99%	15%	35%	92%
HDL_Diet	α_4	100%	91%	91%	89%	91%	91%
	α_2	0%	13%	63%	0%	1%	34%
	δ_1	49%	73%	99%	15%	35%	91%
TG	α_4	6%	41%	71%	2%	11%	43%
	δ_1	10%	53%	86%	2%	26%	52%
TG_Diet	α_4	6%	41%	71%	2%	11%	43%
	δ_1	10%	53%	85%	3%	26%	54%
TG-R	α_4	13%	21%	88%	10%	3%	58%
	δ_1	27%	50%	93%	11%	21%	86%

* : Test d'association après avoir retiré les SNPs significatifs pour le test de stratification.

Comme attendue, pour TG, la moyenne des estimations du χ_1^2 est légèrement plus élevée si le trait est ajusté pour la covariable Diet et aucune différence n'est observée pour HDL. Pour QTDT et MG, les moyennes sont plus élevées lorsque le trait est normal (TG *vs* TG_R). Ceci reste vrai si l'on prend en compte la stratification. La tendance inverse est observée pour QTLD.

D'une manière générale, pour un trait et un SNP donné, l'estimation de la moyenne de la statistique est toujours plus grande pour MG|S et toujours plus petite pour QTDT. Ces moyennes sont entre 1.5 et 6.2 fois plus grandes pour MG|S que pour QTDT. Pour QTLD|S, ce rapport est compris entre 1.0 et 2.4. Ces résultats sont consistants avec la quantité d'information utilisée. En effet, pour le test MG, le nombre effectif d'individus pris en compte dans l'estimation de β est le nombre total d'individus phénotypés et génotypés. Les deux tests d'association basés sur une décomposition orthogonale des scores génotypiques (QTDT et QTLD) n'utilisent qu'un sous effectif de cet échantillon. Seulement les individus (phénotypés et génotypés) ayant au moins un parent hétérozygote, si les deux parents sont typés, ou au moins deux individus (phénotypés et génotypés) de génotypes différents dans la fratrie si moins de deux parents sont génotypés, sont pris en compte dans l'estimation de la composante intra-famille. De plus, la méthode QTDT n'utilise que les individus apparentés pour l'estimation de la composante intra-famille alors que la méthode QTLD utilise tous les individus (apparentés et fondateurs). Nous donnons dans le tableau 33 les caractéristiques de la répartition des individus aux SNPs fonctionnels et le nombre effectif d'individus utilisés pour chacun de ces tests.

Tableau 33 : Caractéristiques de la répartition des individus aux SNPs fonctionnels et nombre d'individus informatif pour chacun des tests d'association.

Chr	Variant	# phéno	# géno	# phéno et géno	# fondateurs	QTDT*	QTLD*	MG**
8	α_4	6009	6130	5854	590	1 846 (**3.17**)	2 436 (**2.40**)	5854
19	α_2	6009	6275	5995	599	2 120 (**2.83**)	2 719 (**2.20**)	5995
19	δ_1	6009	6275	5995	599	2 240 (**2.68**)	2 839 (**2.11**)	5995

* : nombre effectif (Rapport de l'effectif utilisé par MG sur l'effectif du test considéré) d'individus pour l'estimation de la composante intra-famille.

** : nombre effectif d'individus pour l'estimation de la composante β.

Pour les trois variants fonctionnels, les tailles effectives des échantillons varient peu relativement à l'effectif de MG. La plus grande différence est observée en α_4 (ratio = 3.17

et 2.40 pour QTDT et QTLD) et la plus petite en δ_1 (ratio = 2.68 et 2.11 pour QTDT et QTLD).

On observe cependant des grandes différences dans l'estimation des moyennes des statistiques qui ne peuvent être expliquées que par les différences des tailles d'échantillons. Pour HDL, relativement à MG|S, les rapports des moyennes des statistiques des tests d'association sont plus élevés en δ_1 et α_2 qu'en α_4. Par exemple, pour QTLD|S, le ratio des moyennes des statistiques est de 2.42 (=8.62/3.56) en α_2 et de 1.73 (=16.90/9.77) en δ_1. De même, pour QTDT, ce rapport est plus élevé en α_2 (ratio=6.25=8.62/1.38) ou δ_1 (ratio=2.37=6.90/7.13) qu'en α_4 (ratio=1.56). Il faut noter que l'effet taille des deux SNPs α_4 et δ_1 est le même (Tableau 27, $\sigma^2_G(HDL) = 0.3\%$) avec une MAF similaire (21.7% et 24.9% pour α_4 et δ_1 respectivement).

Si on compare les moyennes des statistiques pour un même test, alors on observe également de très grandes variations dans les performances des modèles. Lorsque l'on teste l'association de HDL avec QTDT et QTLD|S, la moyenne du χ^2_1 est environ treize et huit fois plus grande en α_4 qu'en α_2 ; la part de variance attribuable aux variants fonctionnels varie dans le même sens ($\sigma^2_{\alpha_4}(HDL)$=0.3% et $\sigma^2_{\alpha_2}(HDL)$=0.2%), au contraire, la taille effective des individus varie en sens opposé (1 846 vs 2 120 en α_4 vs α_2). Pour les deux variants fonctionnels liés mais non associés du chromosome 19 (α_2, δ_1), les tailles effectives sont similaires (ratio=1.04) et même si les différences des moyennes des statistiques sont moins marquées, les moyennes restent toujours plus élevées en δ_1 qu'en α_2 (ratio=5.17, 2.74, 1.96 pour QTDT, QTLD|S et MG|S respectivement).

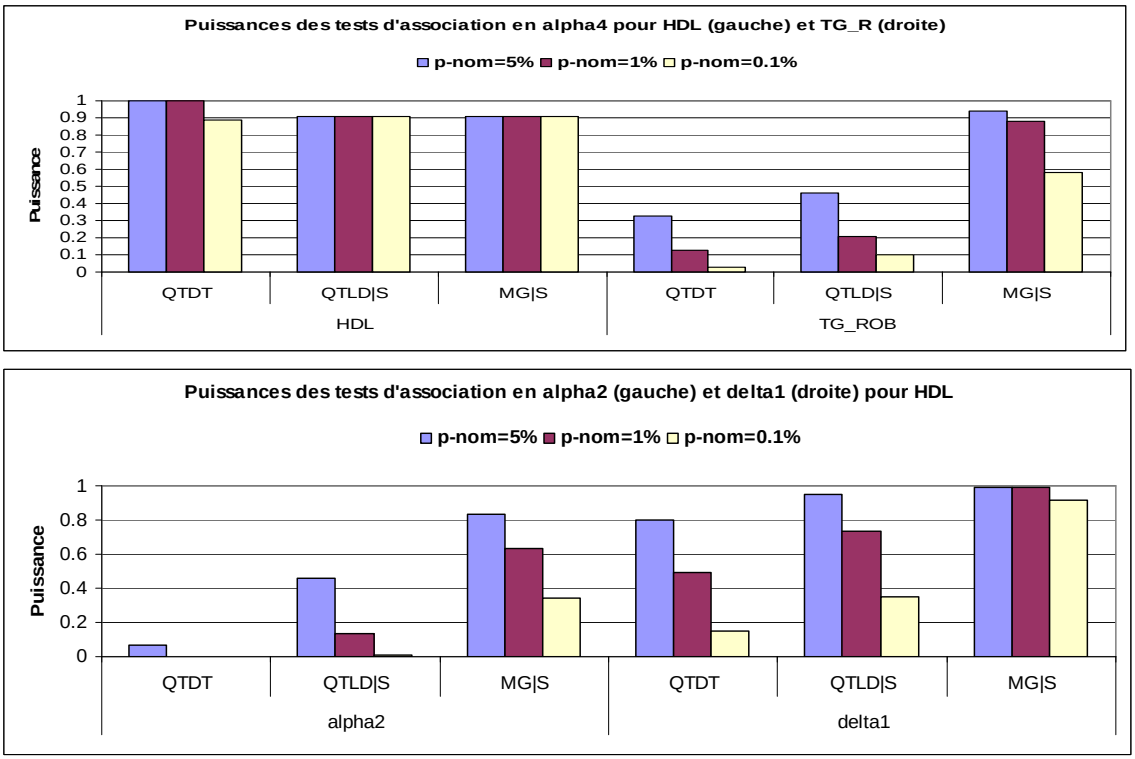
Pour TG, les estimations des moyennes des statistiques en α_4 et δ_1 varient dans le même sens que les tailles effectives d'échantillons (Tableau 33) mais dans le sens inverse de l'effet taille du gène ($\sigma^2_{\alpha_4}(TG) = 0.4\%$ et $\sigma^2_{\delta_1}(TG) = 0.3\%$). L'impact du traitement médicamenteux en δ_1 est relativement mineur, seulement 2% de la population répond au traitement pour HDL et TG.

Le tableau 36 en annexe montre les moyennes des estimations des effets alléliques obtenues pour les trois analyses d'association. L'effet allélique au variant α_4 pour HDL est nettement supérieure aux effets observés aux autres variants fonctionnels. Le sens des estimations des effets alléliques (coefficients de régression) est consistent pour les trois tests d'association et chaque méthode donne en moyenne les mêmes estimations.

Lorsque le seuil nominal diminu, les niveaux de puissance restent bons uniquement avec α_4 pour le trait HDL et MG|S (Tableau 32). Pour QTDT et QTLD|S, les niveaux de puissance diminuent de manière dramatique, il est par exemple impossible de détecter le variant α_2 pour le trait HDL par le test QTDT pour des seuils de 1% et 0.1%. Pour TG, la

puissance reste relativement correct pour δ_1 seulement et avec TG normalisé (TG_R). Les puissances sont illustrées dans la figure 24 pour le variant causal α_4 avec HDL et TG_R et pour les variants α_2 et δ_1 avec HDL.

Figure 24 : Puissance pour des seuils nominaux de 5%, 1% et 0.1% pour HDL et TG_R en α_4 (en haut) et en α_2/δ_1 pour HDL (en bas)



4.4. Conclusions de l'étude GAW16

En conclusion, les résultats montrent que le modèle du Measured Genotype (MG) est plus performant que les modèles d'association utilisant une décomposition orthogonale des scores génotypiques (QTDT et QTLD), même en tenant compte des résultats de stratification. QTDT, quant à lui est le moins performant des trois. Ces résultats sont consistants avec les résultats de simulation obtenus par deux autres études (Havill, Dyer et al. 2005; Aulchenko, de Koning et al. 2007).

Ces résultats montrent aussi que les trois modèles sont similaires dans l'estimation de l'erreur de type 1. Comme cela a été mentionné par Havill (Havill, Dyer et al. 2005),

l'erreur de type 1 de QTDT et QTLD n'est pas augmentée sous l'hypothèse nulle d'absence d'association en présence de liaison. Cependant, il faut noter que les tests sont ici réalisés dans des échantillons de familles de très grande taille. Bien que l'effet gène du variant causal soit très petit (entre 0.2% et 0.4%), les trois tests ont une assez bonne puissance (>90% pour un seuil nominal de 0.1%) de détecter l'association directe pour HDL et deux des trois variants causaux α_4 et δ_1.

Lorsque le seuil nominal diminue, la puissance des tests reste bonne uniquement pour MG. Pour ces trois modèles d'association, la puissance est la plus petite pour le variant fonctionnel avec le plus petit effet taille (α_2) et pour le trait le moins héritable (TG ; h^2=30%). On remarque aussi que les performances de ces modèles d'association diffèrent grandement pour les deux polymorphismes fonctionnels liés mais non associés du chromosome 19 (α_2 et δ_1). La puissance en δ_1 est bien meilleure qu'en α_2 alors que les parts de variance sont peu différentes (Saint Pierre, Vitezica et al. 2009).

Conclusions et perspectives

Comparer différentes méthodes d'analyse génétique pour rechercher des loci potentiellement impliqués dans la variation de traits complexes corrélés a été le fil conducteur de ce travail. Dans le cadre de l'étude des maladies monogéniques, un certains nombres de méthodes on été mises au point et se sont revelés efficaces. Avec les importants progrès des puces à ADN, qui ont permis de recueillir des quantités énormes de données génotypiques, de nouvelles approches pour localiser et identifier des loci génétiques impliqués dans des maladies multifactorielles ont été développées. La composante génétique pour ces maladies est en général incontestable, mais la contribution des loci sur la variation du trait est souvent très faible, ce qui nécessite d'utiliser des méthodes puissantes pour avoir une bonne capacité de détection des loci génétiques impliqués dans la variation phénotypique.

Les progrès de la médecine générale ont permis de mieux comprendre les maladies et grâce au développement de l'ingénierie médicale on a pu relier les pathologies à plusieurs mesures phénotypiques. Cela a conduit à développer des méthodes d'analyses jointes de traits corrélés, qui permettent d'analyser ces données simultanément. L'objectif de ce travail a été de comparer la puissance de plusieurs méthodes pour l'analyse de traits multivariées avec d'autres méthodes, classiquement utilisées dans un cadre univariée, tant du point de vue de l'information apportée par les marqueurs au niveau de la liaison et de l'association.

Ce travail de comparaison s'est d'abord effectué dans le cadre de la base de données familiale du projet collaboratif NEMO, dont l'objectif est de caractériser la composante génétique de la densité osseuse. L'originalité de cet échantillon est que ces familles ont été sélectionnées par des hommes ayant des valeurs basses de la densité osseuse aux sites squelettiques du rachis lombaire (LS) ou du col du fémur (FN). Dans le même contexte, un nouvel échantillon de sujets non apparentés a été constitué regroupant des hommes sélectionnés pour leurs valeurs phénotypiques. Par ailleurs, dans le cadre du congrès GAW16 nous avons comparé plusieurs méthodes d'analyse dans des données familiales, pour des traits quantitatifs mimant au mieux des maladies cardiovasculaires.

La première étape de l'étude NEMO a été la recherche de QTL en utilisant l'information apportée par le marqueur au niveau de la liaison génétique par des méthodes de type « non paramétrique », basées sur le modèle de décomposition de la variance (VC). Nous avons comparé deux approches pour rechercher des QTLs à effets pléiotropiques sur la variation des traits. La première approche repose sur une simple extension du modèle VC

univarié au modèle bivarié pour l'analyse jointe de traits corrélés. Une méthode classique est de contraindre le paramètre de corrélation génétique aux bornes de l'espace des valeurs possibles. La deuxième approche repose sur des combinaisons linéaires des traits d'intérêts LS et FN, obtenus par des méthodes de réduction de la dimension telles que l'analyse en composantes principales (ACP). Notre objectif était d'évaluer le gain apporté par ces deux approches relativement aux analyses de liaisons univariées, qui ignorent les corrélations entre les traits. La plupart du temps, la localisation de gènes à effets pléiotropiques est basée sur la consistance des régions de liaison identifiées d'une part, par l'analyse de liaison du premier axe de variation d'une ACP et d'autre part, par l'analyse univariée des traits d'intérêts LS ou FN. Nous avons donc également comparés les résultats avec ceux obtenus par les deux premiers axes de variation d'une ACP.

Quelque soit l'approche utilisée, dans nos données, les régions de liaison localisées sont relativement similaires. Si en théorie les analyses de liaison bivariées peuvent améliorer la détection de QTLs à effets pléiotropiques, en pratique, le plus grand gain du bivarié est obtenu lorsque la corrélation induite par les effets génétiques et environnementaux est négative ce qui n'est pas le cas dans notre étude. Utiliser des méthodes sophistiquées peut parfois être un plus, mais dans les scénarios étudiés pour la recherche de loci par la liaison, l'ensemble des méthodes évaluées donnent des résultats à peu près équivalents.

Ce travail a été étendu à l'identification de QTL par la recherche de l'association génétique pour la variation de la DMO. Nous avons comparé le modèle d'association bivarié basé sur le modèle SUR (Seemingly Unrelated Regression). La particularité de ce modèle est de permettre des effets génétiques différents sur chacun des traits.

Les résultats ont montré que pour certaines des régions identifiées, la signification de l'association de l'analyse bivariée est bien meilleure que celle de l'analyse univariée. L'analyse d'association jointe semble donc être une approche intéressante pour les études à grande échelle. Ceci nous a amené à développer une étude de simulation dans des échantillons de sujets non apparentés et recensés selon des critères variables vis-à-vis de la valeur des phénotypes étudiés. Nous avons donc évalué les performances relatives des tests d'association univariées et bivariés pour différents modèles génétiques et différents modes de sélection des sujets. Les résultats ont montré que la méthode d'association bivariée, basée sur le modèle SUR, étaient au moins aussi puissante et souvent meilleure, que l'analyse univariée, même quand le QTL n'exerçait pas d'effets pléiotropiques sur les traits.

On peut noter qu'aucun variant localisé dans les régions de liaison identifiées par le criblage du génome n'a été retrouvé significativement associé à la variation de la DMO par le criblage du génome pour l'association. Plusieurs raisons sont possibles. Bien que deux stratégies complémentaires ont été réalisées afin d'augmenter la puissance de l'étude, c'est-à-dire, sélection des individus aux extrêmes des valeurs de la distribution dans la population générale et analyse jointe des traits corrélés LS et FN, pour des questions évidentes de coût, l'étude d'association a été conduite dans un échantillon

relativement petit. On pourrait également envisager que dans ces régions de liaison, des allèles distincts au même loci (hétérogénéité allélique) ou sur différents loci (hétérogénéité non allélique), donnent lieu indépendamment à une augmentation ou à une diminution de la variation de la DMO. Contrairement à la recherche de QTL utilisant l'information apportée les marqueurs au niveau de la liaison, l'hétérogénéité allélique peut diminuer les performances de détecter une association génétique entre le locus du variant causal et le locus marqueur. Les variants causaux contenus dans les régions, identifiées par la liaison, peuvent donc ne pas être détectés par l'association.

En ce qui concerne la comparaison des diverses méthodes utilisées dans le cadre du congrès GAW16, nous avons évalués trois méthodes classiques d'association pour l'analyse de traits quantitatifs dans des données familiales. Ces approches étaient le modèle du Measured Genotype (MG), reposant sur un modèle linéaire mixte dans lequel les relations entre apparentés sont pris en compte dans la matrice des effets polygéniques. La deuxième approche reposait sur une décomposition orthogonale des scores génotypiques, le test du QTDT (*Quantitative Trait Disequilibrium Test*) et son extension, le QTLD (*Quantitative Trait Linkage Disequilibrium*). Nos résultats ont montré que le modèle MG est plus performant que les modèles d'association utilisant une décomposition orthogonale des scores génotypiques (QTDT et QTLD), même en tenant compte de la stratification de population. En effet, si ces trois modèles sont similaires dans leurs erreurs de type 1, nous avons montré que la puissance est bonne uniquement pour le modèle MG.

Nous avons pu remarquer tout au long de ce travail que l'analyse de traits multifactoriels pose un problème d'ordre méthodologique. Au regard des effets modestes que l'on cherche à détecter et du nombre de marqueurs testés, il est important que l'analyse soit menée avec précaution quant aux possibles biais qui peuvent affecter la qualité des résultats. Par exemple, les erreurs de génotypages ou la stratification éventuelle de la population sont autant de facteurs à prendre en considération afin de rendre compte de la validité et de la fiabilité des résultats obtenus.

Au cours de ce travail nous avons souvent été limités par le temps pour mener à bien toutes les comparaisons possibles entre les diverses méthodes. Nous n'avons pas évalué la puissance des tests de liaison bivariés dans nos données NEMO, de futures investigations seraient nécessaires pour comparer les performances empiriques relatives de ces tests de liaison. De même, à partir des données familiales GAW16, il serait intéressant de comparer la puissance des analyses d'association utilisant des échantillons de familles relativement à des échantillons de sujets non apparentés.

Si on veut progresser dans la connaissance des origines génétiques des maladies multifactorielles, il est important d'utiliser des méthodes optimales. Au carrefour entre médecine et statistique, l'évaluation des diverses méthodes utilisées en génétique statistique reste un chantier grand ouvert.

Annexe 1 : Figures et tableaux

Figure 25 : Vue d'un os en coupe

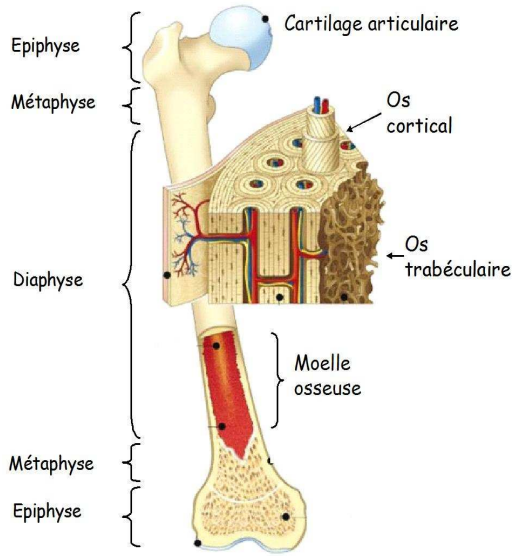


Figure 26 : Distribution de la masse osseuse en fonction de l'âge chez la femme et chez l'homme

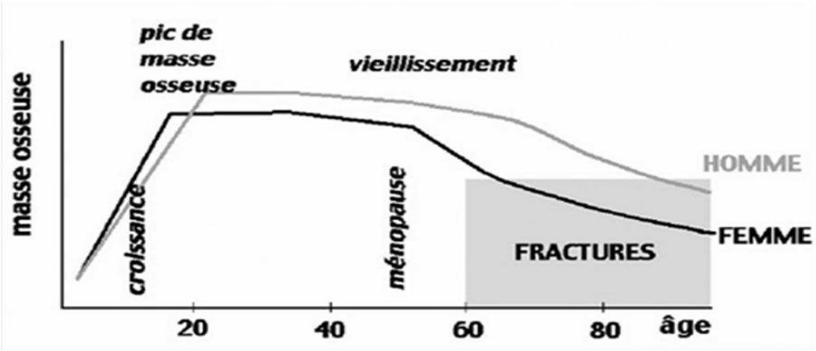
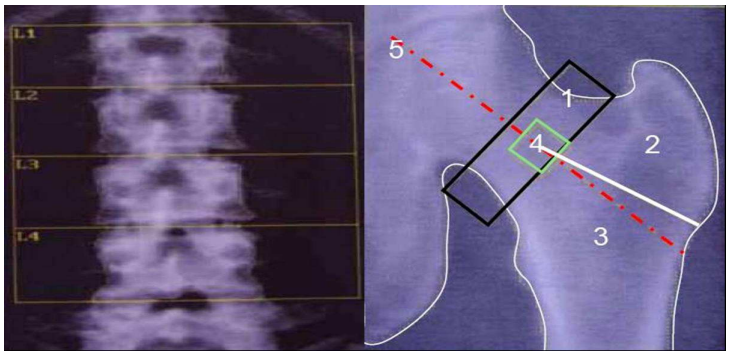
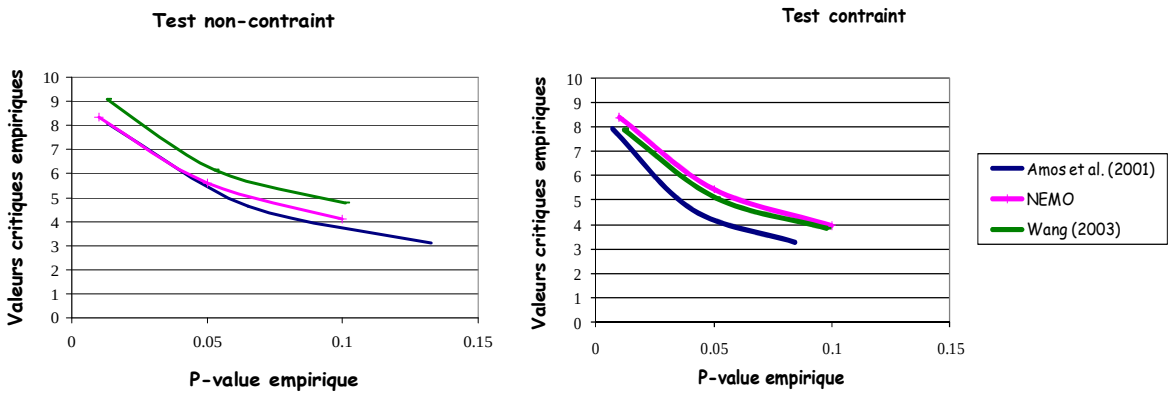


Figure 27 : Zones de mesure de la DMO aux lombaires (gauche) et à la hanche (droite)



Sites de mesure de la DMO à la hanche. n°1 : col du fémur ; n°2 : trochanter ; n°3 : région inter-trochantérienne ; n°4 : Ward ; n°5 : longueur de l'axe du col.

Figure 28 : Valeurs critiques en fonction des seuils empiriques des tests de liaison bivariées



Pour le test Non contraint, Amos suppose le mélange C $(1/4\,\chi_0^2 + 1/2\,\chi_1^2 + 1/4\,\chi_3^2)$; Wang suppose le mélange A $(1/4\,\chi_1^2 + 1/2\,\chi_2^2 + 1/4\,\chi_3^2)$. Pour le test Contraint, Amos suppose le mélange D $(1/4\,\chi_0^2 + 1/2\,\chi_1^2 + 1/4\,\chi_2^2)$ et Wang suppose le mélange E $\left(1/2\,\chi_1^2 + 1/2\,\chi_2^2\right)$. Valeurs telles que reportées pour un échantillon de 100 familles à 4 enfants (600 individus). L'héritabilité polygénique est posée à 67% pour chacun des traits. 1 000 réplicats sont générés sous l'hypothèse nulle pour Amos contre 10 000 pour Wang.

Figure 29 : Histogrammes des distributions des covariables âge et IMC dans les échantillons sélectionnés pour des valeurs basses (TG) ou hautes (TD).

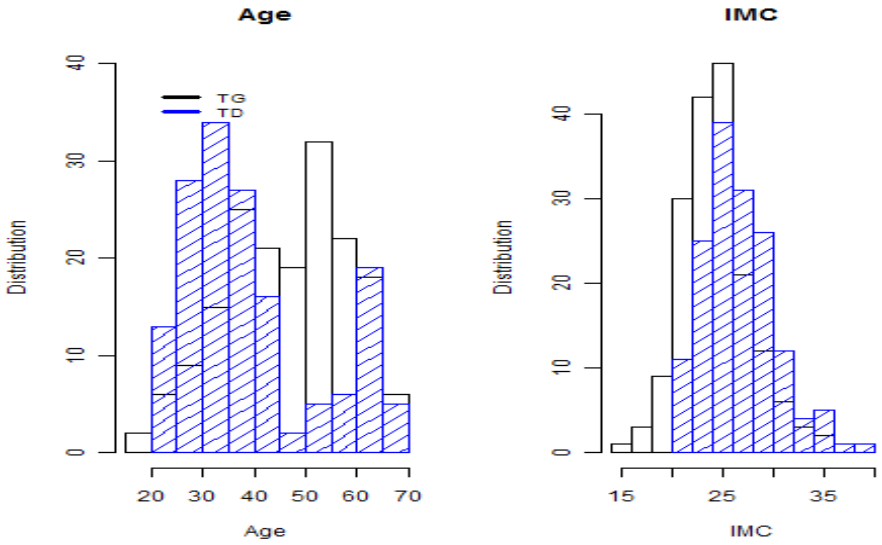
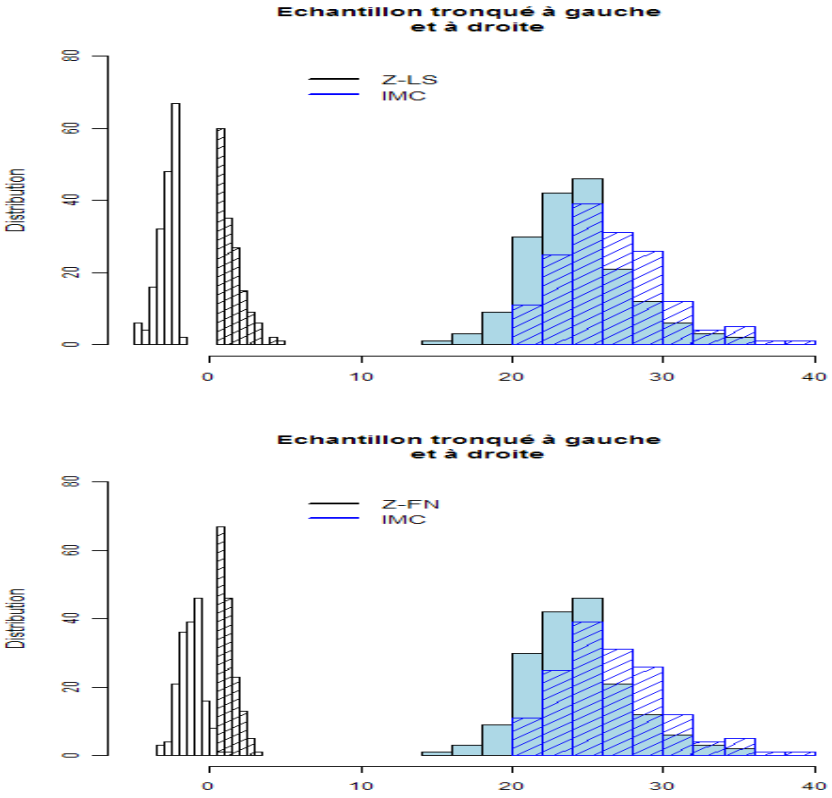
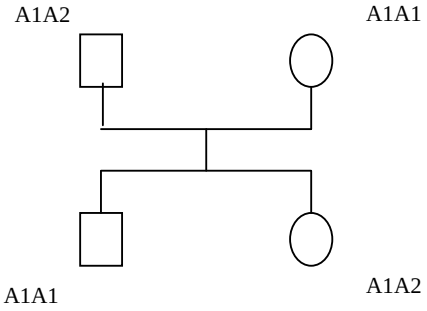


Figure 30 : Histogrammes des distributions de la covariable IMC et Z-LS (en haut) et IMC et Z-FN (en bas) dans les échantillons sélectionnés pour des valeurs basses ou hautes (zone hachurée).



Exemple du modèle VC pour une famille à deux enfants

Supposons la famille ci-dessous :



Si je n'ai aucune information a priori sur les allèles des génotypes au marqueur, les proportions d'allèles identiques par descendance (Π_{jl}) au marqueur entre les individus j et l (j=1,..,4 et l=1,..,4) de la famille sont égales aux proportions attendues a priori. Ces proportions dépendent seulement des liens d'apparentement Φ_{jl} entre les individus j et l et est de la forme :

$$2\Phi = \begin{pmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1/2 & 1 \end{pmatrix} \quad \begin{matrix} \leftarrow\text{Parent 1} \\ \leftarrow\text{Parent 2} \\ \leftarrow\text{Enfant 1} \\ \leftarrow\text{Enfant 2} \end{matrix}$$

Où chaque ligne et chaque colonne de la matrice représentent un individu de la famille. Les éléments de la diagonale sont les proportions partagées d'allèle IBD d'un individu avec lui-même.

Nous supposons que la liaison entre le QTL et le marqueur est complète (θ=0). A partir des génotypes observés au marqueur A (A1/A2), nous pouvons calculer les proportions du nombre d'allèles partagés par descendance entre les paires d'individus j et l : π_{jl}. Les époux sont supposés être non apparentés. Quelque soit le génotype observé au marqueur, ils ne partagent aucun allèle identique par descendance (IBD=0). Un individu avec lui-même partage tous ses allèles IBD (IBD=2). Par contre, il existe une ambiguïté sur l'IBD pour la paire de germains. Les enfants 1 (e1, génotype A1A1) et 2 (e2, génotype A1A2) peuvent partager 0 ou 1 allèle identique par descendance selon que l'allèle A1 transmis par la mère soit le même pour les deux enfants. Admettons que la mère est transmise le même allèle A1 alors la paire de germains e1 et e2 partage 1 allèle en commun (IBD=1) et $\pi_1 = \Pi_{e1,e2} = 1/2$. Si au contraire, elle ne transmet pas le même allèle A1, les deux frères

partagent 0 allèle en commun (IBD=0) et $\pi_0 = \Pi_{e1e2} = 0$. Ces deux évènements sont équiprobables (1/2) sans aucune autre information. Donc en moyenne la proportion d'allèles identiques par descendance entre les deux frères est $\Pi_{e1,e2} = \frac{1}{2} \times \pi_0 + \frac{1}{2} \times \pi_1 = \frac{1}{4}$.

La matrice de variance phénotypique pour cette famille s'écrit alors :

$$\Omega = \sigma_A^2 \times \begin{pmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 1/4 \\ 1/2 & 1/2 & 1/4 & 1 \end{pmatrix} + \sigma_C^2 \times \begin{pmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1/2 & 1 \end{pmatrix} + \sigma^2 \times \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Les effets polygéniques sont basés sur la ressemblance génétique théorique tandis que les effets du QTL reposent sur la ressemblance génétique au sein de chaque famille échantillonnée.

Tableau 34 : Moyenne des estimations des effets alléliques (± écart-type moyen) sous l'hypothèse d'absence d'association et absence de liaison

Trait	Chr	SNP	QTDT		QTLD		QTLD|S *		MG	MG|S *
			β_b	β_w	β'_b	β'_w	β'_b	β'_w	β	β
HDL	7	rs2521760	-0.03 (0.03)	0.00 (0.03)	-0.01 (0.03)	-0.02 (0.02)	-0.01 (0.03)	-0.02 (0.02)	-0.02 (0.02)	-0.02 (0.02)
		rs10215692	-0.02 (0.03)	0.04 (0.03)	0.00 (0.03)	0.01 (0.03)	0.00 (0.03)	0.01 (0.02)	0.00 (0.02)	0.00 (0.02)
HDL_Diet	7	rs2521760	-0.03 (0.03)	0.00 (0.03)	-0.01 (0.03)	-0.02 (0.02)	-0.01 (0.03)	-0.02 (0.02)	-0.02 (0.02)	-0.02 (0.02)
		rs10215692	-0.02 (0.03)	0.04 (0.03)	0.00 (0.03)	0.01 (0.03)	0.00 (0.03)	0.01 (0.02)	0.00 (0.02)	0.00 (0.02)
TG	7	rs2521760	-0.04 (0.02)	0.03 (0.04)	-0.01 (0.02)	-0.02 (0.03)	-0.01 (0.02)	-0.02 (0.03)	-0.01 (0.02)	-0.01 (0.02)
		rs10215692	-0.02 (0.03)	-0.04 (0.05)	-0.01 (0.03)	-0.04 (0.04)	-0.01 (0.03)	-0.04 (0.04)	-0.03 (0.03)	-0.02 (0.03)
TG_Diet	7	rs2521760	-0.04 (0.02)	0.03 (0.04)	-0.01 (0.02)	-0.02 (0.03)	-0.01 (0.02)	-0.02 (0.03)	-0.01 (0.02)	-0.01 (0.02)
		rs10215692	-0.02 (0.03)	-0.04 (0.05)	-0.01 (0.03)	-0.04 (0.04)	-0.01 (0.03)	-0.04 (0.04)	-0.03 (0.03)	-0.03 (0.03)
TG_R	7	rs2521760	-0.03 (0.02)	0.03 (0.04)	-0.03 (0.02)	0.02 (0.03)	-0.02 (0.02)	0.01 (0.03)	-0.01 (0.02)	-0.01 (0.02)
		rs2521760	-0.02 (0.03)	-0.04 (0.05)	-0.05 (0.03)	0.00 (0.04)	-0.04 (0.03)	0.00 (0.04)	-0.02 (0.03)	-0.02 (0.03)

* : Test d'association après avoir retiré les SNPs significatifs pour le test de stratification.

Tableau 35 : Moyenne des estimations des effets alléliques (± écart-type moyen) sous l'hypothèse d'absence d'association en présence de liaison

Trait	Chr	SNP	QTDT		QTLD		QTLD|S *		MG	MG|S *
			β_b	β_w	β'_b	β'_w	β'_b	β'_w	β	β
HDL	8	rs17091651	-0.01 (0.03)	-0.04 (0.04)	-0.01 (0.02)	-0.03 (0.03)	-0.01 (0.02)	-0.03 (0.03)	-0.02 (0.02)	-0.02 (0.02)
		rs4244457	-0.03 (0.02)	0.03 (0.03)	-0.02 (0.02)	0.01 (0.02)	-0.02 (0.02)	0.01 (0.02)	0.00 (0.02)	0.00 (0.01)
	19	rs11083567	0.02 (0.02)	-0.02 (0.03)	0.02 (0.02)	-0.01 (0.02)	0.01 (0.02)	-0.01 (0.02)	0.00 (0.02)	0.00 (0.02)
		rs1631931	0.00 (0.03)	0.05 (0.03)	0.03 (0.02)	0.02 (0.02)	0.03 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)
HDL_Diet	8	rs17091651	-0.01 (0.03)	-0.04 (0.04)	-0.01 (0.02)	-0.03 (0.03)	-0.01 (0.02)	-0.03 (0.03)	-0.02 (0.02)	-0.02 (0.02)
		rs4244457	-0.03 (0.02)	0.03 (0.03)	-0.02 (0.02)	0.01 (0.02)	-0.02 (0.02)	0.01 (0.02)	0.00 (0.02)	0.00 (0.01)
	19	rs11083567	0.02 (0.02)	-0.02 (0.03)	0.02 (0.02)	-0.01 (0.02)	0.01 (0.02)	-0.01 (0.02)	0.00 (0.02)	0.00 (0.02)
		rs1631931	0.00 (0.03)	0.05 (0.03)	0.03 (0.02)	0.02 (0.02)	0.03 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)
TG	8	rs17091651	-0.05 (0.03)	0.02 (0.04)	-0.02 (0.03)	-0.02 (0.03)	-0.02 (0.03)	-0.02 (0.03)	-0.02 (0.02)	-0.02 (0.02)
		rs4244457	0.01 (0.02)	-0.04 (0.03)	0.01 (0.02)	-0.02 (0.02)	0.00 (0.02)	-0.02 (0.02)	-0.01 (0.02)	-0.01 (0.01)
	19	rs11083567	0.01 (0.02)	-0.01 (0.03)	0.03 (0.02)	-0.02 (0.03)	0.03 (0.02)	-0.02 (0.03)	0.01 (0.02)	0.00 (0.02)
		rs1631931	0.01 (0.03)	-0.01 (0.04)	0.02 (0.03)	-0.02 (0.03)	0.02 (0.02)	-0.02 (0.03)	0.00 (0.02)	0.00 (0.02)
TG_Diet	8	rs17091651	-0.05 (0.03)	0.03 (0.04)	-0.02 (0.03)	-0.02 (0.03)	-0.02 (0.03)	-0.02 (0.03)	-0.02 (0.02)	-0.02 (0.02)
		rs4244457	0.01 (0.02)	-0.04 (0.03)	0.01 (0.02)	-0.02 (0.02)	0.00 (0.02)	-0.02 (0.02)	-0.01 (0.02)	-0.01 (0.01)
	19	rs11083567	0.01 (0.02)	-0.01 (0.03)	0.03 (0.02)	-0.02 (0.03)	0.03 (0.02)	-0.02 (0.03)	0.00 (0.02)	0.00 (0.02)
		rs1631931	0.01 (0.03)	-0.01 (0.04)	0.02 (0.03)	-0.02 (0.03)	0.02 (0.02)	-0.02 (0.03)	0.00 (0.02)	0.00 (0.02)
TG_R	8	rs17091651	-0.05 (0.03)	0.03 (0.04)	-0.05 (0.03)	0.02 (0.03)	-0.05 (0.03)	0.02 (0.03)	-0.02 (0.02)	-0.02 (0.02)
		rs4244457	0.02 (0.02)	-0.03 (0.03)	0.00 (0.02)	0.00 (0.03)	0.00 (0.02)	0.00 (0.02)	0.00 (0.02)	0.00 (0.01)
	19	rs11083567	0.00 (0.02)	-0.01 (0.03)	-0.01 (0.02)	0.00 (0.03)	-0.01 (0.02)	0.00 (0.03)	0.00 (0.02)	0.00 (0.02)
		rs1631931	-0.01 (0.02)	-0.02 (0.04)	-0.03 (0.03)	0.00 (0.03)	-0.03 (0.03)	0.00 (0.03)	-0.01 (0.02)	-0.01 (0.02)

* : Test d'association après avoir retiré les SNPs significatifs pour le test de stratification.

Tableau 36 : Moyenne des estimations des effets alléliques (± écart-type moyen) sous l'hypothèse d'association

Trait	Chr	Gène	QTDT		QTLD		QTLD|S *		MG	MG|S *
			β_b	β_w	β'_b	β'_w	β'_b	β'_w	β	β
HDL	8	α_4	-0.12 (0.02)	-0.16 (0.03)	-0.10 (0.02)	-0.17 (0.02)	-0.09 (0.03)	-0.15 (0.05)	-0.13 (0.02)	-0.12 (0.04)
	19	α_2	0.09 (0.02)	0.03 (0.03)	0.08 (0.02)	0.05 (0.02)	0.07 (0.03)	0.05 (0.03)	0.07 (0.02)	0.06 (0.03)
		δ_1	0.10 (0.02)	0.09 (0.03)	0.10 (0.02)	0.09 (0.02)	0.10 (0.02)	0.09 (0.02)	0.09 (0.02)	0.09 (0.02)
HDL_Diet	8	α_4	-0.12 (0.02)	-0.16 (0.03)	-0.10 (0.02)	-0.17 (0.02)	-0.09 (0.03)	-0.15 (0.05)	-0.13 (0.02)	-0.12 (0.04)
	19	α_2	0.09 (0.02)	0.03 (0.03)	0.08 (0.02)	0.05 (0.02)	0.07 (0.03)	0.05 (0.03)	0.07 (0.02)	0.06 (0.03)
		δ_1	0.10 (0.02)	0.09 (0.03)	0.10 (0.02)	0.09 (0.02)	0.10 (0.02)	0.09 (0.02)	0.09 (0.02)	0.09 (0.02)
TG	8	α_4	-0.09 (0.02)	-0.05 (0.04)	-0.08 (0.02)	-0.08 (0.03)	-0.07 (0.03)	-0.07 (0.04)	-0.08 (0.02)	-0.07 (0.03)
	19	δ_1	-0.09 (0.02)	-0.06 (0.03)	-0.08 (0.02)	-0.08 (0.03)	-0.07 (0.03)	-0.08 (0.03)	-0.08 (0.02)	-0.07 (0.03)
TG_Diet	8	α_4	-0.09 (0.02)	-0.05 (0.04)	-0.08 (0.02)	-0.08 (0.03)	-0.07 (0.03)	-0.07 (0.04)	-0.08 (0.02)	-0.07 (0.03)
	19	δ_1	-0.09 (0.02)	-0.06 (0.03)	-0.08 (0.02)	-0.08 (0.03)	-0.07 (0.03)	-0.08 (0.03)	-0.08 (0.02)	-0.08 (0.03)
TG_R	8	α_4	-0.09 (0.02)	-0.06 (0.04)	-0.10 (0.02)	-0.06 (0.03)	-0.10 (0.03)	-0.06 (0.03)	-0.08 (0.02)	-0.08 (0.03)
	19	δ_1	-0.11 (0.02)	-0.08 (0.03)	-0.11 (0.02)	-0.08 (0.03)	-0.11 (0.03)	-0.08 (0.03)	-0.10 (0.02)	-0.09 (0.03)

* : Test d'association après avoir retiré les SNPs significatifs pour le test de stratification

Annexe 2 : Articles publiés et en révision

Articles publiés

Saint-Pierre A., Vitezica Z., Martinez M. (2009). *A comparative study of three methods for detecting association of quantitative traits in samples of related subjects.* BMC Proc., 7:122-128

Kaufman J.M., Ostertag A., **Saint-Pierre A**., Cohen-Solal M., Boland A., Van Pottelbergh I., Toye K., de Vernejoul M.C., Martinez M. (2008). *Genome-Wide linkage screen of bone mass density in European pedigrees ascertained through a male relative with low BMD values: Evidence for QTLs on 17q21-23, 11q12-13, 22q11 and 13q12-14*. J. Clin. Endocrinol. Metab., 93(10) :3755-62.

Saad M., Lesage S., **Saint-Pierre A**., Corvol J.C., Zelenika D., Lambert J.C., Vidailhet M., Mellick G.D., Lohmann E., Durif F., Pollak P., Damier P., Tison F., Silburn P.A., Tzourio C., Forlani S., Loriot M.A., Giroud M., Helmer C., Portet F., Amouyel P., Lathrop M., Elbaz A., Durr A., Martinez M. and Brice A (2010). *Genome-wide association study confirms BST1 and identifies a new locus on 12q24 as risk loci for Parkinson's disease in the European population*. Hum. Mol. Genet.

Article en cours de révision

Saint-Pierre A., Kaufman J.M., Ostertag A., Toye K., Zelenika D., Cohen-Solal M., Lathrop M., de Vernejoul M.C., M. Martinez (2010). *Bivariate association analysis for quantitative traits in unrelated subjects: Performance under varying ascertainment schemes and application to a genome-wide association study of two BMD phenotypes in males with high or low BMD*. Eur. J. Hum. Genet.

BMC Proceedings

BioMed Central

Proceedings

Open Access

A comparative study of three methods for detecting association of quantitative traits in samples of related subjects

Aude Saint Pierre*, Zulma Vitezica and Maria Martinez

Address: INSERM, UL563, University Paul-Sabatier, CPTP, Toulouse F-31300, France

E-mail: Aude Saint Pierre* - Aude.saint-pierre@inserm.fr; Zulma Vitezica - Zulma.vitezica@inserm.fr; Maria Martinez - Maria.martinez@inserm.fr

*Corresponding author

from Genetic Analysis Workshop 16
St Louis, MO, USA 17-20 September 2009

Published: 15 December 2009

BMC Proceedings 2009, 3(Suppl 7):S122 doi: 10.1186/1753-6561-3-S7-S122

This article is available from: http://www.biomedcentral.com/1753-6561/3/S7/S122

Abstract

We used Genetic Analysis Workshop 16 Problem 3 Framingham Heart Study simulated data set to compare methods for association analysis of quantitative traits in related individuals. More specifically, we investigated type I error and relative power of three approaches: the measured genotype, the quantitative transmission-disequilibrium test (QTDT), and the quantitative trait linkage-disequilibrium (QTLD) tests. We studied high-density lipoprotein and triglyceride (TG) lipid variables, as measured at Visit 1. Knowing the answers, we selected three true major genes for high-density lipoprotein and/or TG. Empirical distributions of the three association models were derived from the first 100 replicates. In these data, all three models were similar in error rates. Across the three association models, the power was the lowest for the functional SNP with smallest size effects (i.e., $\alpha 2$), and for the less heritable trait (i.e., TG). Our results showed that measured genotype outperformed the two orthogonal-based association models (QTLD, QTDT), even after accounting for population stratification. QTDT had the lowest power rates. This is consistent with the amount of marker and trait data used by each association model. While the effective sample sizes varied little across our tested variants, we observed some large power drops and marked differences in performances of the models. We found that the performances contrasted the most for the tightly linked, but not associated, functional variants.

Background

For pedigree-based association analysis, several methods have been developed that utilize information about transmission of alleles, such as the orthogonal test for within-family variation (quantitative transmission-disequilibrium test, or QTDT) [1,2]. The quantitative trait linkage-disequilibrium test (QTLD) is a modification of the QTDT method that assigns the founder genotypes to the within-family component rather than to the between-family component [3]. The measured genotype

BMC Proceedings 2009, 3(Suppl 7):S122 http://www.biomedcentral.com/1753-6561/3/S7/S122

(MG) model is a simple fixed-effects regression for which non-independence in the data is accounted for by polygenic effects [4,5]. All three approaches, QTDT, QTLD, and MG, can be applied to the association analysis of quantitative traits in extended pedigrees. They differ in the amount and type of marker information used for testing association. The MG model uses all individuals with available phenotype and genotype data. The family-based models use a subset of this sample. The effective sample size of QTDT is further reduced because founders and spouses are not use to estimate the within-component effect. Thus, QTDT may lack of power compared with QTLD and/or MG but, on the other hand, both MG and QTLD tests may be affected by allelic association due to population stratification. The relative merit of these approaches has been investigated in a few instances [3,6]. Here, we extend these studies to explore type I error and relative power of QTDT, QTLD, and MG tests in a large pedigree-based sample, i.e., Genetic Analysis Workshop 16 Problem 3 Framingham Heart Study (FHS) simulated data set. Our investigation was performed with knowledge of the answers.

Methods

Choice of the quantitative traits studied for association analysis

We studied the two simulated quantitative traits, high-density lipoprotein (HDL) and triglyceride (TG), measured at Visit 1 in FHS simulated data set. All our analyses were conducted using the first 100 replicates. Within each replicate, we adjusted trait values for sex and age using a linear regression. We used the residual values of HDL and TG as the phenotypes of interest for association testing. We then assessed the distributions of each trait using the 100 replicates. We found that HDL, but not TG (kurtosis = 16.21, skewness = 2.49), values were normally distributed. The fit to the normal distribution was obtained using a rank-based transformation of TG values (TG_Rob): kurtosis and skewness were equal to -0.02 and 0.003, respectively.

SNP data preprocessing

Genotype data were obtained from the Affymetrix GeneChip Human Mapping 500 k Array. Individual genotype data were filtered based on BRLMM (Bayesian robust linear model with Mahalanobis distance) confidence scores: we used the standard cutoff of 0.5 for call/no-call. Quality control analyses led to 1) exclusion of SNPs with less than 95% call rates, with unknown map position, or with low minor allele frequency (<1%); 2) zeroing out all genotypes at any DNA sample with <95% call rate; 3) exclusion of SNPs not fitting the Hardy-Weinberg equilibrium (p-value $\leq 10^{-6}$) hypothesis; 4) zeroing out genotypes of all individuals in a family at any SNP that showed mendelian errors.

Pedigree sample data

From the total FHS sample of 940 pedigrees, we selected 704 pedigrees having at least two non-founders individuals with available phenotype and genotype data.

Choice of the SNPs tested for association

In brief, the simulation models for HDL and TG included the effects of major genes (five for HDL and three for TG, each explaining 0.1-0.3% of the total variance), and polygenic effects (58% for HDL and 38% for TG). Here, we limited our study to three (*LPL*, *CYP2B7P1*, and *CYP2B6*) of the HDL major genes. TG variability was also explained by two (*LPL* and *CYP2B6*) of these genes. Table 1 lists the main characteristics of all studied SNPs; within each gene the functional SNP is denoted with its symbol name, h^2g is the rate of the trait variance explained by each functional SNP, and D' is the standardized Lewontin's disequilibrium coefficient between the functional variant and each SNP being tightly linked to it. The total number of subjects with available phenotype and genotype data ranged from 5,826 to 5,995. Note that two functional SNPs (α2 and δ1) are tightly (~120 kb) linked but not associated (D' = 0.003). For each gene, we used the functional SNP and two "non-associated" SNPs, selected from the set of SNPs tightly linked (<100 kb) and not associated (D' < 0.10) to the functional SNP. Finally, we also investigated association tests using SNPs not linked to any of the functional variants. The 'false' gene was randomly drawn on chromosome 7 (position: 24,734,008 bp).

Pedigree-based association tests

All association analyses were performed using the qtld command of SOLAR 4.0.7 [7]. The QTDT model decomposes marker effects into two orthogonal components: the between- (bb) and the within- (bw) family components [1]. The restricted model depends on bb only (bw is set to 0). Significance of association is assessed by computing the likelihood ratio of the restricted vs. unrestricted model. The QTLD model [3] includes the founder genotypes in the within-family component rather than in the between-family component (denoted as $b'w$ and $b'b$). Restricted and unrestricted likelihoods of both the QTDT and QTLD models were maximized as a function of the polygenic component (h^2). The MG model is a classical mixed model in which the marker is included as a covariate, and the correlations between relatives are accounted for by h^2 [5]. The regression coefficient of the marker (b) is the association parameter. The restricted MG model depends on h^2 only (b set to 0). The SNP was coded as the number of rare

BMC Proceedings 2009, **3**(Suppl 7):S122 http://www.biomedcentral.com/1753-6561/3/S7/S122

Table 1: Characteristics of the SNPs selected to test association to HDL and TG

Chr	Gene	Position (bp)	SNP	MAF (%)	Symbol	D' (functional variant)[a]	h^2g HDL	h^2g TG
7	None	24,734,008	rs2521760	12.7	-	-	-	-
8		19,794,163	rs17091651	10.0	-	0.04 (α4)	-	-
	LPL							
		19,868,351	rs3200218	21.7	α4		0.3%	0.4%
		19,943,326	rs4244457	32.9	-	0.04 (α4)	-	-
19		46,010,146	rs11083567	18.2	-	0.07 (α2) - 0.03 (δ1)	-	-
	CYP2B7P1	46,089,501	rs8103444	24.4	α2	0.003 (δ1)	0.2%	
	CYP2B6	46,210,613	rs8192719	24.9	δ1	0.003 (α2)	0.3%	0.3%
		46,335,684	rs1631931	13.5	-	0.01 (α2) - 0.03 (δ1)	-	-

[a]Pairwise linkage disequilibrium coefficient (D'/Dmax) between the functional variant (symbol) and the SNPs in its vicinity (<200 kb).

allele copies across all three methods. The effective sample sizes of these three association tests differ. MG model uses all subjects (founders, spouses, and relatives) with known phenotype and genotype status. From this sample, the two family-based association models discard data on the relatives not fulfilling either one of the two conditions: 1) their two parents are genotyped and at least one of them is heterozygote or 2) they have at least one sibling with a different genotype. The effective sample size of QTDT is further reduced because founders and spouses are not use to estimate the within-component effect.

Evidence for population stratification (denoted here as STRAT) is assessed through the likelihood ratio of the restricted (bw and bb are held equal) to the unrestricted (bw and bb are estimated freely) model. All three association tests, as well as the STRAT test, are assumed to follow a chi-square distribution with one degree of freedom.

We performed single-SNP association analyses. For each trait and each SNP, we computed the three association tests (and STRAT test) in each replicate, and derived the mean and standard deviation of each chi-square statistic over 100 replicates. We also derived mean and standard deviations of the association parameters (regression coefficients). Power and error rates were defined as the proportion of replicates with a chi-square value exceeding a given nominal threshold value. MG and QTLD analyses were also performed accounting for population stratification (denoted as MG_S and QTLD_S tests): MG and QTLD chi-square values were both set to zero in the replicates having a STRAT p-value ≤ 5%. The three association tests were evaluated under varying conditions regarding i) inclusion or exclusion of the dietary effect (covariate "diet" affects TG levels and is correlated among family members) and in the association model, ii) trait distribution, i.e., untransformed vs. transformed trait values. Indeed, these association models assume that trait values are normally distributed, and departures from normality can inflate their type I error or reduce their power.

Results and discussion

Table 2 shows empirical estimates of the mean chi-square statistics and type I error rates, at a nominal p-value of 5%, of QTDT, QTLD, and MG tests when the studied SNP is not associated to the trait. The three association tests were roughly similar in empirical estimates, whether or not the studied SNP is linked to a major gene. In general, error rates were lower or close to the nominal values, except for QTDT with two SNPs. As expected, accounting for population stratification decreased the mean chi-square statistics of both QTLD and MG models. Interestingly, in these data, departure from normality did not yield inflated error rates, except with QTDT for TG and rs4244457. Error rates remained unchanged when diet was included into the model (results not shown).

Table 3 shows empirical estimates of the three association models when the studied SNP is the functional polymorphism. For QTLD and MG models, we chose to report the results obtained after accounting for population stratification. Across the three association models, the power was the lowest for the functional SNP with smallest size effects (i.e., α2), and for the less heritable trait (i.e., TG). For TG, mean chi-square estimates were slightly increased when diet was included into the model. For QTDT and MG_S models, estimates were also increased when the trait was normal (i.e., using TG_Rob), relative to when the trait was non-normal. Reverse trends were observed for QTLD_S. The direction of the association parameters was consistent across the three association models (results not shown). Overall, for a given trait and SNP, the mean chi-square statistic was always the highest with MG_S and the lowest with

BMC Proceedings 2009, **3**(Suppl 7):S122 http://www.biomedcentral.com/1753-6561/3/S7/S122

Table 2: Mean χ^2 statistics (μ-χ^2) and type I error rates (at a nominal p) of QTDT, QTLD, and MG association tests

			μ-χ^2(SD)					p = 5%		
Trait	Gene	SNP	QTDT	QTLD	QTLD_S	MG	MG_S	QTDT	QTLD_S	MG_S
A. No association and no linkage										
HDL	none	rs2521760	0.48 (0.62)	0.82 (1.00)	0.82 (1.00)	0.73 (0.85)	0.72 (0.86)	0%	1%	0%
TG			0.86 (1.10)	1.00 (1.15)	0.97 (1.15)	0.64 (0.70)	0.60 (0.69)	3%	4%	1%
TG_Rob			0.99 (1.28)	0.92 (1.24)	0.76 (1.03)	0.48 (0.63)	0.44 (0.62)	4%	4%	0%
B. No association and linkage										
HDL	*LPL*	rs17091651	1.03 (1.12)	1.36 (1.51)	1.36 (1.51)	0.93 (1.01)	0.93 (1.01)	2%	7%	3%
		rs4244457	1.50 (1.79)	1.03 (1.36)	0.60 (0.92)	0.63 (0.79)	0.51 (0.75)	6%	1%	1%
	CYP27B7P1	r11083567	0.85 (0.97)	0.48 (0.63)	0.38 (0.53)	0.40 (0.51)	0.37 (0.52)	2%	0%	0%
	CYP2B6	rs1631931	1.86 (2.02)	0.72 (1.09)	0.61 (0.85)	1.08 (1.24)	0.99 (1.19)	12%	1%	1%
TG	*LPL*	rs17091651	0.70 (0.88)	0.93 (1.31)	0.88 (1.32)	0.89 (1.10)	0.83 (1.12)	1%	4%	5%
		rs4244457	1.76 (1.86)	1.27 (1.46)	0.99 (1.32)	0.65 (1.13)	0.61 (1.12)	14%	5%	3%
	CYP2B6	rs11083567	0.75 (0.95)	1.11 (1.22)	1.03 (1.18)	0.55 (0.77)	0.53 (0.77)	1%	3%	2%
		rs1631931	0.93 (1.35)	1.19 (1.41)	1.10 (1.33)	0.62 (0.79)	0.61 (0.79)	6%	7%	1%
TG_Rob	*LPL*	rs17091651	0.75 (0.99)	0.89 (1.22)	0.77 (1.16)	0.80 (0.89)	0.74 (0.89)	2%	3%	1%
		rs4244457	1.52 (1.74)	0.75 (1.20)	0.66 (1.20)	0.52 (0.98)	0.49 (0.99)	9%	4%	3%
	CYP2B6	rs11083567	0.73 (1.03)	0.66 (1.15)	0.66 (1.16)	0.5 (0.69)	0.5 (0.69)	2%	1%	0%
		rs1631931	0.97 (1.29)	0.85 (1.04)	0.79 (0.98)	0.75 (0.93)	0.74 (0.93)	2%	2%	1%

QTDT. The mean chi-square of QTDT was 1.6 to 6.2 times lower than that of MG_S. For QTLD_S the ratios ranged from 1.0 to 2.4. This is consistent with the amount of marker and trait information used by each association model. For MG, the effective sample sizes (Ne) ranged from 5854 (α4) to 5995 (α2 and δ1) subjects. For QTLD and QTDT, Ne values ranged from 2436 (α4) to 2839 (δ1), and from 1846 (α4) to 2240 (δ1), respectively. It is worth noting that across the three functional variants, the drops in Ne values relative to that of MG varied little: they were the lowest with δ1 (2.11 vs. 2.68 for QTLD vs. QTDT), and the highest with α4 (2.40 vs. 3.17 for QTLD vs. QTDT). In contrast, and for HDL, differences in the performances of the models were more marked with δ1 than with α4. Indeed, mean chi-square statistic of QTLD_S was 1.73 lower than that of MG_S with δ1, whereas both tests showed same performances with α4. Similarly, drops of the mean QTDT statistic, relative to MG_S, were much greater with α2 (6.25) or δ1 (2.37) than with α4 (1.56). It is worth noting that α4 and δ1 explained similar amount of HDL variability. Thus, these results suggest that the relative performance of the association models can not be simply related to differences in the effective sample sizes.

In conclusion, our results showed that MG outperformed the two orthogonal-based association models (QTLD, QTDT), even after accounting for population stratification. QTDT had the lowest power rates. Similar conclusions were reached by two previous simulation studies [3,6]. It is worth noting that our investigation was conducted in a relatively large pedigree-based sample (>5,850 subjects with known phenotype and genotype status; out of these ~10% are founders). Thus, although

Table 3: Mean χ^2 statistics (μ-χ^2) and power (at a nominal p) of QTDT, QTLD, and MG

		μ-χ^2 (SD)			p = 5%			p = 0.1%		
SNP symbol	Trait	QTDT	QTLD_S	MG_S	QTDT	QTLD_S	MG_S	QTDT	QTLD_S	MG_S
α4	HDL	17.88 (6.28)	27.18 (11.51)	27.88 (11.55)	100%	91%	91%	89%	91%	91%
	HDL_Diet	17.87 (6.28)	27.17 (11.51)	27.88 (11.54)	100%	92%	92%	91%	92%	92%
α2	HDL	1.38 (1.35)	3.56 (2.59)	8.62 (5.05)	7%	46%	83%	0%	1%	34%
	HDL_Diet	1.38 (1.35)	3.56 (2.59)	8.62 (5.05)	7%	46%	83%	0%	1%	34%
δ1	HDL	7.13 (3.80)	9.77 (4.48)	16.90 (6.00)	80%	95%	99%	15%	35%	92%
	HDL_Diet	7.13 (3.80)	9.76 (4.48)	16.89 (6.00)	80%	95%	99%	15%	35%	91%
α4	TG	2.21 (2.46)	5.88 (4.67)	9.92 (6.31)	16%	59%	83%	2%	11%	43%
	TG_Diet	2.21 (2.35)	6.01 (4.6)	10.19 (6.07)	22%	62%	86%	2%	14%	46%
	TG_Rob	3.35 (3.16)	4.67 (4.11)	12.67 (6.34)	33%	46%	94%	3%	10%	58%
δ1	TG	3.11 (2.87)	7.97 (5.29)	12.13 (6.16)	28%	78%	92%	2%	26%	52%
	TG_Diet	3.11 (2.87)	8.04 (5.42)	12.28 (6.22)	32%	77%	92%	3%	26%	54%
	TG_Rob	5.15 (3.58)	7.46 (4.64)	17.46 (7.04)	57%	75%	95%	11%	21%	86%

the major gene-specific effects were very modest (<0.4%), the three association models showed good power (>80%, at $p = 5\%$) to detect direct association for HDL and two ($\alpha 4$ and $\delta 1$) of the three functional variants. At a lower tabulated threshold ($p = 0.1\%$), the power remained good using the MG model only. For TG, good power was obtained with the MG model with one functional SNP ($\delta 1$) and using transformed TG values.

List of abbreviations used

FHS: Framingham Heart Study; HDL: High-density lipoprotein; QTLD: Quantitative trait linkage disequilibrium; QTDT: Quantitative transmission-disequilibrium test; MG: Measured genotype; SNP: Single-nucleotide polymorphism; STRAT: Population stratification; TG: Triglyceride.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ASP carried out the statistical genetic analyses and drafted the manuscript. ZV contributed in the statistical analysis and helped to draft the manuscript. MM conceived the study, coordinated it, and contributed to the draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences.

This article has been published as part of *BMC Proceedings* Volume 3 Supplement 7, 2009: Genetic Analysis Workshop 16. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/3?issue=S7.

References

1. Abecasis GR, Cardon LR and Cookson WO: A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 2000, 66:279–292.
2. Fulker DW, Cherny SS, Sham PC and Hewitt JK: Combined linkage and association sib-pair analysis for quantitative traits. *Am J Hum Genet* 1999, 64:259–267.
3. Havill LM, Dyer TD, Richardson DK, Mahaney MC and Blangero J: The quantitative trait linkage disequilibrium test: a more powerful alternative to the quantitative transmission disequilibrium test for use in the absence of population stratification. *BMC Genet* 2005, 6(Suppl 1):S91.
4. Hopper JL and Mathews JD: Extensions to multivariate normal models for pedigree analysis. *Ann Hum Genet* 1982, 46:373–383.
5. Boerwinkle E, Chakraborty R and Sing CF: The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. *Ann Hum Genet* 1986, 50:181–194.
6. Aulchenko YS, de Koning DJ and Haley C: Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* 2007, 177:577–585.
7. Almasy L and Blangero J: Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 1998, 62:1198–1211.

ORIGINAL ARTICLE

Endocrine Care

Genome-Wide Linkage Screen of Bone Mineral Density (BMD) in European Pedigrees Ascertained through a Male Relative with Low BMD Values: Evidence for Quantitative Trait Loci on 17q21–23, 11q12–13, 13q12–14, and 22q11

Jean-Marc Kaufman, Agnès Ostertag, Aude Saint-Pierre, Martine Cohen-Solal, Anne Boland, Inge Van Pottelbergh, Kaatje Toye, Marie-Christine de Vernejoul, and Maria Martinez

Department of Endocrinology and Unit for Osteoporosis and Metabolic Bone Diseases (J.-M.K., I.V.P., K.T.), Ghent University Hospital, 9000 Ghent, Belgium; Institut National de la Santé et de la Recherche Médicale Unit 606 (A.O., M.C.-S., M.-C.d.V.), 75475 Paris, France; Institut National de la Santé et de la Recherche Médicale Unit 563 (A.S.-P., M.M.), 31024 Toulouse, France; Centre National de Génotypage (A.B.), 91057 Evry, France; and Department of Endocrinology (I.V.P.), Onze-Lieve Vrouwziekenhuis Ziekenhuis Aalst, 9300 Aalst, Belgium

Context: Bone mass is under strong genetic control, with heritability estimates greater than 50% and is likely determined by complex interactions between genetic and environmental factors.

Objective: The objective of the study was to localize genes contributing to bone mineral density (BMD) variation.

Design: An autosomal genome-wide scan for BMD at the lumbar spine and femoral neck was conducted with variance components linkage methods.

Participants: A total of 103 pedigrees (Network in Europe on Male Osteoporosis Family Study) ascertained through a male relative with low (Z-score ≤ -2) BMD values at either lumbar spine or femoral neck.

Main Outcome Measures: Nonparametric multipoint logarithm of the odds ratio scores for lumbar spine and femoral neck BMD values adjusted for age, gender, and body mass index.

Results: We identified a total of eight chromosomal regions with logarithm of the odds ratio score of 1.5 or greater ($P \leq 5 \times 10^{-2}$): on 1q42–43, 11q12–13, 12q23–24, 17q21–23, 21q22, and 22q11 for lumbar spine and on 5q31–33 and 13q12–14 for femoral neck BMD.

Conclusions: Four of our detected quantitative trait loci (QTL) reached the genome-wide criteria for significant (17q,21–23, $P \leq 2 \times 10^{-5}$) or suggestive (11q12–13, 22q11, and 13q12–14, $P \leq 7 \times 10^{-4}$) linkage. Apart from 22q11, which is a novel QTL, all other loci provide consistent replication for previously reported QTLs for BMD and other bone-related traits. Finally, several of our specific-linkage areas encompass prominent candidate genes: type 1 collagen (*COL1A1*) and the sclerosteosis/van Buchem disease (*SOST*) genes on 17q21–23; the low-density lipoprotein receptor-related protein 5 (*LRP5*) gene on 11q12–13; and the *rank ligand* gene on 13q12–14. Further analysis of these positive regions by fine linkage disequilibrium mapping is thus warranted. (*J Clin Endocrinol Metab* 93: 3755–3762, 2008)

Osteoporosis is a common multifactorial disease characterized by reduced bone mass and high susceptibility to low-trauma fractures. Low bone mineral density (BMD) is a major risk factor for osteoporotic fracture. Bone mass is under strong genetic control, with heritability estimates greater than 50% and is likely determined by complex interactions between genetic and

0021-972X/08/$15.00/0
Printed in U.S.A.

doi: 10.1210/jc.2008-0678 Received March 27, 2008. Accepted July 23, 2008.
First Published Online July 29, 2008

Abbreviations: BMD, Bone mineral density; BMI, body mass index; LOD, logarithm of the odds ratio; NEMO, Network in Europe on Male Osteoporosis; QTL, quantitative trait locus.

environmental factors, throughout fetal development childhood and adult life. Several genes may be involved in BMD regulation and/or BMD loss. Evidence from studies in animals and humans suggests that the genetic control of BMD may differ by skeletal sites and between genders and/or ethnic populations (1–3). Numerous molecular association or linkage studies aiming to identify genes for BMD determination have been performed, but to date, no clear consensus has been reached. Candidates might be genes involved in cytokine-signaling pathways, the hormonal regulation of calcium homeostasis, or the function of bone cells. Several positive associations for various BMD phenotypes have been reported with different candidate genes and/or polymorphisms, but the role and effect size of the associated polymorphisms/genes remain unclear (1, 2). Similarly, a large number of chromosomal regions have been reported as positively linked to BMD (4–13), but for the majority of these linkage signals, the contributing specific genes have not been identified. Indeed, few of the genomic regions thus far revealed meet the criteria for genome-wide significance, and/or there has been limited replication between studies. A number of factors may have confounded the studies (small sample size, clinical or genetic heterogeneity), making interpretation of the results difficult. Replication and confirmation of the findings are essential to enable conclusions to be drawn.

Here we undertook a full genome-wide screen for BMD variation in a sample of 103 pedigrees recruited within the thematic Network in Europe on Male Osteoporosis (NEMO) and ascertained through a male with low BMD values (Z-score ≤ -2) at either the lumbar spine or femoral neck. This family sample offers the possibility to investigate the genetics of BMD in a rather unique collection of families collected through a male younger than 67 yr with idiopathic osteoporosis.

Subjects and Methods

Family data

The NEMO family study includes 103 Caucasian pedigrees selected through a male relative with idiopathic osteoporosis. Probands were ascertained from 1995 to 2003 in Belgium and France. The sampling scheme and inclusion/exclusion criteria have been elaborated elsewhere (14–16). Briefly, to be eligible as a proband, the subject had to be a male; needed to have a low bone mass, arbitrarily defined by a bone densitometry Z-score of −2.0 or less at the lumbar spine or femoral neck, secondary causes of osteoporosis having been excluded; and aged between 19 and 67 yr. Z-scores are BMD values expressed as units of SDs from the mean for an age- and gender-matched general referent healthy population. Family information was collected on all living first-degree relatives (parents, siblings, children), the proband's spouse, and second or more distant relatives. Relatives, aged between 19 and 85 yr and who agreed to participate, underwent similar clinical investigation as for probands. Gender, age at examination, weight, and height were measured on the visit when the BMD measurements were performed. From all participants a written informed consent was obtained for the study, which was approved by the Ethical Committee of the Ghent University Hospital and the Lariboisière Hospital, respectively.

Measurements of phenotypes

BMDs (grams per square centimeter) of spine and femoral neck were measured using dual-energy x-ray absorptiometry with a QDR 2000 device (software version 7.20; Hologic Inc., Bedford, MA) in Belgium and a DPX-L densitometer (Lunar Corp., Madison, WI) in France. Machines were calibrated daily, and in both participating centers, the coefficient of variation for measurement of a phantom was less than 1%. For the spine, the quantitative phenotype was combined BMD of L2-L4 and L1-L4 for Lunar and Hologic measures, respectively. Members of the same pedigree were measured in the same center and on the same type of machine; data obtained from the two different osteodensitometers were made compatible by linear regression.

Phenotypes analyzed. Before linkage analyses each BMD trait was adjusted for relevant risk factors, including age, gender, and body mass index (BMI), using multiple regression. We used a quadratic function to investigate BMD variations with age. BMI was calculated as weight in kilograms divided by height in meters squared. To remove the effects of these variables, and all possible interactions among them, regression models were built separately in three groups of family members (probands, male and female relatives), as explained in detail elsewhere (17). From the fitted model, a residual value was derived for each subject and each BMD. The distributions of the residuals displayed significant skewness and kurtosis: $P = 7 \times 10^{-5}$ and $P = 4 \times 10^{-8}$ for lumbar spine-BMD and femoral neck-BMD, respectively. To achieve a normal distribution, we removed the effect of outliers and any residual phenotypic data beyond 3 SD and used a natural logarithm to transform the residual levels. The new logarithmically transformed residuals of lumbar spine BMD and femoral neck BMD exhibited nonsignificant kurtosis: $P = 0.07$ and $P = 0.052$, respectively, and for these new traits, the linkage tests can be assumed to follow the standard distribution of logarithm of the odds ratio (LOD) scores (18).

Molecular analyses

Genotyping was carried out at the Centre National de Génotypage (Evry, France). From the whole NEMO sample, 103 families with at least two siblings with DNA available were initially genotyped with a panel of 441 autosomal markers. The Linkage Marker Set MD 10 (Applied Biosystems, Foster City, CA) formed the core marker set for the genome-wide screen. These microsatellite markers, labeled with fluorescent dyes (FAM, HEX, NED), are distributed at an average marker density of 7.9 cM and have an average heterozygosity of 75%. The Centre National de Génotypage has developed a protocol allowing the coamplification of up to six of these markers in a single reaction to be robust using dual 384-well GeneAmp PCR 9700 cyclers (Applied Biosystems) and an automated procedure for PCR and purification setup. Automatic genotyping was performed based on a series of Genetic Profiler software (version 1.1, Amersham Biosciences, Buckinghamshire, UK).

Genotype interpretation and quality control

Before statistical analysis, rigorous genotype quality assurance was performed to ensure accurate binning of alleles. Automatic genotyping was performed based on a series of software processes implemented in Genetic Profiler software (version 1.1) applied to the raw MegaBACE data: trace processing, fragment sizing, allele calling, and assigning genotype quality scores. Consistency of the data with Mendelian inheritance and lack of recombination between loci was evaluated using Pedcheck (19) and other purpose-written software. Allele frequencies for the 441 markers were estimated from our family data by Vitesse 2.0 program (20). Marker order and intermarker distances were obtained from the published Marshfield maps (https/research.marshfieldclinic.org/genetics). We used the sex-average genetic map in all our linkage analyses.

Statistical linkage methods

Multipoint genomic scans for quantitative traits were performed using variance-components linkage method (21), as implemented in SOLAR (22). We estimated the genetic variance attributable to the region around a specific genetic marker ($\sigma^2 m$) by specifying the expected genetic covariances between arbitrary relatives as a function of the identity-by-descent relationships at a given marker locus assumed to be tightly linked to a locus influencing the quantitative trait. Linkage is evaluated by compar-

ing a model incorporating both a genetic additive variance and a polygenic component with a purely polygenic model (no linkage, $\sigma^2 m = 0$). Minus twice the natural logarithm of this likelihood ratio is assumed to follow a one-sided χ^2 distribution. The LOD score is the χ^2 divided by 2 ln10. True multipoint identity-by-descent probabilities were computed using the Markov chain Monte Carlo algorithm implemented in LOKI (23). The ascertainment scheme of pedigrees based on low BMD values of the probands was accounted for in the analyses by computing the likelihood of the pedigrees conditional on the likelihoods of their respective probands (24).

Results

Sample characteristics

The NEMO sample includes 103 extended pedigrees, ascertained in Belgium (n = 72) or France (n = 31). The family size ranged from four to 64 members in a pedigree (up to four generations), with a mean size of eight members (for a full description of the NEMO data, see Ref. 17). Table 1 shows the main characteristics of the NEMO sample. BMD phenotypic and genotypic data were available for a total of 589 and 610 individuals, respectively. Of a total of 3269 relative pairs, 566 (17%) and 540 (16%) are either siblings or cousins. Probands had a mean age of 47.0 yr and a mean BMI of 24.2 kg/m^2 and, as expected, had on average lower BMD values than their relatives. Male and female relatives had similar age and BMI distributions. Initial analyses of the full cohort revealed significant relationships between each BMD trait and the covariates gender, age, and BMI. Together, these variables accounted for 16 and 25% of the total variation of lumbar spine-BMD and femoral neck-BMD, respectively. In the relatives in our sample, the mean (SE) of bone densities adjusted for age, sex, and BMI at lumbar spine and femoral neck are, respectively, 0.89 (0.13) and 0.88 (0.11) g/m^2. In the probands, the mean (SE) of bone densities adjusted for age and BMI at the lumbar spine and femoral neck are, respectively, 0.71 (0.08) and 0.76 (0.09) g/m^2. The adjusted bone densities are significantly ($P < 10^{-15}$) lower at both sites in the probands, reflecting our sampling scheme through low BMD values. Subsequent analyses were conducted using the log-transformed residual values of lumbar spine-BMD and femoral neck-BMD. Both adjusted traits were found to have high heritability estimates: 61 ± 0.07% ($P < 2 \times 10^{-18}$) and 42 ± 0.08% ($P < 6 \times 10^{-10}$), respectively.

Genome-wide linkage analysis

Multipoint linkage analyses were performed across all 22 autosomes. The genome-wide linkage test results for the adjusted lumbar spine and femoral neck BMD phenotypes are shown in Fig. 1. We identified eight chromosomal regions with multipoint LOD score greater than 1.5 (Table 2), on chromosomes 1 (LOD = 1.75, position = 252 cM, close to marker D1S2800), 11 (LOD = 2.64, position = 60 cM at D11S4191), 12 (LOD = 1.65, position = 118 cM close to marker D12S78), 17 (LOD = 3.63, position = 76 cM, at marker D17S787), 21 (LOD = 2.05, position = 44 cM, close to marker D21S266), and 22 (LOD = 2.74, position = 24 cM, close to marker D22S315) for lumbar spine-BMD and on chromosomes 5 (LOD = 1.53, position = 163 cM at D5S422) and 13 (LOD = 2.71, position = 36 cM close to D13S218) for femoral neck-BMD.

Each of the loci detected on these regions appears to affect primarily either spine or hip BMD, not both of these two skeletal sites. Overall, our linkage scan results do not reveal loci with substantial effect on BMD (*i.e.* LOD score above 1) at both skeletal sites (Fig. 1).

We conducted secondary analyses aiming to evaluate, within our eight identified regions, the hypothesis of quantitative trait locus (QTL) with gender-specific effects on BMD variability. We used the strategy, which has been mostly used to identify, in humans, gender-specific QTLs on BMD variation. Gender-specific linkage results were obtained by setting the BMD values of either women or men as missing values, when building the men-strata or women-strata, respectively. Phenotypes of probands were, however, not altered to compute gender-specific LOD scores corrected for ascertainment. The men-strata was slightly smaller than the women-strata: number of informative pedigrees (86 *vs.* 88) and subjects with known phenotypes (301 *vs.* 354), respectively. The results of gender-specific linkage analyses in the eight identified linkage peaks are shown in Fig. 2. Overall, within five regions, similar linkage trends were obtained in each gender strata. The most notable differences in the gender-specific LOD scores were observed on chromosomes 1q42–43 (LOD score = 0.96 *vs.* 0.11), 5q31–33 (LOD score = 0.06 and 1.26), and 22q11 (LOD score = 3.54 *vs.* 0.52) in women and men, respectively. The evidence for linkage was found increased within one region only: on 22q11 (LOD score = 3.54, in women *vs.* LOD score = 2.74 in the full data).

TABLE 1. Characteristics of the NEMO sample

	Individuals			
	Total	Probands	Males	Females
n	821	103	332	386
With DNA	610	103	219	288
Measured covariates and BMD phenotypes	589	103	215	271
Mean age (yr)	43.3 ± 15.8	47.0 ± 11.7	40.2 ± 15.9	43.8 ± 16.7
Range	19–82	19–67	19–80	19–82
Mean BMI (kg/m^2)	24.5 ± 4.1	24.2 ± 3.4	24.6 ± 3.6	24.6 ± 4.7
Range	14.8–45.2	14.8–34.5	16.8–37.0	16.4–45.2
Mean BMD (g/cm^2)				
Lumbar spine	0.89 ± 0.15	0.75 ± 0.09	0.93 ± 0.13	0.92 ± 0.14
Femoral neck	0.75 ± 0.13	0.67 ± 0.11	0.79 ± 0.13	0.74 ± 0.12

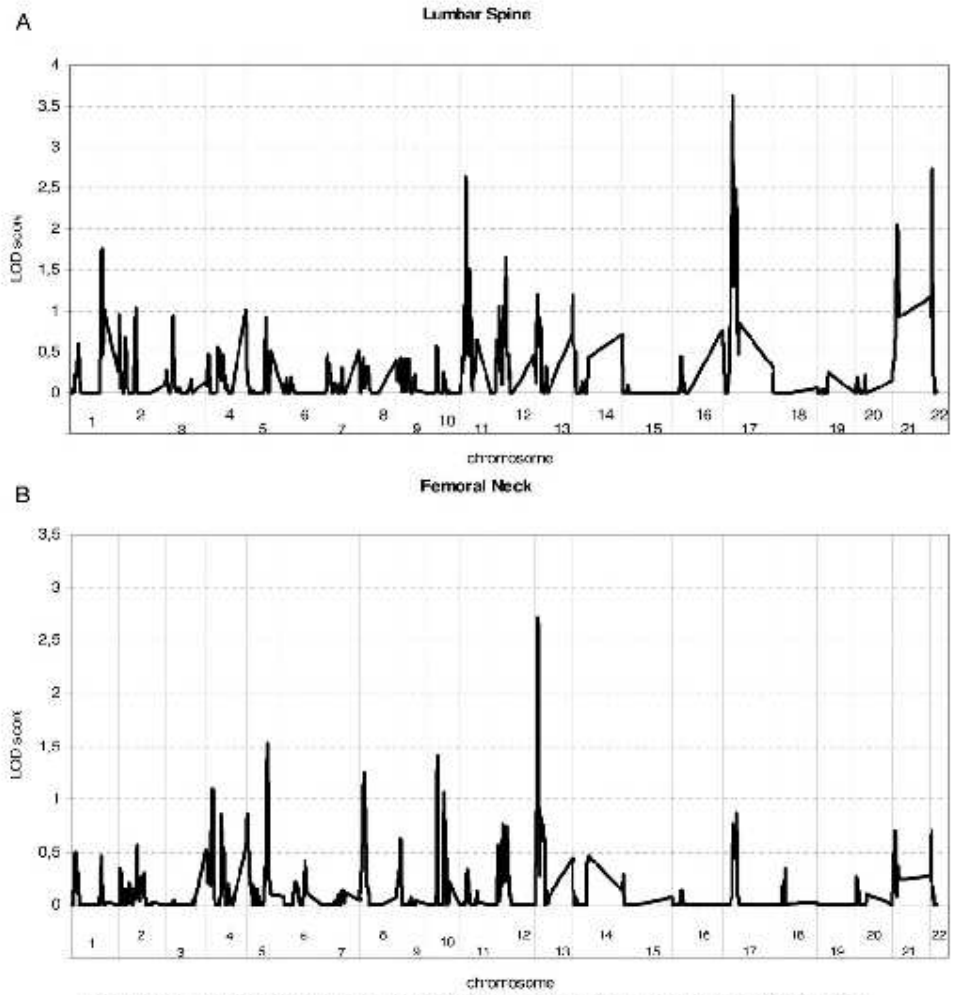


FIG. 1. Multipoint results of the genome-wide linkage scan for adjusted lumbar spine and femoral neck BMD values in 103 NEMO pedigrees.

Discussion

Our multipoint genome scan identified eight chromosomal regions positively linked to lumbar spine-BMD or femoral neck-BMD with LOD score values of 1.5 or greater. Based on the theoretical genome-wide thresholds (25), we obtained one region with significant (*i.e.* point-wise $P \leq 2.2 \times 10^{-5}$) and three regions with suggestive (*i.e.* point-wise $P \leq 7.4 \times 10^{-4}$) evidence for linkage. It is worth noting that several of our eight positive findings overlap with major QTL identified in previous genome-wide scans for BMD and/or encompass prominent candidate genes for BMD variation.

We found little overlap between QTL for lumbar spine and femoral neck. This is consistent with previous whole-genome studies (4, 6, 8, 11, 13) that have reported linkage on different chromosomal regions to BMD at the spine or hip.

Previous genome-wide linkage studies have also suggested that some of the genes regulating BMD may act in a gender-specific manner. Only one of our identified linkage regions supports this hypothesis: evidence for linkage on 22q11 was obtained in women (LOD score = 3.54) but not men (LOD score = 0.52).

TABLE 2. Chromosomal regions with a maximal multipoint LOD score greater than 1.5 for lumbar spine (LS) or femoral neck (FN) BMD in NEMO data

Chromosome	Position (cM)[a]	LODmax	Marker	Pointwise *P* value
LS				
1	252	1.75	D1S2800	2.25×10^{-3}
11	60	2.64	D11S4191	2.46×10^{-4}
12	118	1.65	D12S78	2.93×10^{-3}
17	76	3.63	D17S787	2.19×10^{-5}
21	44	2.05	D21S266	1.05×10^{-3}
22	24	2.74	D22S315	1.90×10^{-4}
FN				
5	163	1.53	D5S422	3.98×10^{-3}
13	36	2.71	D13S218	2.05×10^{-4}

[a] Marker positions using sex-average genetic maps from the Center for Medical Genetics, Marshfield Medical Research Foundation (http://research.marshfieldclinic.org/genetics).

Our highest linkage peak was found on 17q21–23 (LOD = 3.63, at marker D17S787). The same region has been previously identified but for other bone-related phenotypes: wrist bone size (26) and femur head width (27), with a multipoint LOD score of 3.01 and 3.6, respectively. The one-unit support interval surrounding our peak has a chromosomal location in the range of 67–80 cM. It encompasses two prominent candidate genes for BMD: type 1 collagen (*COL1A1*) and the sclerosteosis/van Buchem disease (*SOST*) gene. *COL1A1* is one of the most widely studied candidate genes. It has been significantly associated with osteoporotic fracture risk, but its role on BMD variation remains unclear (2, 28–30). So far, and to our knowledge, two association studies investigated the impact of polymorphisms in the *SOST* gene on BMD and came to contradicting results. In a sample of 619 women, lumbar spine-BMD was not found to be associated with *SOST* sequence variants (31). The second study used a larger cohort (1939 men and women) and showed that the polymorphisms associated with BMD differed in women and men and that the association was mainly observed in the older subjects (32). It also reported significant interaction effects between polymorphisms at the SOST and *COL1A1* genes. Altogether it is possible that either one or both of these two candidate genes explain our linkage signal on 17q21. However, the associated polymorphisms, so far identified by association studies, seem to have very modest effects on BMD variation. Under such conditions, it is striking to observe such a relatively high linkage peak as we obtained on 17q21. We plan to further explore the contribution of these two candidate genes on BMD variation through linkage disequilibrium mapping and also to estimate the amount of the linkage signal that can be explained by these candidate genes' polymorphisms.

We obtained suggestive evidence for a QTL affecting lumbar spine-BMD variation on chromosomes 22q11–12 (LOD score = 2.74, close to D22S315) and 11q12–13 (LOD score = 2.64, close to D11S4191). The QTL on 22q11 is novel and does not overlap with major QTLs reported by other studies. Linkage for lumbar spine-BMD to the 11q12–13 genomic region is supported by two previous studies (6, 33). The first study (33) studied a single large kindred with autosomal dominant high lumbar spine-BMD, and found a linkage peak (maximum LOD score = 5.74) at D11S987, which is about 7 cM telomeric to our peak. The second study reported a multipoint LOD score of 1.97 close to D11S1313 (~2 cM centromeric to our peak) for lumbar spine-BMD in a sample of 835 normal premenopausal Caucasian and Afro-American sisters (6). Other studies identified different regions on chromosome 11: at 47 cM, close to D11S4148, in a sample of Irish families selected through a proband with low BMD (34); and at 109 cM, close to D11S908 (12). The 11q12-q13 contains several putative candidate genes, as the T cell immune regulator 1 (*TCIRG1*) and the low-density lipoprotein receptor-related protein 5 (*LRP5*) gene. Mutations in *LRP5* have been shown to lead to severe Mendelian bone phenotypes and can cause either markedly high or low bone mass traits (35, 36). Thereafter a number of studies have demonstrated that common polymorphic variants in *LRP5* are associated with normal bone mineral density (37–41). Some studies suggested, however, that the contribution of *LRP5* might depend on gender or be limited to women (42, 43). Our secondary linkage analyses within the 11q12–13 region did not favor the hypothesis of a QTL acting in a gender-specific manner on lumbar spine-BMD variation (Fig. 2).

Our third suggestive QTL was obtained on chromosome 13q12–14 (LOD score = 2.71, near D13S218) and for femoral neck-BMD. Previous genome-scan studies have identified linkage peaks in the 13q14 region for different BMD phenotypes. A linkage peak (multipoint LOD score = 1.67, position = 46–55 cM) for distal forearm BMD has been obtained in a Chinese sample of 96 nuclear families (5). The same region has also been identified by gender-specific linkage analyses only but in the opposite gender-strata (9, 44). The study of 29 Mexican-American families (9) revealed, in men only, 13q linkage peaks for neck (LOD = 2.51, position 50 cM) and trochanter (LOD = 3.46, position 45 cM) BMD. Conversely, the study of 941 Asian nuclear families (44) found, in women only, significant evidence for a QTL on 13q for lumbar spine-BMD (LOD score = 3.62, position 40 cM). Potential candidate genes include TNF ligand superfamily, member 11 (*TNFSF11*) gene, which encodes receptor activator of nuclear factor-κB ligand (*RANKL*), the receptor of TRANCE (TNF-related activation-induced cytokine), a TNF family member. Both are critical regulators of dendritic cell and osteoclast function, bone resorption, and calcium homeostasis. A few studies have investigated the association of *RANKL* polymorphisms with BMD and come to conflicting results. A genetic

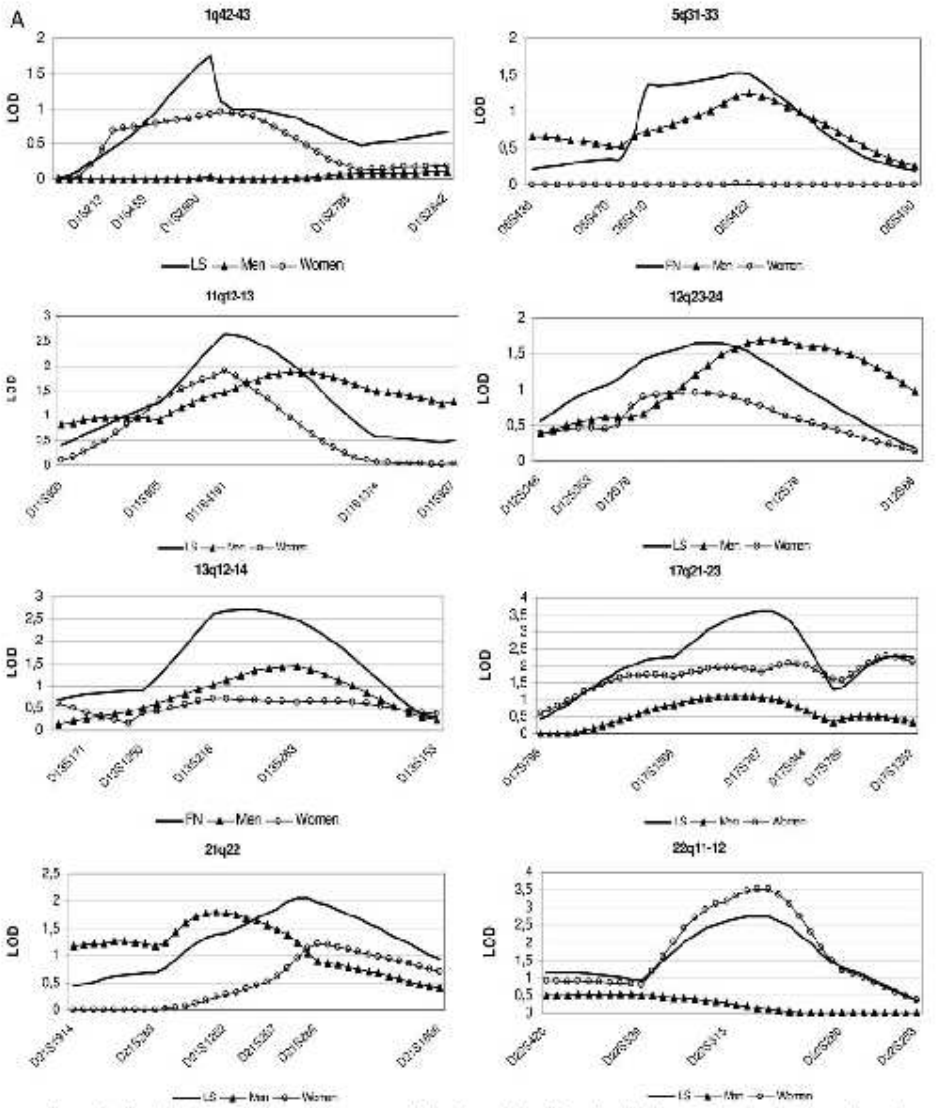


FIG. 2. Linkage peaks identified in the NEMO data: multipoint LOD scores for lumbar spine (LS) and femoral neck (FN) BMD in the combined and by gender-specific strata.

screen of the promoter region identified several polymorphisms, and some were found positively associated with femoral neck-BMD in a sample of postmenopausal women (45). On the contrary, positive association with hip BMD has been found in men but not women (46). These association studies are, however, based on a limited number of polymorphisms in *RANKL*, which could explain their conflicting results. A recent family-based association study used a more systematic approach to screen the underlying genetic variation and showed strong positive association with hip-BMD (47). Additional sex-stratified analyses sug-

gested, however, that the observed association was mainly driven by the male subjects. In our NEMO data, gender-specific linkage analyses showed a slightly higher LOD score value in the men (LOD score = 1.42) than the women (LOD score = 0.75) strata (Fig. 2), but evidence for linkage is weaker in the men strata than the combined data, suggesting that both strata contribute to the linkage peak on 13q12–14.

All our next linkage peaks are consistent with previous studies. Our finding for lumbar spine on 12q23–24 (LOD score = 1.65, close to D12S78) overlaps with the linkage peak (LOD = 1.63, close to D12S79, position = 125 cM) reported for femoral neck in a sample of 40 multiplex Caucasian pedigrees ascertained through a proband with osteoporosis (48). Our finding on 1q42–43 coincides with the linkage peak (LOD score = 2, position = 255 cM, close to D1S2800) identified in a sample of 715 European pedigrees (13). Similarly, our linkage peak on 5q31–33 for femoral neck-BMD was previously identified in a sample of 595 sister-pairs for the same skeletal site and close to the same marker D5S422 (LOD score = 2.23) (6). This 5q31–33 peak encompasses the cytokine cluster and the reversion-induced LIM gene (*RIL*), which has been shown associated with radial BMD in adult Japanese women (49). Our 21q22 peak has also been previously identified. The first study obtained suggestive linkage (LOD score = 3.14, close to D21S1446) of trochanteric BMD (8). A more recent work performed gender-specific linkage analyses and obtained, in men only, linkage evidence at 21q22 (LOD score = 3.36) for spine BMD (50). Possible candidates from this region are collagen VI, α-1, and α-2 (*COL6A1*, *COL6A2*) polypeptide genes.

In conclusion, our study is the first genome-wide linkage screen for genes underlying BMD variability performed, to date, in European pedigrees ascertained through a male relative with low BMD levels. Our results show novel linkage regions and also support some of the linkage regions previously reported. The site-specific differences in the heritability of BMD are already well acknowledged. Similar to other studies, we found differential linked regions for BMD at specific skeletal sites, supporting the view that different genes regulate BMD at different skeletal sites and that BMD as measured by dual-energy x-ray absorptiometry is a complex phenotype, a composite reflection of volumetric BMD and bone geometry. These observations highlighted the complexity of the interplay between genetic and environmental factors to determine the final BMD variance at specific sites. We found significant linkage on chromosome 17q21–23 and suggestive linkage on chromosomes 11, 13, and 22 for QTLs contributing to BMD variation at the lumbar spine or femoral neck. Further analysis of these positive regions by fine association mapping is thus warranted. We are planning to develop linkage disequilibrium mapping studies using two complementary strategies. The first approach is a gene-centered association study design aiming to scrutinize the prominent candidate genes located in our best-linked regions. Extending the study to search for interacting effects between candidate genes is also warranted. For instance, our linkage signals on 17q21–23 and 11q12–13 could result from epistatic effects of *LRP5* and *SOST* genes on lumbar spine-BMD variation, as suggested by a recent study (51). Indeed, both proteins interact at the extracellular domain of LRP5 that has a role in the wnt canonical pathway involved in bone formation (51). The second project is to conduct a whole genome association study in the NEMO data. This approach may help to delineate the genetic determinants of BMD variation, by identifying additional QTLs, not revealed by linkage approaches.

Acknowledgments

Address all correspondence and requests for reprints to: Dr. Maria Martinez, Ph.D., Institut National de la Santé et de la Recherche Médicale Unit 563, Hôpital Purpan, BP 31024 Toulouse, France. E-mail: maria.martinez@toulouse.inserm.fr.

This work was supported in part by the Network in Europe on Male Osteoporosis (European Commission Grant QLG-CT-2002-00491), the Flemish Fund for Scientific Research (FWO Vlaanderen Grants G.0331.02 and G.0662.07); the Société Française de Rhumatologie, and the French National Agency of Research.

Disclosure Statement: The authors have nothing to disclose.

References

1. Ralston SH 2005 Genetic determinants of osteoporosis. Curr Opin Rheumatol 17:475–479
2. Liu YJ, Shen H, Xiao P, Xiong DH, Li LH, Recker RR, Deng HW 2006 Molecular genetic studies of gene identification for osteoporosis: a 2004 update. J Bone Miner Res 21:1511–1535
3. Zmuda JM, Sheu YT, Moffett SP 2006 The search for human osteoporosis genes. J Musculoskelet Neuronal Interact 6:3–15
4. Devoto M, Shimoya K, Caminis J, Ott J, Tenenhouse A, Whyte MP, Sereda L, Hall S, Considine E, Williams CJ, Tromp G, Kuivaniemi H, Ala-Kokko L, Prockop DJ, Spotila LD 1998 First-stage autosomal genome screen in extended pedigrees suggests genes predisposing to low bone mineral density on chromosomes 1p, 2p and 4q. Eur J Hum Genet 6:151–157
5. Niu T, Chen C, Cordell H, Yang J, Wang B, Wang Z, Fang Z, Schork NJ, Rosen CJ, Xu X 1999 A genome-wide scan for loci linked to forearm bone mineral density. Hum Genet 104:226–233
6. Koller DL, Econs MJ, Morin PA, Christian JC, Hui SL, Parry P, Curran ME, Rodriguez LA, Conneally PM, Joslyn G, Peacock M, Johnston CC, Foroud T 2000 Genome screen for QTLs contributing to normal variation in bone mineral density and osteoporosis. J Clin Endocrinol Metab 85:3116–3120
7. Deng HW, Xu FH, Huang QY, Shen H, Deng H, Conway T, Liu YJ, Liu YZ, Li JL, Zhang HT, Davies KM, Recker RR 2002 A whole-genome linkage scan suggests several genomic regions potentially containing quantitative trait loci for osteoporosis. J Clin Endocrinol Metab 87:5151–5159
8. Karasik D, Myers RH, Cupples LA, Hannan MT, Gagnon DR, Herbert A, Kiel DP 2002 Genome screen for quantitative trait loci contributing to normal variation in bone mineral density: the Framingham Study. J Bone Miner Res 17:1718–1727
9. Kammerer CM, Schneider JL, Cole SA, Hixson JE, Samollow PB, O'Connell JR, Perez R, Dyer TD, Almasy L, Blangero J, Bauer RL, Mitchell BD 2003 Quantitative trait loci on chromosomes 2p, 4p, and 13q influence bone mineral density of the forearm and hip in Mexican Americans. J Bone Miner Res 18:2245–2252
10. Styrkarsdottir U, Cazier JB, Kong A, Rolfsson O, Larsen H, Bjarnadottir E, Johannsdottir VD, Sigurdardottir MS, Bagger Y, Christiansen C, Reynisdottir I, Grant SF, Jonasson K, Frigge ML, Gulcher JR, Sigurdsson G, Stefansson K 2003 Linkage of osteoporosis to chromosome 20p12 and association to BMP2. PLoS Biol 1:E69
11. Wilson SG, Reed PW, Bansal A, Chiano M, Lindersson M, Langdown M, Prince RL, Thompson D, Thompson E, Bailey M, Kleyn PW, Sambrook P, Shi MM, Spector TD 2003 Comparison of genome screens for two independent cohorts provides replication of suggestive linkage of bone mineral density to 3p21 and 1p36. Am J Hum Genet 72:144–155
12. Shen H, Zhang YY, Long JR, Xu FH, Liu YZ, Xiao P, Zhao LJ, Xiong DH, Liu YJ, Dvornyk V, Rocha-Sanchez S, Liu PY, Li JL, Conway T, Davies KM, Recker RR, Deng HW 2004 A genome-wide linkage scan for bone mineral

density in an extended sample: evidence for linkage on 11q23 and Xq27. J Med Genet 41:743–751
13. Ralston SH, Galwey N, MacKay I, Albagha OM, Cardon L, Compston JE, Cooper C, Duncan E, Keen R, Langdahl B, McLellan A, O'Riordan J, Pols HA, Reid DM, Uitterlinden AG, Wass J, Bennett ST 2005 Loci for regulation of bone mineral density in men and women identified by genome wide linkage scan: the FAMOS study. Hum Mol Genet 14:943–951
14. Cohen-Solal ME, Baudoin C, Omouri M, Kuntz D, De Vernejoul MC 1998 Bone mass in middle-aged osteoporotic men and their relatives: familial effect. J Bone Miner Res 13:1909–1914
15. Baudoin C, Cohen-Solal ME, Beaudreuil J, De Vernejoul MC 2002 Genetic and environmental factors affect bone density variances of families of men and women with osteoporosis. J Clin Endocrinol Metab 87:2053–2059
16. Van Pottelbergh I, Goemaere S, Zmierczak H, De Bacquer D, Kaufman JM 2003 Deficient acquisition of bone during maturation underlies idiopathic osteoporosis in men: evidence from a three-generation family study. J Bone Miner Res 18:303–311
17. Pelat C, Van Pottelbergh I, Cohen-Solal M, Ostertag A, Kaufman JM, Martinez M, de Vernejoul MC 2007 Complex segregation analysis accounting for GxE of bone mineral density in European pedigrees selected through a male proband with low BMD. Ann Hum Genet 71(Pt 1):29–42
18. Blangero J, Williams JT, Almasy L 2000 Robust LOD scores for variance component-based linkage analysis. Genet Epidemiol 19(Suppl 1):S8–S14
19. O'Connell JR, Weeks DE 1998 PedCheck: a program for identification of genotype incompatibilities in linkage analysis. Am J Hum Genet 63:259–266
20. O'Connell JR, Weeks DE 1995 The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nat Genet 11:402–408
21. Amos CI 1994 Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54:535–543
22. Almasy L, Blangero J 1998 Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198–1211
23. Heath SC 1997 Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am J Hum Genet 61:748–760
24. Boehnke M, Lange K 1984 Ascertainment and goodness of fit of variance component models for pedigree data. Prog Clin Biol Res 147:173–192
25. Lander E, Kruglyak L 1995 Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247
26. Deng HW, Shen H, Xu FH, Deng H, Conway T, Liu YJ, Liu YZ, Li JL, Huang QY, Davies KM, Recker RR 2003 Several genomic regions potentially containing QTLs for bone size variation were identified in a whole-genome linkage scan. Am J Med Genet A 119:121–131
27. Koller DL, Liu G, Econs MJ, Hui SL, Morin PA, Joslyn G, Rodriguez LA, Conneally PM, Christian JC, Johnston Jr CC, Foroud T, Peacock M 2001 Genome screen for quantitative trait loci underlying normal variation in femoral structure. J Bone Miner Res 16:985–991
28. Mann V, Ralston SH 2003 Meta-analysis of COL1A1 Sp1 polymorphism in relation to bone mineral density and osteoporotic fracture. Bone 32:711–717
29. Grant SF, Reid DM, Blake G, Herd R, Fogelman I, Ralston SH 1996 Reduced bone density and osteoporosis associated with a polymorphic Sp1 binding site in the collagen type Ia1 gene. Nat Genet 14:203–205
30. Van Pottelbergh I, Goemaere S, Nuytinck L, De Paepe A, Kaufman JM 2001 Association of the type I collagen αSp1 polymorphism, bone density and upper limb muscle strength in community-dwelling elderly men. Osteoporos Int 12:895–901
31. Balemans W, Foernzler D, Parsons C, Ebeling M, Thompson A, Reid DM, Lindpaintner K, Ralston SH, Van Hul W 2002 Lack of association between the SOST gene and bone mineral density in perimenopausal women: analysis of five polymorphisms. Bone 31:515–519
32. Uitterlinden AG, Arp PP, Paeper BW, Charmley P, Proll S, Rivadeneira F, Fang Y, van Meurs JB, Britschgi TB, Latham JA, Schatzman RC, Pols HA, Brunkow ME 2004 Polymorphisms in the sclerosteosis/van Buchem disease gene (SOST) region are associated with bone-mineral density in elderly whites. Am J Hum Genet 75:1032–1045
33. Johnson ML, Gong G, Kimberling W, Recker SM, Kimmel DB, Recker RB 1997 Linkage of a gene causing high bone mass to human chromosome 11 (11q12–13). Am J Hum Genet 60:1326–1332
34. Wynne F, Drummond FJ, Daly M, Brown M, Shanahan F, Molloy MG, Quane KA 2003 Suggestive linkage of 2p22–25 and 11q12–13 with low bone mineral density at the lumbar spine in the Irish population. Calcif Tissue Int 72:651–658
35. Gong Y, Slee RB, Fukai N, Rawadi G, Roman-Roman S, Reginato AM, Wang H, Cundy T, Glorieux FH, Lev D, Zacharin M, Oexle K, Marcelino J, Suwairi W, Heeger S, Sabatakos G, Apte S, Adkins WN, Allgrove J, Arslan-Kirchner M, Batch JA, Beighton P, Black GC, Boles RG, Boon LM, *et al.* 2001 LDL receptor-related protein 5 (LRP5) affects bone accrual and eye development. Cell 107:513–523
36. Little RD, Carulli JP, Del Mastro RG, Dupuis J, Osborne M, Folz C, Manning SP, Swain PM, Zhao SC, Eustace B, Lappe MM, Spitzer L, Zweier S, Braunschweiger K, Benchekroun Y, Hu X, Adair R, Chee L, FitzGerald MG, Tulig C, Caruso A, Tzellas N, Bawa A, Franklin B, McGuire S, Nogues X, Gong G, Allen KM, Anisowicz A, Morales AJ, Lomedico PT, Recker SM, Van Eerdewegh P, Recker RR, Johnson ML 2002 A mutation in the LDL receptor-related protein 5 gene results in the autosomal dominant high-bone-mass trait. Am J Hum Genet 70:11–19
37. Urano T, Shiraki M, Enura Y, Fujita M, Sekine E, Hoshino S, Hosoi T, Orimo H, Emi M, Ouchi Y, Inoue S 2004 Association of a single-nucleotide polymorphism in low-density lipoprotein receptor-related protein 5 gene with bone mineral density. J Bone Miner Metab 22:341–345
38. Ferrari SL, Deutsch S, Choudhury U, Chevalley T, Bonjour JP, Dermitzakis ET, Rizzoli R, Antonarakis SE 2004 Polymorphisms in the low-density lipoprotein receptor-related protein 5 (LRP5) gene are associated with variation in vertebral bone mass, vertebral bone size, and stature in whites. Am J Hum Genet 74:866–875
39. Ferrari SL, Deutsch S, Baudoin C, Cohen-Solal M, Ostertag A, Antonarakis SE, Rizzoli R, de Vernejoul MC 2005 LRP5 gene polymorphisms and idiopathic osteoporosis in men. Bone 37:770–775
40. Zhang ZL, Qin YJ, He JW, Huang QR, Li M, Hu YQ, Liu YJ 2005 Association of polymorphisms in low-density lipoprotein receptor-related protein 5 gene with bone mineral density in postmenopausal Chinese women. Acta Pharmacol Sin 26:1111–1116
41. Xiong DH, Lei SF, Yang F, Wang L, Peng YM, Wang W, Recker RR, Deng HW 2007 Low-density lipoprotein receptor-related protein 5 (LRP5) gene polymorphisms are associated with bone mass in both Chinese and whites. J Bone Miner Res 22:385–393
42. Koller DL, Ichikawa S, Johnson ML, Lai D, Xuei X, Edenberg HJ, Conneally PM, Hui SL, Johnston CC, Peacock M, Foroud T, Econs MJ 2005 Contribution of the LRP5 gene to normal variation in peak BMD in women. J Bone Miner Res 20:75–80
43. Crabbe P, Balemans W, Willaert A, van Pottelbergh I, Cleiren E, Coucke PJ, Ai M, Goemaere S, van Hul W, de Paepe A, Kaufman JM 2005 Missense mutations in LRP5 are not a common cause of idiopathic osteoporosis in adult men. J Bone Miner Res 20:1951–1959
44. Hsu YH, Xu X, Terwedow HA, Niu T, Hong X, Wu D, Wang L, Brain JD, Bouxsein ML, Cummings SR, Rosen CJ 2007 Large-scale genome-wide linkage analysis for loci linked to BMD at different skeletal sites in extreme selected sibships. J Bone Miner Res 22:184–194
45. Mencej S, Prezelj J, Kocijancic A, Ostanek B, Marc J 2006 Association of TNFSF11 gene promoter polymorphisms with bone mineral density in postmenopausal women. Maturitas 55:219–226
46. Hsu YH, Niu T, Terwedow HA, Xu X, Feng Y, Li Z, Brain JD, Rosen CJ, Laird N 2006 Variation in genes involved in the RANKL/RANK/OPG bone remodeling pathway are associated with bone mineral density at different skeletal sites in men. Hum Genet 118:568–577
47. Xiong DH, Shen H, Zhao LJ, Xiao P, Yang TL, Guo Y, Wang W, Guo YF, Liu YJ, Recker RR, Deng HW 2006 Robust and comprehensive analysis of 20 osteoporosis candidate genes by very high-density single-nucleotide polymorphism screen among 405 white nuclear families identified significant association and gene-gene interaction. J Bone Miner Res 21:1678–1695
48. Devoto M, Spotila LD, Stabley DL, Wharton GN, Rydbeck H, Korkko J, Kosich R, Prockop D, Tenenhouse A, Sol-Church K 2005 Univariate and bivariate variance component linkage analysis of a whole-genome scan for loci contributing to bone mineral density. Eur J Hum Genet 13:781–788
49. Omasu F, Ezura Y, Kajita M, Ishida R, Kodaira M, Yoshida H, Suzuki T, Hosoi T, Inoue S, Shiraki M, Orimo H, Emi M 2003 Association of genetic variation of the RIL gene, encoding a PDZ-LIM domain protein and localized in 5q31.1, with low bone mineral density in adult Japanese women. J Hum Genet 48:342–345
50. Streeten EA, McBride DJ, Pollin TI, Ryan K, Shapiro J, Ott S, Mitchell BD, Shuldiner AR, O'Connell JR 2006 Quantitative trait loci for BMD identified by autosome-wide linkage scan to chromosomes 7q and 21q in men from the Amish Family Osteoporosis Study. J Bone Miner Res 21:1433–1442
51. Semenov MV, He X 2006 LRP5 mutations linked to high bone mass diseases cause reduced LRP5 binding and inhibition by SOST. J Biol Chem 281:38276–38284

HMG Advance Access published December 6, 2010

Human Molecular Genetics, 2010 1–13
doi:10.1093/hmg/ddq497

Genome-wide association study confirms BST1 and suggests a locus on 12q24 as the risk loci for Parkinson's disease in the European population

Mohamad Saad[1,2,†], Suzanne Lesage[3,4,5,†], Aude Saint-Pierre[1,2], Jean-Christophe Corvol[3,4,5,6], Diana Zelenika[7], Jean-Charles Lambert[8,9,10], Marie Vidailhet[3,4,5], George D. Mellick[11,12], Ebba Lohmann[3,4,5], Franck Durif[13], Pierre Pollak[14], Philippe Damier[15], François Tison[16], Peter A. Silburn[11,12], Christophe Tzourio[17,18], Sylvie Forlani[3,4,5], Marie-Anne Loriot[19,20,21], Maurice Giroud[22], Catherine Helmer[23], Florence Portet[24], Philippe Amouyel[8,9,10,25], Mark Lathrop[7], Alexis Elbaz[17,18], Alexandra Durr[3,4,5,26], Maria Martinez[1,2,*] and Alexis Brice[3,4,5,26,*] for the French Parkinson's Disease Genetics Study Group[‡]

[1]INSERM U563, CPTP, CHU Purpan, 31024 Toulouse, France, [2]Paul Sabatier University, Toulouse, France, [3]Université Pierre et Marie Curie-Paris6, Centre de Recherche de l'Institut du Cerveau et de la Moelle épinière, UMR-S975, Paris, France, [4]INSERM U975, Paris, France, [5]CNRS, UMR 7225, Paris, France, [6]INSERM CIC-9503, Hôpital Pitié-Salpêtrière, Paris, France, [7]Centre National de Génotypage, Institut Génomique, Commissariat à l'Energie Atomique, Evry, France, [8]INSERM U744, Lille, France, [9]Institut Pasteur de Lille, Lille, France, [10]Université de Lille Nord, Lille, France, [11]National Centre for Adult Stem Cell Research, Eskitis Institute for Cell and Molecular Therapies, Griffith University, Brisbane, Queensland, Australia, [12]Department of Neurology, Princess Alexandra Hospital, Brisbane, Queensland, Australia, [13]Service de Neurologie, Hôpital Gabriel Montpied, Clermont-Ferrand, France, [14]Service de Neurologie, CHU de Grenoble, Grenoble, France, [15]CHU Nantes, CIC0004, Service de Neurologie, Nantes, France, [16]Service de Neurologie, Hôpital Haut-Lévêque, Pessac, France, [17]INSERM U708, Paris, France, [18]Université Pierre et Marie Curie Paris6, Paris, France, [19]UMR-S775, Paris, France, [20]Université Paris Descartes, Paris, France, [21]AP-HP, Hôpital Européen Georges Pompidou, Paris, France, [22]Centre Hospitalier Dijon, Dijon, France, [23]INSERM, CR897, Université Victor Segalen Bordeaux-2, Bordeaux, France, [24]INSERM U888, Montpellier, France, [25]CHRU de Lille, Lille, France, and [26]Department of Genetics and Cytogenetics, AP-HP, Pitié-Salpêtrière Hospital, Paris, France

Received July 16, 2010; Revised and Accepted November 10, 2010

We performed a three-stage genome-wide association study (GWAS) to identify common Parkinson's disease (PD) risk variants in the European population. The initial genome-wide scan was conducted in a French sample of 1039 cases and 1984 controls, using almost 500 000 single nucleotide polymorphisms (SNPs). Two SNPs at SNCA were found to be associated with PD at the genome-wide significance level ($P < 3 \times 10^{-8}$). An additional set of promising and new association signals was identified and submitted for immediate replication in two independent case–control studies of subjects of European descent. We first carried out an *in silico* replication study using GWAS data from the WTCCC2 PD study sample (1705 cases, 5200 WTCCC controls). Nominally replicated SNPs were further genotyped in a third sample of 1527 cases and 1864 controls from France and Australia. We found converging evidence of association with PD on 12q24 (rs4964469,

*To whom correspondence should be addressed. Tel: +33 562744587; Fax: +33 562744558; Email: maria.martinez@inserm.fr (M.M.); Tel: +33 142162189; Fax: +33 144243658; Email: alexis.brice@upmc.fr (A.B.)
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
‡The French Parkinson's Disease Genetics Study Group includes Y. Agid, M. Anheim, A.-M. Bonnet, M. Borg, A. Brice, E. Broussolle, J.-C. Corvol, Ph. Damier, A. Destée, A. Dürr, F. Durif, S. Klebe, E. Lohmann, M. Martinez, C. Penet, P. Pollak, O. Rascol, F. Tison, C. Tranchant, M. Vérin, F. Viallet and M. Vidailhet.

combined $P = 2.4 \times 10^{-7}$) and confirmed the association on 4p15/BST1 (rs4698412, combined $P = 1.8 \times 10^{-6}$), previously reported in Japanese data. The 12q24 locus includes RFX4, an isoform of which, named *RFX4_v3*, encodes brain-specific transcription factors that regulate many genes involved in brain morphogenesis and intracellular calcium homeostasis.

INTRODUCTION

Parkinson's disease (PD) is the second most common degenerative disease, affecting 1–2% of individuals older than 65 years. Clinical features of PD result primarily from the loss of dopaminergic neurons in the substantia nigra. Although the common form of PD is sporadic, six genes have been identified, mainly by linkage analyses of Mendelian forms of the disease. Two genes, SNCA (encoding α-synuclein) and LRRK2, have an autosomal dominant inheritance and four other genes, PARK2 (parkin), PARK6 (PINK1), PARK7 (DJ-I) and PARK13 (ATP13A2), have an autosomal recessive inheritance (1). Frequently, mutations in these genes are found in patients with early-onset PD, particularly those with autosomal recessive inheritance. However, in most populations, Mendelian forms of parkinsonism are rare when compared with the most common form of PD, a frequent and complex disorder probably explained by the interaction between genetic and environmental factors.

The first two genome-wide association studies (GWASs) in PD (2,3) provided evidence of association with several loci but most often not at the genome-wide significant level, and most initial association findings were not confirmed by subsequent replication analyses (4). Two recent GWASs (5,6) reported strong or genome-wide significant associations with one or more of the known PD genes (i.e. SNCA, MAPT and/or LRRK2). So far, only two 'new' loci have been identified, 1q32/PARK16 and 4p15/BST1, in the Japanese data (5). The US/UK/German GWAS (6) replicated positive association with variants at PARK16 but failed to replicate the association at BST1.

To identify additional variants that affect PD risk in the European population, we designed a three-stage GWAS of PD in three case–control samples from France, the UK and Australia (total of >13 300 subjects). A set of 50 top association signals was identified in the scan sample (1039 cases and 1984 controls from France) using the Illumina-610Quad chip. Promising and new signals were followed-up for stepwise replication in two further UK and French/Australian case–control studies (>3200 cases and 7000 controls).

RESULTS

The genome-wide association results from logistic test corrected for genomic inflation (GC) revealed two single nucleotide polymorphisms (SNPs) with $P_{GC} < 10^{-7}$, and a substantial number of SNPs with strong ($P_{GC} < 10^{-4}$) evidence of association (Fig. 1 and Table 1). For practical reasons, we focus our attention on the 50 best-associated SNPs to prioritize for immediate *in silico* replication (Table 1). Secondary logistic analyses, adjusted for the first two principal componenets (PCs) led to similar rank order of SNPs, albeit slightly weaker association signals (Table 1). This suggests that the significant results, revealed by our primary analyses, are not biased by residual population substructure within our French scan sample. The 50 best-associated SNPs spanned 23 distinct genomic loci, and were associated with $P_{GC} < 5.6 \times 10^{-5}$. Sixteen associations were found within two well-known PD genes, i.e. SNCA (4q22, 4 SNPs) and MAPT (17q12–q21, 11 SNPs), or within BST1 (4p15, one SNP), a recently reported PD risk locus established at the genome-wide significance level in a Japanese population (5). The remaining 34 SNPs were located in 20 distinct previously unreported putative PD loci. The two genome-wide significant SNPs were located on 4q22/SNCA [rs356220, $P_{GC} = 2.82 \times 10^{-8}$, OR = 1.37; 95% CI (1.22–1.53) and rs2736990, $P_{GC} = 2.88 \times 10^{-8}$, OR = 1.35 (1.22–1.50)]. The next most significant SNP was on chromosome 12q21/LOC401725 [rs7954761, $P_{GC} = 2.09 \times 10^{-7}$, OR = 1.34 (1.20–1.50)].

The 50 top SNPs were tested for *in silico* replication in the WTCCC2 PD study data (Table 1). For the sake of clarity, OR values are reported as a function of the number of risk alleles as identified in the stage-1 data. Associations for all 15 SNPs in SNCA and MAPT genes were replicated at nominal P-values $< 4 \times 10^{-5}$. Association with the BST1 variant was also replicated but at a weaker significance level (OR = 1.08, $P = 0.025$). For all SNPs at SNCA, MAPT and BST1, the results in the French scan and the UK replication samples were highly congruent in terms of risk alleles and allele frequencies. As expected, ORs estimated in our scan study tend to be higher than those obtained in the replication-stage data, especially for BST1. Of the remaining 20 loci, association signals at three loci (four SNPs) were replicated with nominal $P < 5\%$ and with the same direction of effect. These SNPs were located on chromosomes 2q21.3 (rs621341, OR = 1.08, $P = 0.028$ and rs6723108, OR = 1.11, $P = 0.005$), 12p13.3 (rs11064524, OR = 1.08, $P = 0.045$) and 12q24 (rs4964469, OR = 1.11, $P = 0.0045$). Differences in allele frequencies across the data from France and UK were notable for the 2q21–q22 SNPs. Indeed, the region encompasses the LCT (lactase) gene whose SNPs are known to vary in frequency across Europe, and rs6723108 has been shown to have different allele frequencies in the French and the UK–Irish populations (7).

We further followed-up the five replicated SNPs (from three newly identified loci and from BST1) in the second replication dataset (1527 cases and 1864 controls from France and Australia) (Table 2). In stage 3, evidence of association was assessed with the Mantel–Haenszel test to control for the potential confounding owing to the different geographical origins (France versus Australia). Evidence of association was replicated for two SNPs located on 12q24 (rs4964469, OR = 1.12, $P = 0.0175$) and on 4p15/BST1

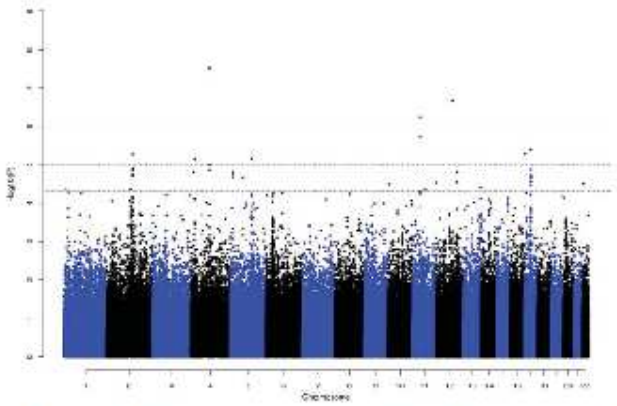


Figure 1. Manhattan plot of the genome-wide association results for 492 929 SNPs. Logistic analysis corrected for genomic inflation (GC results).

(rs4698412, OR = 1.10, $P = 0.029$). Association signals from joint analysis of the two replication datasets were improved for the same two SNPs only: at 4p15/BST1 (stage 2 + stage 3, $P = 0.0033$) and at 12q24 (stage 2 + stage 3, $P = 0.00036$) loci. Notably, joint analysis of the three datasets showed a consistently greater support for association for the newly identified locus on chromosome 12q24 ($P = 2.38 \times 10^{-7}$) than for 4p15/BST1 ($P = 1.79 \times 10^{-6}$). Additional analyses showed that our initial association signals were not confounded by age and they did not appear to be driven either by the subgroup of cases having an early age (<50) of onset of the disease or by those having a positive family history of PD (results not shown). The population-attributable risk (PAR) associated with SNCA, MAPT, BST1 and the 12q24 locus estimated in stage-1 data was 11, 20, 13 and 8%, respectively; in the combined data, PAR was 7% and 4% for BST1 and 12q24, respectively.

Finally, we also examined 18 SNPs from five loci, previously reported to be associated with PD at a genome-wide significance level from two published GWASs of PD (5,6) (Table 3). We added three SNPs (at SNCA and BST1) that were found strongly associated in our stage-1 data. The table also shows the results for a suggestive PD risk locus (GAK) reported by the published GWAS of PD from familial cases (8). As for the previously reported PD loci, two loci only (SNCA and MAPT) have been identified with genome-wide significance at the screen stage: SNCA in both the Japanese and European populations and MAPT in the European population only. The two new PD risk loci (BST1 and PARK16) were identified in the Japanese population: association signals were strong ($P < 10^{-5}$) in the discovery sample, and exceeded $P < 10^{-8}$ in the combined data (5). As already reported here, our GWAS provided genome-wide significance for two SNCA variants and replicated positive associations for variants at MAPT and BST1 loci. It is worth noting that allele frequencies may differ markedly between the Japanese and the European datasets, especially for SNPs at the SNCA and BST1 loci. Saliently, the directions of effects (i.e. risk allele) and effect sizes at SNCA variants are rather congruent across the European and Japanese datasets. For the remaining three PD loci, evidence of association was nominal (PARK16, $P_{GC} = 0.03$; LRKK2, $P_{GC} = 0.04$; GAK $P_{GC} = 0.008$) in the France-GWAS data.

DISCUSSION

Our genome-wide association analyses in the French scan data revealed two SNPs with genome-wide significance ($P_{GC} < 10^{-7}$), and a number of additional SNPs with suggestive evidence ($P_{GC} < 10^{-4}$). Here, we focused on the 50 top associated SNPs for immediate replication in two independent case–control samples. We used a stepwise replication design. To refine the set of most promising results, we first conducted *in silico* replication for the 50 SNPs in the WTCCC2 PD data (1705 cases and 5200 WTCCC controls). Replicated SNPs were genotyped and tested in a third dataset of 1527 cases and 1864 controls from France and Australia. Our scan stage showed genome-wide significance of association with PD for two SNPs at the 4q22/SNCA locus ($P_{GC} < 2.88 \times 10^{-8}$). Indeed, out of the 50 top associated SNPs, 15 are located in genomic regions of two known PD genes (SCNCA, MAPT) and one is located on 4p15/BST1, a risk locus recently reported with genome-wide significance in Japanese samples. SNPs at SNCA and MAPT were all significantly associated with PD in the UK-GWAS data (SNCA, $P < 8 \times 10^{-5}$; MAPT, $P < 2.75 \times 10^{-6}$). Evidence of association with 4p15/BST1 was also replicated in the UK sample but at a lower (rs4698412, $P = 0.025$) significance level. Out of the remaining 34 SNPs, four SNPs (three loci) showed significant ($P < 0.05$) and consistent evidence of association in the UK data. A total of five SNPs (four loci: 2q21.3, 12p13.3, 12q24 and BST1) were followed-up for replication in the third case–control sample. Two of the four tested regions were replicated: 4p15/BST1 ($P = 0.03$) and 12q24

Table 1. Top 50 SNPs in scan stage and *in silico* replication results

Chromosome (gene)	Position (bp)	SNP	Stage-1: scan (France) data					Stage-2: replication (UK) data		
			RA[a]	RAF[b]	OR	P_{GC} (two-tailed)[c]	P_{2PCs} (two-tailed)[d]	RAF	OR	P (one-tailed)[e]
Known PD genes/previously published loci										
4q22 (SNCA)	90858538	rs11931074	T	0.07	1.52	1.35E-05	9.04E-05	0.07	1.33	4.01E-05
	90860363	rs356220	T	0.35	1.37	2.82E-08	6.26E-07	0.36	1.27	2.59E-09
	90894261	rs3857059	G	0.07	1.54	1.00E-05	6.32E-05	0.07	1.33	3.95E-05
	90897564	rs2736990	G	0.44	1.35	2.88E-08	1.32E-07	0.45	1.24	3.98E-08
17q12–21 (MAPT)	41074926	rs393152	A	0.75	1.32	2.68E-05	1.43E-04	0.76	1.31	2.20E-08
	41279463	rs12185268	A	0.76	1.32	3.44E-05	1.62E-04	0.76	1.30	3.59E-08
	41279910	rs12373139	G	0.76	1.33	1.81E-05	7.60E-05	0.76	1.30	2.77E-08
	41281077	rs17690703	C	0.72	1.34	3.94E-06	6.61E-06	0.72	1.24	1.37E-06
	41347100	rs17563986	A	0.75	1.34	1.30E-05	5.65E-05	0.76	1.31	2.58E-08
	41412603	rs1981997	G	0.76	1.33	2.20E-05	8.81E-05	0.76	1.30	4.61E-08
	41436901	rs8070723	A	0.75	1.33	2.19E-05	8.91E-05	0.76	1.30	2.61E-08
	41544850	rs7225002	A	0.59	1.27	2.72E-05	4.08E-05	0.57	1.23	1.14E-07
	41602941	rs2532274	A	0.75	1.33	2.21E-05	1.06E-04	0.75	1.28	2.92E-07
	41605885	rs2532269	T	0.75	1.33	1.90E-05	8.58E-05	0.76	1.29	1.11E-07
	41648797	rs2668692	G	0.76	1.33	1.97E-05	8.20E-05	0.76	1.29	1.22E-07
4p15 (BST1)	15346446	rs4698412	A	0.52	1.28	6.88E-06	1.96E-06	0.55	1.08	0.0247
Newly identified loci										
1p36.22	11880226	rs12724129	T	0.34	1.26	4.35E-05	4.45E-05	0.40	0.99	[f]
1p36.11	26448619	rs10902724	A	0.05	1.55	5.13E-05	5.00E-04	0.07	0.96	[f]
2q14.3	126112282	rs1365783	G	0.42	1.26	4.33E-05	9.10E-05	0.46	0.94	[f]
2q21.3	135011650	rs621341	T	0.28	1.30	5.11E-06	4.37E-04	0.48	1.08	0.0277
	135196450	rs6723108	G	0.31	1.25	6.85E-05	3.20E-03	0.51	1.11	0.0053
	135318626	rs6729702	G	0.46	1.26	1.75E-05	6.02E-04	0.64	1.04	0.15
	135339278	rs6430552	C	0.46	1.26	1.89E-05	6.95E-04	0.64	1.05	0.14
	135367236	rs6714498	T	0.46	1.26	1.84E-05	6.20E-04	0.64	1.05	0.14
2q22	136611978	rs4954564	A	0.52	1.27	1.21E-05	4.07E-04	0.74	1.03	0.27
	136722668	rs6430612	T	0.41	1.27	1.35E-05	1.89E-03	0.65	1.03	0.23
	136730076	rs10221893	T	0.41	1.27	1.31E-05	1.88E-03	0.65	1.03	0.22
2q35	216463846	rs6741233	C	0.87	1.51	9.34E-06	4.46E-04	0.93	1.01	0.46
4p16	11054284	rs368039	A	0.11	1.41	1.52E-05	2.11E-05	0.13	0.86	[f]
5p15.2	10016889	rs1428954	G	0.53	1.27	1.65E-05	3.73E-04	0.57	1.04	0.18
	10026935	rs10072891	G	0.53	1.27	2.02E-05	6.68E-04	0.57	1.04	0.16
	10037418	rs38065	A	0.65	1.29	1.50E-05	2.19E-04	0.69	1.02	0.33
5q12.1	60896208	rs1423326	T	0.60	1.27	2.10E-05	6.54E-04	0.65	1.02	0.36
5q22.2	112814742	rs26990	C	0.13	1.41	6.67E-06	8.67E-06	0.19	0.94	[f]
6q12	70963882	rs9360414	T	0.38	1.25	5.34E-05	4.00E-05	0.38	1.05	0.13
10p14	6933911	rs10905042	C	0.06	1.53	3.08E-05	3.82E-04	0.07	1.02	0.42
11p12	36589978	rs12419750	A	0.89	1.47	4.96E-05	1.65E-05	0.90	0.97	[f]
	36600652	rs1391542	A	0.89	1.47	5.44E-05	1.89E-05	0.90	0.98	[f]
	36613848	rs7128419	A	0.89	1.47	4.91E-05	1.71E-05	0.90	0.98	[f]
	36618299	rs12271660	A	0.90	1.50	5.64E-05	4.56E-05	0.92	0.96	[f]
	36684837	rs12294719	T	0.79	1.44	5.42E-07	6.72E-07	0.81	1.00	0.48
	36687460	rs1533588	A	0.79	1.41	1.79E-06	3.78E-06	0.82	0.97	[f]
11q13.5	75709727	rs12295401	T	0.06	1.53	4.40E-05	5.16E-05	0.06	1.04	0.29
12p13.3	760163	rs11064524	G	0.20	1.32	2.80E-05	1.21E-04	0.24	1.08	0.0447
12q21.31	82691472	rs7954761	A	0.60	1.34	2.09E-07	2.59E-07	0.61	0.99	[f]
12q24	105474117	rs4964469	A	0.33	1.27	2.73E-05	1.30E-04	0.37	1.11	0.0045
	105513235	rs1035767	T	0.11	1.42	1.50E-05	2.15E-05	0.11	0.98	[f]
13q34	113253980	rs2259599	G	0.83	1.39	3.74E-05	5.61E-03	0.88	0.96	[f]
17p13.2	4376339	rs9899558	G	0.73	1.34	5.04E-06	3.82E-04	0.77	0.98	[f]
22q11.23	22917303	rs9608247	A	0.16	1.33	2.99E-05	9.67E-05	0.17	0.95	[f]

[a]Risk allele in stage-1 data.
[b]Risk allele frequency in controls.
[c]Logistic tests corrected for genomic inflation.
[d]Logistic tests including 2PCs as covariates.
[e]P-values shown when the direction of effect in stage-1 and stage-2 data are consistent.
[f]One-tailed $P > 0.5$.

($P = 0.018$). Of the four regions, only one (12p13.3) showed no evidence of association from the combined analysis of the two replication datasets. Overall, evidence of association was consistently stronger with the region of the newly identified PD risk locus than with BST1, in each replication sample as well as in the combined (genome-wide and two replication samples) data (12q24, $P = 2.38 \times 10^{-7}$; BST1, $P = 1.79 \times 10^{-6}$).

Table 2. GWAS and replication: loci considered to follow-up

Locus	Stage e (case/control) Position (bp)	SNP	Stage-1 (France) (1039/1984) RA[a]	RAF[b]	%OR (95% CI)	P (two-tailed)	Stage-2 (UK) (1905/3200) RAF	OR (95% CI)	P (one-tailed)[c]	Stage-3 (France/Australia) (1237/1864) RAF	OR (95% CI)	P (one-tailed)[c,d]	Combined Stage 2 + 3 (3225/5064) OR (95% CI)	P (one-tailed)[c,d]	Stage 1 + 2 + 3 (4271/8048) OR (95% CI)	P (one-tailed)[d]
BST1	15346646	rs4698412	A	0.52	1.28 (1.15–1.42)	6.9E-06	0.55	1.08 (1.00–1.17)	0.0307	0.56	1.10 (1.00–1.21)	0.0290	1.09 (1.02–1.16)	0.0033	1.14 (1.08–1.20)	1.79E-06
3q21.3	135011600	rs823141	T	0.28	1.30 (1.16–1.46)	5.1E-06	0.48	1.08 (1.00–1.17)	0.0277	0.35	1.03 (0.93–1.15)	0.3159	1.06 (1.00–1.13)	0.0300	1.12 (1.04–1.18)	7.90E-05
	135196400	rs6772108	G	0.31	1.25 (1.12–1.40)	6.9E-05	0.51	1.11 (1.02–1.20)	0.0063	0.30	1.03 (0.93–1.14)	0.3047	1.08 (1.01–1.15)	0.0096	1.12 (1.06–1.18)	2.09E-05
11p15.3	748143	rs11028534	G	0.30	1.33 (1.16–1.50)	3.8E-05	0.24	1.08 (0.99–1.18)	0.0467	0.25	0.93 (0.83–1.06)	e	–	–	–	–
12q24	105479117	rs4964469	A	0.33	1.27 (1.13–1.41)	2.7E-05	0.37	1.11 (1.03–1.21)	0.0063	0.34	1.12 (1.01–1.24)	0.0175	1.12 (1.05–1.19)	0.0004	1.16 (1.09–1.22)	2.38E-07

[a]Risk allele in stage-1 data.
[b]Risk allele frequency in controls; %odds ratio computed for the stage-1 risk allele.
[c]P-values shown when the direction of effect in stage-1 and each replication data are consistent.
[d]P-values from stratified association tests.
[e]One-tailed $P > 0.5$.

The evidence of association ($P_{GC} < 1.35 \times 10^{-5}$, Tables 1 and 3) that we detected with several SNPs in the 3′ block of linkage disequilibrium (LD) of the SNCA locus (Fig. 2A), including the two SNPs reaching genome-wide significance in our scan sample, is highly consistent with previous PD GWAS studies (5,6).

MAPT is located in a large block of LD on chromosome 17q12–q22, which contains several additional genes (Fig. 2B). Previous studies (9,10) have identified a large haplotypic block associated with PD, with H1 and H2 being the at-risk and the protective haplotype, respectively. Our two most associated SNPs in the 17q12–q22 region are located within this haplotypic block: rs17690703 ($P_{GC} = 3.9 \times 10^{-6}$) and rs17563986 ($P_{GC} = 1.3 \times 10^{-5}$), the latter being at MAPT. In addition, H2 is tagged by the minor alleles of four of our genotyped SNPs: rs12185268/G, rs12373139/A, rs1981997/A and rs8070723/G. In our scan data, we found the same H1/H2 association signals, with all minor alleles of these four SNPs being significantly associated ($P_{GC} < 3.44 \times 10^{-5}$) with a decreased risk of PD (Table 1).

The BST1 gene has previously been associated with PD in a Japanese GWAS at a genome-wide significance level (5). Strong evidence of association for rs4698412 was found in the Japanese scan ($P = 5.3 \times 10^{-5}$, OR = 1.25) and in the combined (scan + replication) data ($P = 1.8 \times 10^{-8}$, OR = 1.24) (5). A much weaker signal was obtained in the US/UK/German data, in both the scan ($P = 0.09$, OR = 1.07) and the combined ($P = 0.03$, OR = 1.06) data (6). Here, we report strong evidence of association of PD with BST1 (combined $P = 1.79 \times 10^{-6}$, OR = 1.14). The most associated SNP (rs4698412) maps to a 15 kb LD-block (Fig. 2C) and is in high LD ($r^2 = 0.74/0.79$) with the next top two BST1 variants (Table 3). Despite the variation in the allele frequency of the risk allele between the Japanese (RAF = 0.33) and the European (RAF = 0.52–0.56) samples (Tables 2 and 3), we found marked homogeneity in the direction of effects across the groups, but effect sizes seemed to be lower in European than in Japanese samples. BST1 has been proposed to play a role in generating cyclic ADP-ribose that serves as a second messenger for Ca^{2+} mobilization in endoplasmic reticulum and thus Ca homeostasis-related BST1 could be a cause of selective vulnerability of dopaminergic neurons in PD (11).

Our most associated SNP, on 12q24 (combined $P = 2.38 \times 10^{-7}$, OR = 1.16), is 26 kb centromeric of RFX4 (Regulatory factor X4) (Fig. 2D). Two other close genes, POLR3B (Polymerase RNA III polypeptide B) and RIC8B (Resistance to inhibitors of cholinesterase 8 homolog B), are 200 kb centromeric and telomeric of the 12q24 SNP, respectively. The RFX proteins belong to the winged-helix subfamily of helix–turn–helix transcription factors. The *RFX4_v3* transcript variant is the only *RFX4* isoform that is significantly expressed in the fetal and adult brain, and its expression is restricted to the brain. In addition, it has a role in the transcription of many genes involved in brain morphogenesis, such as the signaling components in the wnt, bone morphogenetic protein (BMP) and retinoic acid (RA) pathways. In particular, cx3cl1, a CX3C-type chemokine gene, which is highly expressed in brain in response to injury or infection and regulates intracellular calcium concentration, was downregulated

Table 3. Association results of previously reported PD loci

S1: Table 1 from Satake *et al., Nat. Genet.*, 2009; Japanese data – scan phase (1078 PD/2521 controls); S2: Table 2 from Simon-Sanchez *et al., Nat. Genet.*, 2009; US/GE/UK data – scan phase (1745 PD/4047 controls); S3: Table 2 from Pankratz *et al., Hum. Genet.*, 2009; US (PROGENI + GenePD) data – (857 familial PD cases/867 controls).
[a]Genome-wide significant values ($<10^{-7}$) are italicized.

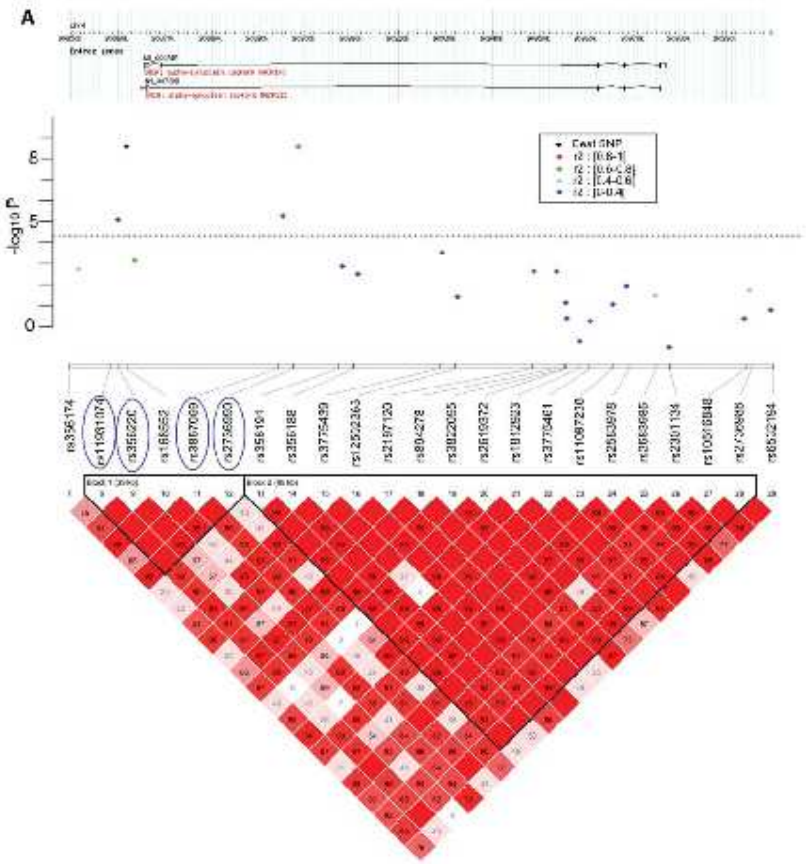


Figure 2. Regional association plots and LD structure for the four PD risk loci (**A**) 4q22/SNCA, (**B**) 17q12–q22/MAPT, (**C**) 4p15/BST1 and (**D**) 12q24/RFX4. The −log10 *P*-values (logistic regression tests corrected for genomic inflation) in the GWAS stage. In each panel, the blue horizontal line indicates a *P*-value of 5×10^{-5}. Pairwise linkage disequilibrium (*D'*) values are displayed and the SNPs with the strongest association signals are circled. SNPs are color-coded for LD relationships (r^2) to the best (colored in black) SNP: red, $0.8 \leq r^2 < 1$; green, $0.6 \leq r^2 < 0.8$; gray, $0.4 \leq r^2 < 0.6$; blue, $0 \leq r^2 < 0.4$. Positions are NCBI build 36 coordinates. Intron and exon structures of genes are taken from the UCSC Genome Browser.

in RFX4_v3-null mice (12). This allows speculation that RFX4 and BST1 are functionally linked and indirectly involved in the regulation of intracellular Ca^{2+} concentrations, which plays an important role in various cellular functions and cell death. Finally, polymorphisms in RFX4 have been shown to be risk factors for the bipolar disorder, manic-depressive illness (13). A recent study showed that a substantial proportion (10–15%) of top GWAS hits, so far identified, are e-quantitative trait loci (eQTLs), i.e. associated with gene expression levels (14). We have initiated eQTL analysis using an existing brain expression database (15), but so far failed to identify any association of the PD-associated rs4964469 SNP with the expression of known genes contained within the 12q24 region.

In conclusion, we have conducted a large GWAS of PD in three case–control samples from France, the UK and Australia. The GWAS stage has 75% and 33% power to detect the loci of the effect sizes observed in stage-1 data for the 12q24 variant (OR = 1.27) at a significance of $P < 5 \times 10^{-5}$ and $P < 10^{-7}$, respectively. In the scan-step, we detected genome-wide significance of association with PD for two SNPs on 4q22, and strong evidence of association with 17q12–q22 SNPs. The two regions encompass previously reported loci: SNCA and MAPT, respectively.

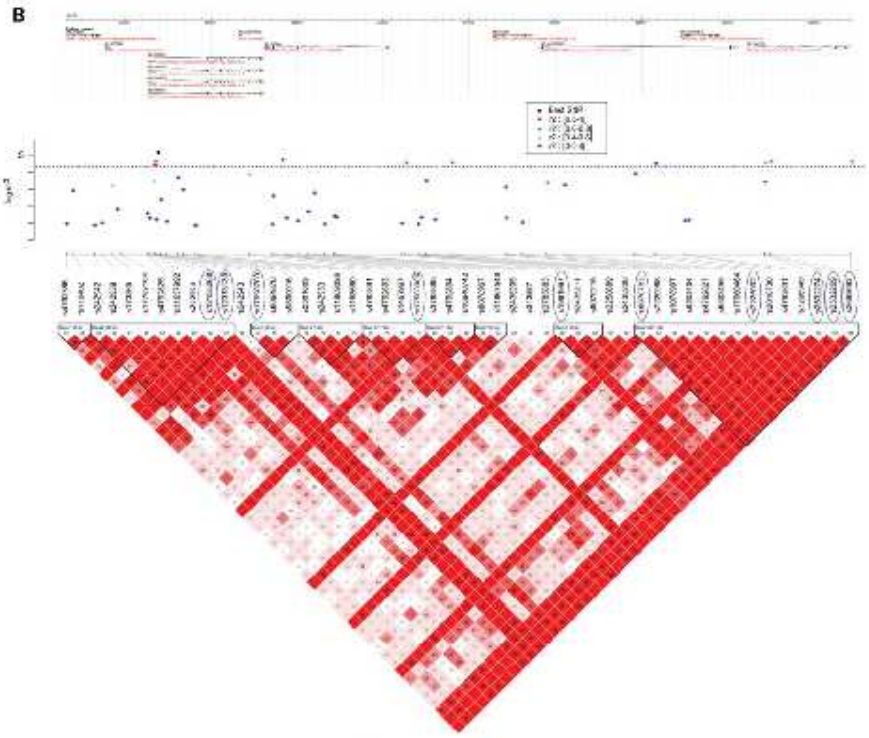


Figure 2. Continued

In addition, we confirmed, for the first time in subjects of European ancestry, the association of PD with 4p15/BST1, recently identified in Japanese samples. Finally, we identified a new locus on 12q24, potentially associated with PD. Further replication studies conducted in large case–control samples are warranted to evaluate the contribution of this locus to PD risk.

MATERIALS AND METHODS

Sample ascertainment and diagnostic criteria

The main characteristics of the three case–control samples are shown in Table 4.

Stage-1 subjects. The total number of cases and controls from France included in stage 1 was 1070 and 2023 controls, respectively.

- **PD subjects:** Patients were recruited through the French network for the study of Parkinson's disease Genetics (PDG) that comprises 15 university hospitals across France. Definite and probable PD was defined according to standard criteria. Definite PD required at least two of three cardinal signs (akinesia and/or rigidity and/or tremor) and absence of exclusion criteria (ophthalmoplegia, pyramidal or cerebellar signs, early dementia, urinary incontinence or postural instability and prior exposure to neuroleptic drugs), and a positive and sustained response to levodopa therapy. Probable PD required at least two of the five following criteria: the parkinsonian triad, a good response to levodopa therapy and asymmetrical onset. Most (>80%) of the PD cases fulfilled the criteria for definite PD. Patients were selected in an effort to enrich for individuals who may have greater genetic predisposition to PD, through selection of cases with a positive family history of PD (Table 4). Cases were of European origin, mostly French ($n = 930$). Subjects diagnosed genetically with known *PARK* mutations (*SNCA*, *LRRK2*, *parkin* and *PINK1*) were excluded.

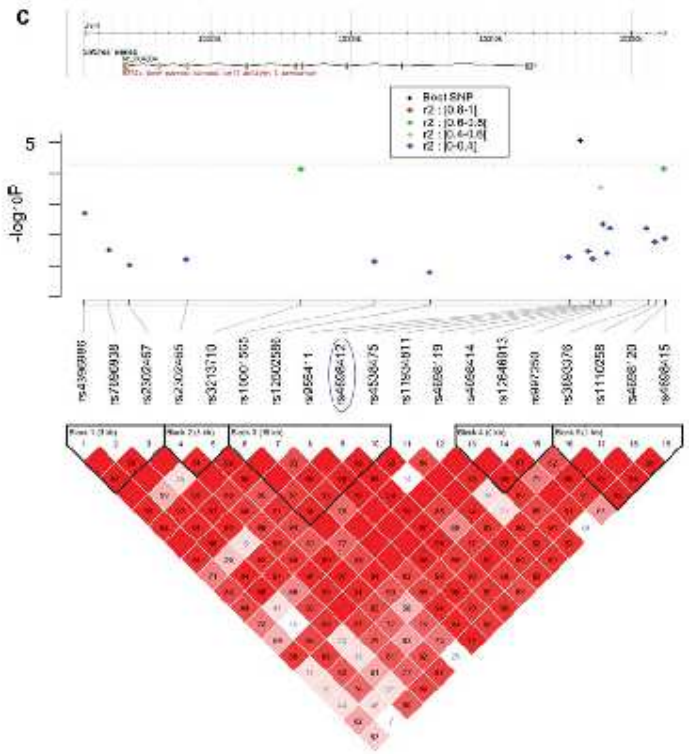


Figure 2. Continued

- 3C neurologically normal controls: The French Three-City (3C) cohort is a population-based, prospective (4-year follow-up) study of the relationship between vascular factors and dementia, carried out in three French cities: Bordeaux (Southwest France), Dijon (central eastern France) and Montpellier (Southeast France) (16). Participants (>9000) are non-institutionalized subjects, over 65 years of age, randomly selected from the electoral rolls of each city. Patients with Alzheimer's disease or other types of dementia, and individuals for whom information on their dementia status during the 4-year follow-up was missing were further excluded. Here, we used a sample of 2023 neurologically normal subjects matched on gender with PD cases, randomly selected from all the participants.

Stage-2 subjects. In silico replication sample: we exchanged genome-wide association data with the WTCCC2 PD study group (Spencer *et al.*, submitted). This case–control study consisted of 1705 PD cases and 5200 controls from the 1958 Birth Cohort and from the OK Blood Services Controls (17).

Stage-3 subjects. De novo genotyping was conducted in two independent case–control datasets from France (872 PD, 1440 controls) and Australia (655 PD, 424 controls). The subjects from France were combined from three French studies: TERRE (207 cases, 468 controls), PARTAGE (313 cases, 593 controls) and an extension of PDG (352 cases, 378 controls). The extension PDG study includes patients who were not available at the time of the stage-1 genotyping execution and neurologically normal spouses of PDG patients. In cases, the mean age at examination and the mean age of onset of PD is 59 (30–86) and 50 (20–84) years, respectively. The mean age of controls is 60 (31–85) years. In PARTAGE, patients and controls were identified among affiliates of the Mutualité Sociale Agricole (MSA) from five French districts. Parkinsonism was defined as the presence of at least two cardinal signs (rest tremor, bradykinesia, rigidity, impaired

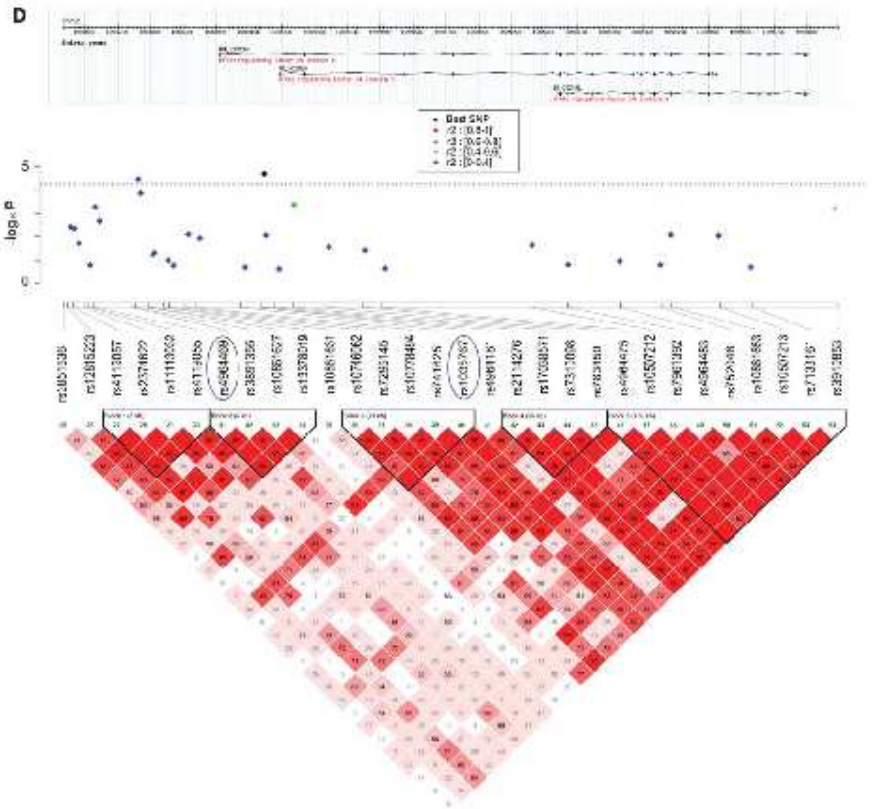


Figure 2. Continued

postural reflexes); PD was defined as the presence of parkinsonism after exclusion of other causes of parkinsonism. Controls were randomly selected from all MSA affiliates in the same districts and matched for sex and age (±2 years). DNA was collected from saliva (Oragene kit). Cases and controls have a mean age of 67 (37–79) years, and the mean age of onset of disease is 63 (35–75) years. TERRE is based on a similar protocol (18), but DNA was collected from blood; the mean age in cases and controls is 73 (46–82) years, and the mean age of onset is 66 (39–80) years in cases.

Australian study: Subjects with PD were recruited from one private and two public movement disorder clinics in Brisbane. Controls were electoral roll volunteers and patient spouses, excluding the subjects demonstrating signs of parkinsonism (19). The mean age is 72 (34–105) and 74 (33–107) years in controls and cases, respectively; the mean age of onset is 59 (23–96) years in cases. Only Caucasian subjects were included in stage 3; in the Australian study, analyses were restricted to participants who reported having four European grandparents (>85% British). There was no overlap between the subjects used in the replication datasets and those included in the stage-1 data. Written informed consent was obtained for all participating subjects and research protocols were approved by local ethics committees.

Genotyping

Stage-1 genotyping. DNA samples of PDG cases and 3C controls were transferred to the French Centre National de Génotypage. First-stage samples that passed DNA quality control (QC) (1064 PD cases and 2023 controls) were genotyped with Illumina Human610-Quad BeadChip and subjected to standard QC procedures.

Table 4. Samples used (post-QC) in this study

Center Genotyping platform	Stage-1 Scan French Illumina 610-Quad	Stage-2 Replication UK Illumina 650Y	Stage-3 Replication French–Australian Illumina GoldenGate	Total
Cases	1039	1705	1527	4271
Sex ratio: M/F	1.42	1.37	1.42	
Age: mean ± SD (n)[a]	57.5 ± 16.6 (1003)	NA	69.0 ± 12.7 (1365)	
AOO: mean ± SD (n)[a]	48.9 ± 12.8 (970)	65.2 ± 11.3 (1109)	61.3 ± 12.4 (1351)	
FH+ (%)	47	0	17	
Controls	1984	5200	1864	9048
Sex ratio (M/F)	1.33	1.02	1.05	
Age: mean ± SD (n)[a]	73.7 ± 5.4 (1984)	51	68.1 ± 10.0	
Total	3023	6905	3391	13 319

[a]Number of subjects for which age/age of onset of disease is known.

Stage-2 genotyping. This WTCCC2 PD study sample was genotyped by the Welcome Trust Case–Control Consortium using the Illumina 650Y genotyping array (Spencer *et al.*, submitted).

Stage-3 genotyping. Genotyping in the extended PDG sample was carried out in the UMR/S 975 laboratory, using predesigned TaqMan probes (C_537709_10/ rs621341; C_29330880_10/ rs6723108; C_12096605_10/ rs11064524; C_2775670_10/ rs4964469; C_1216796_10/ rs4698412) on an ABI 7500 Real-Time PCR system Applied Biosystems, Foster City, CA, USA), according to the manufacturer's instructions. Data were then analyzed using the 7500 software v.2.0.1. The TERRE/PARTAGE and Australian samples were genotyped using the Sequenom MassARRAY platform, with the iPLEX protocol (Genoscreen, France). The basic protocol involves a multiplex primer extension followed by matrix-assisted laser desorption ionization-time of flight mass spectroscopy detection. In order to avoid any genotyping bias, cases and controls were randomly mixed when genotyping and, laboratory personnel were blinded to case–control status.

Quality control of France GWAS scan data

Various stringent QC filters were applied to remove poorly performing SNPs and samples using tools implemented in PLINK version 1.7 (20).

SNP QC: Markers were removed if they had a genotype-missing rate >0.03 or a minor allele frequency (MAF) < 0.05 or a Hardy–Weinberg $P \leq 10^{-5}$. This SNP QC step led to the removal of 74 660 autosomal SNPs. Thus, subsequent analyses were based on 492 929 SNPs.

Individual QC: Samples were removed based on standard exclusion criteria: call rate of <96% (22 subjects), inconsistencies between reported gender and genotype-determined gender (11 subjects) and genetic relatedness (identity-by-descent estimate >0.14; 6 subjects). Applying these QC filters led to the removal of 39 subjects (14 cases, 25 controls).

Population stratification and principal component analysis: To detect individuals of non-European ancestry, we thinned the SNPs to reduce LD to a set of 55 193 SNPs. To this end, we removed SNPs from the extensive regions of LD (CHR2, CHR5, CHR6, CHR8, CHR11) (21), and excluded SNPs if any pair within a 1000-SNP window had $r^2 > 0.2$. Our stage-1 genotype data were then merged with genotypes at the same SNPs from 381 unrelated European (CEU), Yoruban (YRI) and Asian (CHB and JPT) samples from the HapMap project. Principal component analysis was applied using EIGENSTRAT (22). The two PCs clearly separated the HapMap data into three distinct clusters according to ancestry, and most of our stage-1 samples were clustered with the HapMap European samples (Fig. 3). Thirty-two samples appeared to be ethnic outliers (including one subject clearly sharing African ancestry) from the European cluster and were excluded from further analysis. The final post-QC scan sample comprised 1039 PD cases and 1984 controls.

Statistical analysis

Association analysis of the genotype data was conducted with PLINK (20).

Stage-1 association analyses. Logistic regression was used to study the allelic association between each SNP and PD assuming an additive genetic model. Our analysis was based on 492 929 SNPs, and on a conservative genome-wide significance threshold of $0.05/492\,929 = 10^{-7}$. The distribution of the association results was found to be marginally inflated (median $\chi^2 = 0.521$); genomic inflation factor $\lambda = 1.14$ ($\lambda_{1000} = 1.10$). Logistic regression analysis adjusted for the two first PCs of the EIGENSTRAT analysis revealed a genomic inflation of 1.03 (median $\chi^2 = 0.472$). As for our primary analyses, we applied the genomic inflation correction method (23); the median of the GC-corrected χ^2 value was 0.447.

Sensitivity analyses: Two further analyses were conducted to assist in the interpretation of results of the identified GWAS SNPs. We performed age-adjusted regression analysis and conducted subgroup analyses of two subtypes of cases against all controls. Cases with a disease onset before 50 years ($n = 428$) were classified as 'early AOO', and cases

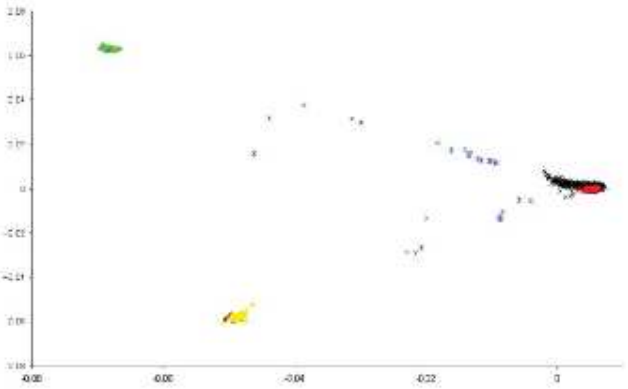


Figure 3. Principal components for our genome-wide stage 1. Plot of the first two principal components from the analysis of our stage-1 (post-QC) data combined with HapMap data. Ethnicity of HapMap samples indicated by color: Africa (YRI) in green, Japan (JPT) in brown, Chinese (6) in yellow and Europe (CEU) in red. Study samples identified to be non-European or not clustering with European samples (outliers) are colored in blue and the remaining study samples assumed to be of European origin are colored in black.

having at least one first-degree relative with PD ($n = 452$) were classified as 'FH+'.

Stage-2 in silico *association analyses*. Statistical data (ORs, effective sample sizes and nominal *P*-values for each of the 50 top SNPs) in the UK sample were obtained from the WTCC2 PD study group that used similar analytical methods (Spencer *et al.*, submitted).

Stage-3 association analyses. For the *de novo* replication stage, we computed association statistics with the Mantel–Haenszel test to control for the potential confounding owing to the geographical center (France versus Australia) for the five SNPs replicated at stage-2. Using raw genotypes from all the study samples, we computed similar stratified (France versus Australia versus UK) association statistics in the combined (stage-2 + stage-3 and stage-1 + stage-2 + stage-3) data.

The PAR associated with the detected variants was estimated with the following formula: $PAR = p\ (OR-1)/[p(OR-1)+1]$, where p is the frequency of the risk allele in controls, and OR is the odds ratio associated with the risk allele.

AUTHOR CONTRIBUTIONS

S.L. supervised DNA sampling; J.C.C., M.V., E.B., F.D., P.P., P.D., F.T., A.D. and A.B. recruited patients; J.C.C., A.D. and AB supervised clinical work; D.Z. and M.L. supervised PD and 3C GWAS genotyping and DNA QC work; S.L. and J.C.L. supervised genotyping of stage-3 samples. A.E., J.C.L., M.A.L., C.T., G.D.M. and P.A.S. contributed to stage-3 replication; M.S., A.S.P. and M.M. executed QC analyses and performed statistical association analyses; A.E., A.B. and M.M. were involved in obtaining funding; M.M. drafted the manuscript and S.L., A.B. and A.E. contributed to the writing of the final version; A.B. and M.M. conceived and oversaw the design and execution of the GWAS.

ACKNOWLEDGEMENTS

The authors are grateful to the patients and their families. They thank the DNA and Cell Bank of UMR_S975 for sample preparation. We thank the members of the 3C consortium: Drs Annick Alpérovitch, Claudine Berr and Jean-Francois Dartigues for giving us the possibility to use part of the 3C cohort. This study makes use of data generated by the Wellcome Trust Case–Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk.

Conflict of Interest statement. The authors declare no competing financial interests.

FUNDING

This work was supported by the French National Agency of Research (ANR-08-MNP-012).

REFERENCES

1. Lesage, S. and Brice, A. (2009) Parkinson's disease: from monogenic forms to genetic susceptibility factors. *Hum. Mol. Genet.*, **18**, R48–R59.
2. Maraganore, D.M., De Andrade, M., Lesnick, T.G., Strain, K.J., Farrer, M.J., Rocca, W.A., Pant, P.V., Frazer, K.A., Cox, D.R. and Ballinger, D.G. (2005) High-resolution whole-genome association study of Parkinson disease. *Am. J. Hum. Genet.*, **77**, 685–693.
3. Pang, H.C., Scholz, S., Matarin, M., Simon-Sanchez, J., Hernandez, D., Britton, A., Gibbs, J.R., Langefeld, C., Stiegert, M.L., Schymick, J. *et al.* (2006) Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol.*, **5**, 911–916.

4. Myers, R.H. (2006) Considerations for genomewide association studies in Parkinson disease. *Am. J. Hum. Genet.*, **78**, 1081–1082.
5. Satake, W., Nakabayashi, Y., Mizuta, I., Hirota, Y., Ito, C., Kubo, M., Kawaguchi, T., Tsunoda, T., Watanabe, M., Takeda, A. *et al.* (2009) Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nat. Genet.*, **41**, 1303–1307.
6. Simon-Sanchez, J., Schulte, C., Bras, J.M., Sharma, M., Gibbs, J.R., Berg, D., Paisan-Ruiz, C., Lichtner, P., Scholz, S.W., Hernandez, D.G. *et al.* (2009) Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat. Genet.*, **41**, 1308–1312.
7. Tian, C., Kosoy, R., Nassir, R., Lee, A., Villoslada, P., Klareskog, L., Hammarstrom, L., Garchon, H.J., Pulver, A.E., Ransom, M. *et al.* (2009) European population genetic substructure: further definition of ancestry informative markers for distinguishing among diverse European ethnic groups. *Mol. Med.*, **15**, 371–383.
8. Pankratz, N., Wilk, J.B., Latourelle, J.C., Destefano, A.L., Halter, C., Pugh, E.W., Doheny, K.F., Gusella, J.F., Nichols, W.C., Foroud, T. *et al.* (2009) Genomewide association study for susceptibility genes contributing to familial Parkinson disease. *Hum. Genet.*, **124**, 593–605.
9. Healy, D.G., Abou-Sleiman, P.M., Lees, A.J., Casas, J.P., Quinn, N., Bhatia, K., Hingorani, A.D. and Wood, N.W. (2004) Tau gene and Parkinson's disease: a case-control study and meta-analysis. *J. Neurol. Neurosurg. Psychiatry*, **75**, 962–965.
10. Zhang, J., Song, Y., Chen, H. and Fan, D. (2005) The tau gene haplotype h1 confers a susceptibility to Parkinson's disease. *Eur. Neurol.*, **53**, 15–21.
11. Zhang, D., Strumpo, D.L., Graves, J.P., Degraff, L.M., Grissom, S.F., Collins, J.B., Li, L., Zeldin, D.C. and Blackshear, P.J. (2006) Identification of potential target genes for RFX4_v3, a transcription factor critical for brain development. *J. Neurochem.*, **98**, 860–875.
12. Chan, C.S., Gertler, T.S. and Surmeier, D.J. (2009) Calcium homeostasis, selective vulnerability and Parkinson's disease. *Trends Neurosci.*, **32**, 249–256.
13. Glaser, B., Kirov, G., Bray, N.J., Green, E., O'donovan, M.C., Craddock, N. and Owen, M.J. (2005) Identification of a potential bipolar risk haplotype in the gene encoding the winged-helix transcription factor RFX4. *Mol. Psychiatry.*, **10**, 920–927.
14. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. and Lathrop, M. (2009) Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.*, **10**, 184–194.
15. Myers, A.J., Gibbs, J.R., Webster, J.A., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., Leung, D., Bryden, L., Nath, P. *et al.* (2007) A survey of genetic human cortical gene expression. *Nat. Genet.*, **39**, 1494–1499.
16. Alperovitch, A. (2003) Vascular factors and risk of dementia: design of the Three-City Study and baseline characteristics of the study population. *Neuroepidemiology*, **22**, 316–325.
17. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature*, **447**, 661–678.
18. Elbaz, A., Clavel, J., Rathouz, P.J., Moisan, F., Galanaud, J.P., Delemotte, B., Alperovitch, A. and Tzourio, C. (2009) Professional exposure to pesticides and Parkinson disease. *Ann. Neurol.*, **66**, 494–504.
19. Sutherland, G.T., Halliday, G.M., Silburn, P.A., Mastaglia, F.L., Rowe, D.B., Boyle, R.S., O'sullivan, J.D., Ly, T., Wilton, S.D. and Mellick, G.D. (2009) Do polymorphisms in the familial Parkinsonism genes contribute to risk for sporadic Parkinson's disease? *Mov. Disord.*, **24**, 833–838.
20. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
21. Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., Ge, D., Rotter, J.I., Torres, E., Taylor, K.D. *et al.* (2008) Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.*, **83**, 132–135. author reply 135–139.
22. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
23. Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Bivariate association analysis in selected samples: Application to a GWAS of two Bone Mineral Density phenotypes in males with high or low BMD

Saint-Pierre Aude[1], Kaufman Jean-Marc[2], Ostertag Agnes[3], Cohen-Solal Martine[3], Boland Anne[4], Toye Kaatje[2], Zelenika Diana[4], Lathrop Mark[4], de Vernejoul Marie-Christine[3], Martinez Maria[1].

[1] INSERM U563, Toulouse, France; [2] Department of Endocrinology and Unit for Osteoporosis and Metabolic Bone Diseases, Ghent University Hospital, Belgium; [3] INSERM U606, Paris, France; [4] Centre National de Génotypage, Evry, France

Running Title: **Bivariate association analysis in selected samples**

Key words: Bivariate association; GWAS; BMD; Osteoporosis

Corresponding author: Aude Saint Pierre, INSERM U563, CHU Purpan, 31024, Toulouse, France; Tel: (33)563 74 45 87; Fax: (33)562 74 45 58; Email: aude.saint-pierre@inserm.fr

Abstract

Our specific aims were to evaluate the power of bivariate analysis and to compare its performance with traditional univariate analysis in samples of unrelated subjects under varying sampling selection designs. Bivariate association analysis was based on the Seemingly Unrelated Regression (SUR) model that allows different genetic models for different traits. We conducted extensive simulations for the case of two correlated quantitative phenotypes, with the Quantitative Trait Locus making equal or unequal contributions to each phenotype. Our simulation results confirmed that the power of bivariate analysis is affected by the size, direction and source of the phenotypic correlations between traits. They also showed that the optimal sampling scheme depends on the size and direction of the induced-genetic correlation. In addition, we demonstrated the efficacy of SUR-based bivariate test by applying it to a real Genome-Wide Association Study of Bone Mineral Density values measured at the Lumbar Spine and at the Femoral Neck in a sample of unrelated males with low Bone Mineral Density (LS Zscores <=-2) and with high Bone Mineral Density (LS and FN Zscores >0.5). A substantial amount of top hits in bivariate analysis did not reach nominal significance in any of the two single-trait analyses. Altogether, our studies suggest that bivariate analysis is of practical significance for GWAS of correlated phenotypes.

INTRODUCTION

With the availability of high-density maps of single nucleotide polymorphisms (SNPs), association studies have become popular tools for identifying genes underlying complex human traits and diseases. For most current population-based genome-wide association studies (GWAS) statistical power is often limited due to the complex interplay among factors that influence the etiology of diseases [1]. Increasing sample size and multilocus or multivariate statistical analyses can improve the power for detecting association. Sample size is often restricted due to genotyping costs and limited sample resources. Several studies have demonstrated that analyzing samples selected with extreme values can be more powerful than analyzing samples randomly selected from the population[2-4]. In addition to using selected samples, another approach to increasing association test power is to perform joint analysis of multiple correlated phenotypes. For many common multifactorial traits, several correlated phenotypes are usually recorded for each individual during sample collection, but most often the phenotypes are analyzed separately in a univariate framework. Joint analysis of correlated phenotypes can theoretically provide greater power than that provided by analysis of individual phenotypes [3,5-7]. Multivariate analysis can also alleviate the multiple testing problem, caused by testing different traits separately, and thereby improve the ability to detect genetic variants whose effects are too small to be detected in univariate analysis[8]. Several multivariate approaches have been applied to linkage studies of correlated complex phenotypes, as osteoporosis and bone-related phenotypes [9-12]. Similarly, various methods, often based on Generalized Estimating Equations (GEE), have been proposed for performing multivariate association tests on population- or family-based data [13-20]. Of the two studies that have investigated the power of bivariate association test in population-based data, one applied the restricted bivariate association test that

assumes same Quantitative Trait Locus effects on each trait [16, 18]. Such constraints in the model may have overestimated or underestimated the relative performance of bivariate over univariate analysis. Finally, GWAS studies using multivariate analysis are rare, especially in samples of subjects selected through their phenotype values, and further investigations using this approach are warranted [4].

To this aim, we evaluated the statistical properties of joint association analysis of two correlated quantitative traits in samples of unrelated subjects through simulation studies, using the Seemingly Unrelated Regression (SUR) bivariate model that allows for different QTL effects on traits. The evaluation was conducted under different situations according to the sample selection design, genetic effects and residual correlation between the traits. We demonstrate the efficacy of SUR-based bivariate test by applying it to simultaneous GWAS analysis of two correlated bone phenotypes, Bone Mineral Density (BMD) at the Lumbar Spine and at the Femoral Neck, which are major risk factors of osteoporosis.

METHODS

SUR-based bivariate model

The Seemingly Unrelated Regression (SUR) model [21] is a generalization of a classical linear regression model that consists of several regression equations with potentially different sets of explanatory variables. It thus allows for a differential effect of explanatory variables on phenotypes as well as the possibility that some variables might be associated with only one trait. Let's N be the total number of unrelated subjects (i=1,..,N), each having observations on two phenotypes y_{ji} (j=1,2). Consider a system of 2 equations, where the jth equation is of the form: $y_j = X_j \times \beta_j + e_j$; y_j is a Nx1 vector of the phenotypic values, X_j is a (K_j+1)xN matrix of explanatory variables with K_j

representing the number of explanatory variables in the model for phenotype j excluding the intercept; $\beta_j = (\beta_0^j, \beta_1^j, ..., \beta_{K_j}^j)'$ is the (K_j+1)x1 vector of coefficients and e_j is a Nx1 vector of the residuals errors. The system of SUR can be written as:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \qquad (1)$$

The SUR model allow for cross-equation correlation of the residual terms. The covariance matrix of all the residuals is assumed to be normally distributed with mean 0 and covariance matrix $E(\varepsilon\varepsilon') = \Omega = \Sigma \otimes I_N$ where I_N is a NxN unit matrix and Σ a 2x2 matrix with the following form:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & rE \times \sigma_1 \sigma_2 \\ rE \times \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix} \qquad (2)$$

σ_1^2 and σ_2^2 are the residual variances of Y_1 and Y_2 respectively and rE is the residual correlation between Y_1 and Y_2.

The SUR model is estimated using the generalized least squares method where the covariance matrix Ω is first estimated using ordinary least squares regression in system (1). Linear restrictions on coefficients can be tested by an F test. The F statistic for systems of equations is: $F = [(C\beta')^t (C \, cov(\beta') C^t)^{-1} (C\beta')] / 2$

where, C is the matrix of restrictions on coefficients. Under the null hypothesis, the F statistic has a central Fisher distribution with 2 and 2xN-K degrees of freedom where K is the total number of estimated coefficients ($K=K_1+K_2+2$). The goodness of fit of the whole system can be measured by the McElroy's r-square (R^2). R^2 is the proportion of co-variance due to X taking into account the residual matrix covariance Ω^{22}.

Here, we applied the SUR model to test association to two continuous phenotypes in unrelated subjects genotyped at one SNP marker, and X_j is the Nx1 vector of genotypes at the SNP. Under an additive model, the genotype for each individual i, noted g_i, is

coded as a function of the number of minor alleles, that is, 0, 1 or 2. We computed the SUR model free of constraints on the regression coefficients, that is, $\beta 1$ and $\beta 2$ were freely estimated. Under the null hypothesis of no association to either one or both phenotypes, the F statistic has a central Fisher distribution with 2 and 2 x (N-2) degrees of freedom. Separate association analyses of Y_1 and Y_2 can be conducted using traditional univariate linear regression model: $y_j = g \times \beta_j + e_j$, where y_j, g, and β_j are as described above but now e_j is assumed to follow a normal distribution N (0, σ_j^2). The null hypothesis of no association ($\beta_j = 0$) can be tested against the alternative ($\beta_j \neq 0$) with a Student statistic (t-test) with N-2 degrees of freedom.

Simulation study:

We considered genetic models of complex traits and specifically tried to generate correlated data mimicking as much as possible our real Bone Mineral Density (BMD) GWAS data (see below). Since a strong (~0.5) and positive phenotypic co-variation exists for BMD values at the Lumbar Spine (LS) and at the Femoral Neck (FN) [23], we generated data for two positively correlated quantitative phenotypes. Further, in real datasets, as causal loci usually contribute a small proportion to the total phenotypic correlation, residual correlation approximates phenotypic correlation between traits. It is also more realistic to assume that the investigator has *a priori* knowledge on the magnitude and sign of the co-variation of the studied phenotypes than on the magnitude and sign of the QTL effect on each phenotype. Therefore, in all our scenarios, the sign of the residual correlation (rE) was positive, but the sign of the induced QTL correlation (rG) was either positive or negative. Also, our BMD GWAS study used a sampling design, with extreme truncate selection of unrelated males, aiming to improve power.

Therefore, we also generated samples of subjects drawn from the extremes of the phenotype(s) population distribution.

The main scenarios and parameter settings are shown in Table 1. The different settings allowed us to generate data for a QTL having same or different effect on the two positively correlated phenotypes, and the two sources of co-variation (QTL and residual) have same or opposite sign. Briefly, we assumed a bi-allelic QTL having additive effects (a_j) on Y_j (j=1,2), with minor and major allele frequency q and p, respectively. The QTL contribution to Y_j is the trait-specific QTL heritability, h^2_j. Here, we focussed our power investigation to QTLs explaining a relatively small part of the trait variance, i.e., from 0.5% to 3% which, for complex traits, seemed to us more realistic. The genotypic means (m_k) of Yj are equal to $2q \times a_j$, $(q-p) \times a_j$ and $-2p \times a_j$ when k, the number of minor alleles, is equal to 0, 1 and 2 respectively and with $a_j = \sqrt{[h^2_j./2pq]}$. We varied the sign of a_j: both were of same or opposite sign and the QTL correlation (rG) was, thus, equal to +1 or -1, respectively. We first generated samples of subjects unselected for their traits values (denoted as Su). Second, we generated subjects selected from the 2.5% (i.e., trait value ≤-2) and 30% (i.e., trait value >0.5) left and right tail of the population distribution of Y_1 (denoted as S1). Third, we included Y_2 in the selection design, that is, we selected subjects from the 2.5% and 30% left and right tail of the population distribution of Y_1 and Y_2 (denoted as S2). These truncate selection criteria (trait value ≤-2 or >0.5) are the values we have used in our real BMD GWAS. Under S1 and S2, we generated samples with equal number of subjects drawn from the left (N/2) and the right (N/2) side of the phenotypes distributions.

Traits values of N (300, 1000) unrelated subjects were generated as follows. For a given combination of parameter values (rE, h^2_1, h^2_2, rG), we first draw QTL alleles from a binomial distribution with parameter q, and built genotypes under Hardy-Weinberg

equilibrium. Then, conditionally on the generated genotype, g_k (k=0,1,2), we jointly drew the values of Y_1 and Y_2 via a bivariate normal distribution with mean $(m_{1k}, m_{2k})^t$ and variance matrix Ω, given in equation (2). Third, under sampling S1 or S2, we applied the corresponding truncate selection, that is individuals not fulfilling the selection criteria were withdrawn from the sample. Steps 1 to 3 were repeated until reaching the required left and right truncated sample sizes of (N/2) subjects.

Each replicate was analyzed with SUR-based bivariate and with two separate univariate analyses using the systemfit package of R software (http://www.r-project.org/) using the genotypes at the QTL, that is, the SNP is the causal variant. The mean and standard deviation of each association statistic (F test and t_1, t_2 tests) were derived from K replicates. Power and type I error rates of each association test were calculated as the proportion of replicates with a test statistic exceeding a given theoretical threshold (Rα) value, at nominal significance levels, α=5%, 1%, 0.1% and 10^{-5}. Type 1 errors were estimated in the settings were $h^2_1=h^2_2=0$ with K=20 000 replicates. Power rates were derived with K=1 000 replicates. To compare the performance of bivariate and that of univariate association analysis, we computed the proportion of replicates where t_1 and t_2 were both lower than Rα. One minus this proportion estimated the probability to detect association to either one of the two phenotypes. To adjust for the two univariate association tests, we applied the Bonferroni correction, that is, we used the theoretical thresholds Rα/2.

RESULTS

SIMULATION STUDY

Tables 2 and 3 present the mean (and sd) association statistic of the SUR-based bivariate (F test) and of the traditional univariate tests (t test), respectively when N=1

000 for 66 scenarios under the alternative hypothesis and when q=0.4. For a given QTL heritability value, the results did not vary, as expected, with q.

Bivariate association statistics: In randomly selected samples, the results in Table 2 show several well-established power figures. First, mean F statistics of bivariate association analysis increase with the size of the trait-specific QTL heritability (h^2_1 and/or h^2_2) irrespective of rG and rE. Second, the power is highest in presence ($rG \neq 0$) than in absence ($rG=0$) of pleitropic effects: the highest power is achieved when $rG=-1$, that is, when the correlation induced by the QTL effect and the residual correlation are opposite in sign. Third, the results also confirm that the power of bivariate association test varies with the size of the residual correlation: when $rG=0$ or $rG=-1$, the power increases with rE; conversely, when $rG=+1$ it decreases with rE. These general trends are observed irrespective of the sampling selection designs. Applying extreme truncate selection increases the power of bivariate association analysis, but the optimal selection design depends on the true genetic model. When $rG=0$ or $rG=-1$, extreme selection on one trait (S1) is more efficient than extreme selection on both traits (S2). Conversely, when $rG=+1$, S2 is more efficient than S1. Overall, under Su or S1, the highest mean F statistics are obtained when $rG=-1$, irrespective of rE. Under S2, the highest power is achieved when $rG=+1$ or when $rG=-1$, depending on the size of rE. Interestingly, when the traits are moderately ($rE=0.20$) correlated, mean F statistics have greater values when $rG=+1$ than when $rG=-1$.

Univariate association statistics Table 3 shows again several well-established power figures. In randomly selected samples, the power of univariate analysis increases with the QTL heritability (h^2_1/h^2_2) and varies little with the size of the residual correlation, rE. For phenotype Y_1, under a given QTL heritability (h^2_1) value, the mean statistic values of all models are similar in the randomly selected samples. Applying extreme

truncate selection increases the power of univariate association analysis of Y_1. Under S1, the power remains similar whichever rG. Under S2, the power is the highest and the lowest for the pleiotropic models rG=+1 and rG=-1, respectively. When rG=-1 or rG=0, the power of univariate association analysis is greater under S1 than under S2. The reverse trend is obtained when rG=+1. For phenotype Y_2, the power of univariate analysis depends on rG and rE. Further, applying extreme selection does not always lead to a gain in power. Indeed, when rG=-1 the power of univariate analysis is the greatest in the unselected samples (Su). When rG=0 the mean t statistic values in the selected samples are biased and inflated. The magnitude of the bias is greater under S2 than under S1. Under S1, the bias increases with rE.

Overall, applying selection criteria on one or both traits is an optimal sampling design when rG=+1: the power of each separate univariate analysis is improved over that in randomly selected samples. When rG=-1, applying extreme truncate selection leads to both a substantial gain and decrease in power for Y_1 and Y_2, respectively. For the situations in which the QTL does not exert pleiotropic effects (rG=0), the highest power of univariate analysis of Y_1 is obtained in the selected samples. However, the mean t statistic values for Y_2, the trait no associated to the QTL, are also increased. Type I error rates of separate univariate analyses may thus be inflated, especially in selected samples and when the residual correlation is high.

Type I error rates: When the QTL/SNP has no effect on Y_1 and Y_2, the values of the mean and standard deviation of both bivariate and univariate association tests are close to the theoretical values, regardless of the residual correlation, minor allele frequency of the studied SNP and of the selection sampling design (Supplementary Table 1.A). Indeed, SUR-based bivariate and each separate univariate association tests have correct type I error rates (Supplementary Table 1.B). However, the false-positive rates of

univariate association analyses for detecting association to either or both the two traits are, as expected, inflated: the estimated rates are roughly two times higher than the theoretical rates. Applying a Bonferroni correction (denoted as U_b) leads to slightly conservative significance levels, especially when the residual correlation between the traits is strong.

Power comparisons : The power to detect association to either or both of the two traits using SUR-based bivariate analysis was compared to the power of separate univariate analysis of Y_1 and Y_2 adjusted for multiple testing by the Bonferroni correction (denoted as U_b). Figure 1.A shows the power curves (at significance of 10^{-5}) against the QTL heritability (h^2_1, h^2_2) when N=1 000, for moderately (rE=0.2) or strongly (rE=0.6) correlated traits. Power curves under S1 and S2 are shown in Figure 1.B, when $h^2_1 = h^2_2 = 0.005$, N=1 000 and rE=0.2 or 0.6.

In randomly selected samples (Figure 1.A), the relative advantage of SUR-based bivariate over univariate association analysis is more obvious when rG=-1 and/or the traits are strongly correlated (rE=0.6) but also when rG=+1 and the traits are moderately correlated (rE=0.2). Under S1 (Figure 1.B), SUR-based bivariate is slightly less powerful than univariate analysis when rG=+1 and rE=0.6 or when rG=0 and rE=0.2. For strongly correlated traits, the power rates are equal to 94.5% (SUR) vs. 29.3% (U_b) when rG=-1; 44.0% (SUR) vs 32.3% (U_b) when rG=0; 36.8% (SUR) vs 39.9% (U_b) when rG=+1. For moderately correlated traits, the power rates are equal to 64.6% (SUR) vs 31.7% (U_b) when rG=-1; 32.9% (SUR) vs 34.9% (U_b) when rG=0; 43.7% (SUR) vs 32.6% (U_b) when rG=+1. Under S2 (Figure 1.B), SUR-based bivariate shows same or slightly lower power than univariate analysis except when rG=-1 or when rG=0 and rE=0.6 where it outperforms univariate test. As already noted above selecting on Y_1 (S1) is the most efficient sampling design when rG=-1 or when rG=0

and the traits are strongly correlated (rE=0.6). Selecting on both traits (S2) is the most efficient design when rG=+1. Overall, when rE=0.6, the power of SUR is the greatest (94.5%) when rG=-1 and under S1, while the power of univariate analysis is the greatest (56.8%) when rG=+1 and under S2. When rE=0.2, the power of SUR and univariate analysis are both the greatest (72.5% and 72.9%) when rG=+1 and under S2. As shown in Supplementary Table 2, all these trends are confirmed under various parameter settings.

ANALYSES OF EMPIRICAL BMD-GENOME-WIDE ASSOCIATION DATA

BMD GWAS data: Subjects were recruited from the Network in Europe on Male Osteoporosis study [24,25]. Subjects selected from this cohort were unrelated males > 18 and < 68 years of age. In addition, the subjects were selected by bone densitometry (measured at the Lumbar Spine and at the Femoral Neck) criteria, having either low BMD (LS-Z-scores ≤-2, n=175) or high BMD (both LS- and FN-Z-scores >0.50, n=155). Further details of the study sample are provided in Supplementary Table 3. Genotyping was carried out at the Centre National de Génotypage (CNG, Evry, France) using the Illumina 370K platform. SNPs and DNA data were subjected to standard quality control analyses with PLINK [26] (details are provided in Supplementary Methods).

Association analysis: Our primary analysis was the joint association analysis of LS-Zscores and FN-Zscores by means of SUR-based bivariate test. For comparison purpose, we also applied separate univariate association analyses of LS and FN Z-scores. We used single marker analysis assuming additive genetic effects. The mean F statistic of our SUR-based genome-wide association analysis was equal to 1.018 (sd=1.022, median= 0.70). The mean t statistic of LS and FN were -0.0167 (sd=1.011,

median=-0.0165) and -0.0129 (sd=1.006, median=0.0104), respectively. These results indicated that there was no meaningful inflation of univariate as well as bivariate association analyses.

Results: SUR-based bivariate analyses identified a substantial number (35) of SNPs with strong evidence of association (P-value<10^{-4}). Interestingly, several of the identified SNPs failed to reach nominal (P-value <5%) significance under separate univariate analyses for either one or the two BMD phenotypes. Genome-wide bivariate and univariate association results were compared in terms of statistical significance and ranks of the SNPs identified in either one of the two approaches. For each SNP, we kept the lowest P-value (denoted as Best_U) of LS or FN univariate association analysis. Univariate P values were not corrected for multiple testing. We ranked the Best_U P-values from the lowest to the highest. We similarly ranked the P-values from SUR-based bivariate analysis of LS and FN. Figure 2 plots the significance levels in each procedure for the top 100 most associated SNPs identified from SUR-based (Figure 2.A) or from univariate (Figure 2.B) analyses. We found that a majority (52) of the top SNPs in SUR-based bivariate analysis also show strong (P<$3x10^{-4}$) association signal in univariate analyses. For a substantial number (16) of the remaining SNPs, univariate analyses fail to reach nominal (P<5%) significance (Figure 2.A) On the other hand, all of the top 100 SNPs in univariate analyses (Figure 2.B) are also highly significant (P<$8x10^{-4}$) in bivariate analysis. Table 4 shows details of the association results for the top 10 SNPs in SUR-based and in each separate univariate analysis. The table also shows P-values and ranks found in each of the two other procedures. The genetic contributions (R^2 values) of the 10 top SNPs are not great, as expected for any relatively common polymorphic locus. Three of the top 10 SNPs from bivariate analysis also rank well (i.e., are in the set of top 300 SNPs) in univariate analyses of LS and/or FN. They

are located on 6q25: rank=2, P=1.3x10^{-5} (LS) and rank=1, P=1.2x10^{-5} (FN); on 15q14-q15: rank= 2 635, P=8.4x10^{-3} (LS) and rank=3, P=1.7x10^{-5} (FN); and on 22q13: rank=1, P=3.5x10^{-6} (LS) and rank=8, P=3x10^{-5} (FN). All the remaining 7 SNPs show a much stronger association signal in bivariate than in univariate analyses, including 2 of the 3 best SUR-based association signals. For the most significant result, on 22q11.2 (P=5.44x10^{-6}), the QTL explains 3.85% of the joint (co)variance of LS and FN. This value likely over-estimates the contribution in unselected populations. Nonetheless, univariate analyses failed to detect association (P>0.07) with this SNP. Conversely, all top 20 SNPs identified from univariate analysis of either LS or FN belong to the set of top 42 SNPs from SUR-based bivariate analysis. Overall, our analyses showed that univariate analysis did not identify new strongly associated SNPs as compared to those detected in bivariate analysis. Conversely, SUR-based analysis identified strongly associated SNPs that were not detected in univariate analysis.

Our study used a design, with extreme truncate selection of unrelated males, aiming to improve power. The approach of studying samples drawn from the extremes of the population distribution of BMD has been used in several linkage studies of BMD variation [25,27], but rarely in association studies[28], and to our knowledge, never in samples drawn from the population of males. Due to our relatively small GWA sample size, no SNP showed evidence of association to either one or both BMD phenotypes at genome-wide significance threshold of 1.7x10^{-7} (0.05/ 298 783 SNPs). However, we used an extreme truncate selection design that, as shown by our simulation studies, has increased power over unselected samples. Our SUR-based bivariate association analyses identified strong association (P<8.4x10^{-6}) with 3 genomic regions (6q22.1, 15q14 and 22q11). These SNPs have not yet been reported to be associated with bone density in previous GWAS [29-31]. Two of them, on 15q14-15 and 22q11, are located in genes that

are known to be expressed in skeletal muscle [32-33]: *GLUT 11* encoded by *SLC2A11* on 22q11 and *RYR3*, on 15q14-15. Because muscle contraction has a major impact on bone density, this might represent an indirect role of these genes on bone density.These genetic variants, whether they are site-specific or possibly shared (pleiotropic), may warrant further follow-up genetic studies on BMD and other bone-related phenotypes.

DISCUSSION

We have evaluated the performance of bivariate association analysis based on the Seemingly Unrelated Regression (SUR) model, which allows different genetic models for different traits. To our knowledge, this is the first study to specifically derive the power and the relative performance of bivariate association analysis in selected samples of unrelated subjects. Our main results coincide with well-known power figures [6-8] and confirmed that bivariate association analysis outperforms univariate analysis when the QTL exerts pleiotropic effects and the relative increase in power is greatest when correlation of the QTL is opposite in sign to the residual correlation. The most powerful sampling selection design varied with the genetic model, specifically with the size and the direction of the induced-QTL correlation. Applying truncate selection on one trait was found the most efficient sampling design when the genetic and the residual correlations are opposite in signs. The same most efficient design was found when the QTL does not exert pleiotropic effects: the power of the SUR-based bivariate association test was found as good as or better than that of univariate association test, depending on the size of the residual correlation. Finally, when the QTL exerts pleiotropic effects and both sources (QTL and residual) of co-variation are of same sign, applying selection criteria on both traits was found the optimal sampling selection

design. Under this sampling design, the performance of SUR-based bivariate test relatively to univariate analysis decreases with the size of the residual correlation.

So far, two studies have investigated the power of bivariate association in unselected population-based data, and they both applied bivariate association test based on Generalized Estimating Equations [16,18]. The former applied a general GEE-based model that allows, as the SUR model, for different QTL effects on the two traits. The second study used a GEE-based bivariate model that assumed same QTL effects on the phenotypes. Our results are congruent with those reported by the first study. The restricted bivariate test estimates, as the univariate test, a single parameter (i.e., the SNP regression coefficients on each trait are all set equal). Under the restricted bivariate model, the gain in power of bivariate analysis is enhanced and reduced when the QTL has similar effect and when it affects one trait only, respectively. Clearly, rarely, knowledge of this magnitude about a complex trait is known *a priori*. Thus, we do not recommend using restricted bivariate models even in unselected data.

Our bivariate genome-wide association analysis of Lumbar Spine and Femoral Neck BMD values, conducted in a sample of unrelated males with low BMD (LS Zscores $\leq$-2) and high BMD (LS and FN Zscores >0.5), consistently demonstrated the advantage of the SUR-based bivariate test over separate univariate analysis. All of the top hits in univariate analysis also showed strong evidence of association in bivariate analysis. Conversely, additional SNP associations were detected with the bivariate method that did not reach nominal significance in single-trait analyses: this was achieved without adjusting significance of univariate analyses for multiple testing.

In conclusion, our results showed that SUR-based models are useful to detect association for correlated phenotypes. However, our results also showed that similar power levels can be achieved whether the QTL exerts or not pleiotropic effects. Thus,

disentangling pure pleiotropic from residual covariation remains a challenge even in bivariate association analysis.

ACKNOWLEDGMENTS

Part of this work was supported by the Network in Europe on Male Osteoporosis (European Commission grant QL6-CT-2002-00491), by the Flemish Fund for Scientific Research (FWO Vlaanderen grants G.0331.02 and G.0662.07); by the Société Française de Rhumatologie (SFR), and by the French National Agency of Research (ANR).

Conflict of interest: none

REFERENCES

1 Hirschhorn JN, Daly MJ: Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 2005; **6**: 95-108.

2 Allison DB: Transmission-disequilibrium tests for quantitative traits. *Am J Hum Genet* 1997; **60**: 676-690.

3 Allison DB, Thiel B, St Jean P, Elston RC, Infante MC, Schork NJ: Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *Am J Hum Genet* 1998; **63**: 1190-1201.

4 Abecasis GR, Cookson WO, Cardon LR: The power to detect linkage disequilibrium with quantitative traits in selected samples. *Am J Hum Genet* 2001; **68**: 1463-1474.

5 Amos C, de Andrade M, Zhu D: Comparison of multivariate tests for genetic linkage. *Hum Hered* 2001; **51**: 133-144.

6 Almasy L, Dyer TD, Blangero J: Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkages. *Genet Epidemiol* 1997; **14**: 953-958.

7 Amos CI, Laing AE: A comparison of univariate and multivariate tests for genetic linkage. *Genet Epidemiol* 1993; **10**: 671-676.

8 Jiang C, Zeng ZB: Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 1995; **140**: 1111-1127.

9 Wang L, Liu YJ, Xiao P *et al*: Chromosome 2q32 may harbor a QTL affecting BMD variation at different skeletal sites. *J Bone Miner Res* 2007; **22**: 1672-1678.

10 Pan F, Xiao P, Guo Y *et al*: Chromosomal regions 22q13 and 3p25 may harbor quantitative trait loci influencing both age at menarche and bone mineral density. *Hum Genet* 2008; **123**: 419-427.

11 Wang XL, Deng FY, Tan LJ *et al*: Bivariate whole genome linkage analyses for total body lean mass and BMD. *J Bone Miner Res* 2008; **23**: 447-452.

12 Liu XG, Liu YJ, Liu J *et al*: A bivariate whole genome linkage study identified genomic regions influencing both BMD and bone structure. *J Bone Miner Res* 2008; **23**: 1806-1814.

13 Lange C, Silverman EK, Xu X, Weiss ST, Laird NM: A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics* 2003; **4**: 195-206.

14 Lange C, van Steen K, Andrew T *et al*: A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. *Stat Appl Genet Mol Biol* 2004; **3**: Article17.

15 Jung J, Zhong M, Liu L, Fan R: Bivariate combined linkage and association mapping of quantitative trait loci. *Genet Epidemiol* 2008; **32**: 396-412.

16 Liu J, Pei Y, Papasian CJ, Deng HW: Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. *Genet Epidemiol* 2009; **33**: 217-227.

17 Pei YF, Zhang L, Liu J, Deng HW: Multivariate association test using haplotype trend regression. *Ann Hum Genet* 2009; **73**: 456-464.

18 Yang F, Tang Z, Deng H: Bivariate association analysis for quantitative traits using generalized estimation equation. *J Genet Genomics* 2009; **36**: 733-743.

19 Zhang L, Bonham AJ, Li J *et al*: Family-based bivariate association tests for quantitative traits. *PLoS One* 2009; **4**: e8133.

20 Zhang L, Pei YF, Li J, Papasian CJ, Deng HW: Univariate/multivariate genome-wide association scans using data from families and unrelated samples. *PLoS One* 2009; **4**: e6502.

21 Zellner A: An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association* 1962; **57**: 348-368.

22 McElroy MB: Goodness of Fit for Seemingly Unrelated Regressions. *Journal of Econometrics* 1977; **6**: 381-387.

23 Livshits G, Deng HW, Nguyen TV, Yakovenko K, Recker RR, Eisman JA: Genetics of bone mineral density: evidence for a major pleiotropic effect from an intercontinental study. *J Bone Miner Res* 2004; **19**: 914-923.

24 Pelat C, Van Pottelbergh I, Cohen-Solal M *et al*: Complex segregation analysis accounting for GxE of bone mineral density in European pedigrees selected through a male proband with low BMD. *Ann Hum Genet* 2007; **71**: 29-42.

25 Kaufman JM, Ostertag A, Saint-Pierre A *et al*: Genome-wide linkage screen of bone mineral density (BMD) in European pedigrees ascertained through a male

relative with low BMD values: evidence for quantitative trait loci on 17q21-23, 11q12-13, 13q12-14, and 22q11. *J Clin Endocrinol Metab* 2008; **93**: 3755-3762.

26 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559-575.

27 Sims AM, Shephard N, Carter K *et al*: Genetic analyses in a sample of individuals with high or low BMD shows association with multiple Wnt pathway genes. *J Bone Miner Res* 2008; **23**: 499-506.

28 Kung AW, Xiao SM, Cherny S *et al*: Association of JAG1 with bone mineral density and osteoporotic fractures: a genome-wide association study and follow-up replication studies. *Am J Hum Genet* 2010; **86**: 229-239.

29 Richards JB, Kavvoura FK, Rivadeneira F *et al*: Collaborative meta-analysis: associations of 150 candidate genes with osteoporosis and osteoporotic fracture. *Ann Intern Med* 2009; **151**: 528-537.

30 Rivadeneira F, Styrkarsdottir U, Estrada K *et al*: Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies. *Nat Genet* 2009; **41**: 1199-1206.

31 Styrkarsdottir U, Halldorsson BV, Gretarsdottir S *et al*: Multiple genetic loci for bone mineral density and fractures. *N Engl J Med* 2008; **358**: 2355-2365.

32 Doege H, Bocianski A, Scheepers A *et al*: Characterization of human glucose transporter (GLUT) 11 (encoded by SLC2A11), a novel sugar-transport facilitator specifically expressed in heart and skeletal muscle. *Biochem J* 2001; **359**: 443-449.

33 Bertocchini F, Ovitt CE, Conti A *et al*: Requirement for the ryanodine receptor type 3 for efficient contraction in neonatal skeletal muscles. *Embo J* 1997; **16**: 6956-6963.

Titles and legends to figures

Figure 1: Power rates at $\alpha=10^{-5}$ of SUR-based bivariate analysis and univariate analysis adjusted for multiple testing by Bonferroni correction (U_b), in samples of N=1 000 subjects and under various parameters settings: QTL heritability (h^2_1/h^2_2), sign of the induced genetic correlation (rG), residual correlation (rE).

(A) Power estimates against QTL heritability for moderately (rE=0.2) or strongly (rE=0.6) correlated traits, in randomly selected samples (Su)

(B) Power estimates under extreme selection (S1 or S2) for moderately (rE=0.2) or strongly (rE=0.6) correlated traits and QTL heritability ($h^2_1=h^2_2=0.005$)

Figure 2: Overlap in significance of results from bivariate and univariate (Best_U) association analysis.

(A) Top 100 hits in SUR-based bivariate association test: $-\log_{10}$ P-values of univariate analysis against $-\log_{10}$ P-values of SUR-based bivariate analysis

(B) Top 100 hits in univariate association test: $-\log_{10}$ P-values of SUR-based bivariate analysis against $-\log_{10}$ P-values of univariate analysis

Table 2: Mean (and sd) of the SUR-based bivariate association statistic (F test) in samples of N=1 000 subjects for various parameter settings: QTL heritability (h^2_1/h^2_2), sign of the induced genetic correlation (rG), residual correlation (rE), and sampling selection design.

			[1]Sampling		
			Su	S1	S2
r_E	r_G	h^2_1/h^2_2	µF (sd)	µF (sd)	µF (sd)
0.2	0	0.005/0	3.69 (2.61)	10.10 (4.34)	9.23 (4.18)
		0.01/0	6.17 (3.32)	18.94 (5.86)	17.97 (5.74)
		0.03/0	17.02 (6.08)	59.02 (10.56)	55.3 (10.32)
	+1	0.005/0.005	5.08 (2.99)	11.42 (4.86)	15.17 (5.41)
		0.005/0.01	7.45 (3.66)	14.07 (5.04)	19.08 (6.58)
		0.01/0.01	9.50 (4.42)	22.89 (6.86)	29.84 (7.89)
		0.03/0.03	26.90 (7.34)	72.04 (12.89)	92.88 (14.41)
	-1	0.005/0.005	7.26 (3.92)	13.91 (5.30)	9.57 (4.48)
		0.005/0.01	10.57 (4.36)	17.22 (5.65)	11.05 (4.88)
		0.01/0.01	13.69 (5.42)	27.72 (7.40)	19.29 (6.56)
		0.03/0.03	39.95 (9.40)	89.83 (13.86)	68.63 (13.23)
0.6	0	0.005/0	4.88 (3.04)	11.26 (4.51)	9.96 (4.47)
		0.01/0	8.79 (4.04)	22.54 (6.75)	19.78 (6.16)
		0.03/0	25.04 (7.34)	69.67 (11.85)	62.92 (11.30)
	+1	0.005/0.005	4.09 (2.69)	10.41 (4.49)	12.22 (4.78)
		0.005/0.01	6.36 (3.60)	12.67 (5.06)	15.60 (5.52)
		0.01/0.01	7.33 (3.85)	20.35 (6.34)	23.81 (6.76)
		0.03/0.03	20.42 (6.53)	63.56 (11.19)	73.11 (11.61)
	-1	0.005/0.005	13.70 (5.32)	20.94 (6.59)	16.02 (5.84)
		0.005/0.01	19.71 (6.52)	27.35 (7.55)	21.20 (6.92)
		0.01/0.01	26.06 (7.40)	42.83 (9.58)	34.26 (8.87)
		0.03/0.03	78.65 (14.17)	143.61 (19.15)	124.18 (17.77)

[1]Su: unselected sample; S1: sample selected on Y_1 distribution; S2: sample selected on Y_1 and Y_2 distributions.

22

Table 3: Mean (and sd) of the traditional univariate association statistic (t test) in samples of N=1 000 subjects for various parameter settings: QTL heritability (h^2_1/h^2_2), sign of the induced genetic correlation (rG), residual correlation (rE), and sampling selection design.

			[1]Sampling					
			SU		S1		S2	
r_E	r_G	h^2_1/h^2_2	Y_1 µt (sd)	Y_2 µt (sd)	Y_1 µt (sd)	Y_2 µt (sd)	Y_1 µt (sd)	Y_2 µt (sd)
0.2	0	0.005/0	2.26 (1.01)	-0.03 (1.00)	4.23 (0.99)	0.98 (1.02)	4.02 (1.00)	2.59 (1.05)
		0.01/0	3.15 (1.01)	-0.02 (0.97)	5.95 (0.98)	1.41 (1.00)	5.78 (0.96)	3.60 (1.00)
		0.03/0	5.53 (1.06)	0.00 (1.00)	10.69 (0.96)	2.33 (0.96)	10.34 (0.98)	6.34 (1.01)
	+1	0.005/0.005	2.23 (1.00)	2.20 (0.98)	4.15 (0.97)	3.21 (1.07)	4.99 (0.97)	4.83 (1.03)
		0.005/0.01	2.19 (0.97)	3.23 (1.02)	4.19 (0.97)	4.21 (0.99)	5.36 (1.02)	5.69 (1.07)
		0.01/0.01	3.18 (1.00)	3.19 (1.04)	5.96 (1.01)	4.72 (1.04)	7.06 (0.96)	6.93 (1.09)
		0.03/0.03	5.57 (1.01)	5.57 (0.98)	10.64 (1.01)	8.60 (1.07)	12.28 (0.96)	12.57 (1.13)
	-1	0.005/0.005	2.20 (1.01)	-2.26 (1.01)	4.18 (1.00)	-1.26 (0.99)	3.15 (0.99)	0.36 (1.02)
		0.005/0.01	2.26 (0.99)	-3.22 (0.97)	4.21 (0.92)	-2.16 (0.99)	2.69 (1.03)	-0.60 (0.99)
		0.01/0.01	3.18 (1.04)	-3.17 (1.04)	5.95 (0.98)	-1.95 (1.01)	4.56 (0.96)	0.44 (0.97)
		0.03/0.03	5.57 (1.01)	-5.58 (1.04)	10.68 (0.97)	-3.91 (1.01)	8.51 (0.96)	0.29 (0.95)
0.6	0	0.005/0	2.23 (1.00)	0.00 (0.96)	4.19 (0.97)	2.35 (0.98)	3.89 (0.99)	2.69 (0.97)
		0.01/0	3.13 (0.99)	-0.05 (0.98)	6.01 (1.00)	3.26 (0.97)	5.64 (0.97)	3.87 (0.98)
		0.03/0	5.55 (1.02)	0.01 (0.98)	10.62 (0.98)	5.44 (0.96)	10.08 (0.96)	6.63 (0.96)
	+1	0.005/0.005	2.22 (1.00)	2.24 (0.99)	4.17 (1.00)	4.06 (1.02)	4.58 (0.99)	4.59 (1.01)
		0.005/0.01	2.25 (1.01)	3.24 (1.03)	4.17 (1.00)	4.79 (1.03)	4.90 (0.94)	5.40 (0.99)
		0.01/0.01	3.17 (1.03)	3.18 (1.00)	5.97 (1.01)	5.83 (1.01)	6.52 (0.96)	6.58 (1.00)
		0.03/0.03	5.55 (1.01)	5.60 (1.03)	10.70 (0.98)	10.47 (0.99)	11.48 (0.93)	11.75 (0.99)
	-1	0.005/0.005	2.20 (1.02)	-2.30 (1.02)	4.13 (0.97)	0.48 (0.95)	3.23 (0.97)	0.81 (0.92)
		0.005/0.01	2.23 (1.00)	-3.21 (1.00)	4.15 (1.00)	-0.20 (1.00)	2.97 (0.99)	0.03 (0.97)
		0.01/0.01	3.10 (1.02)	-3.22 (1.01)	5.99 (0.96)	0.70 (0.97)	4.75 (0.97)	1.11 (0.95)
		0.03/0.03	5.52 (1.04)	-5.60 (1.05)	10.68 (0.95)	0.57 (0.93)	8.89 (0.92)	1.69 (0.88)

[1]Su: unselected sample; S1: sample selected on Y_1 distribution; S2: sample selected on Y_1 and Y_2 distributions.

Table 1: Outline of the main scenarios and varying parameter values in the bivariate data simulations

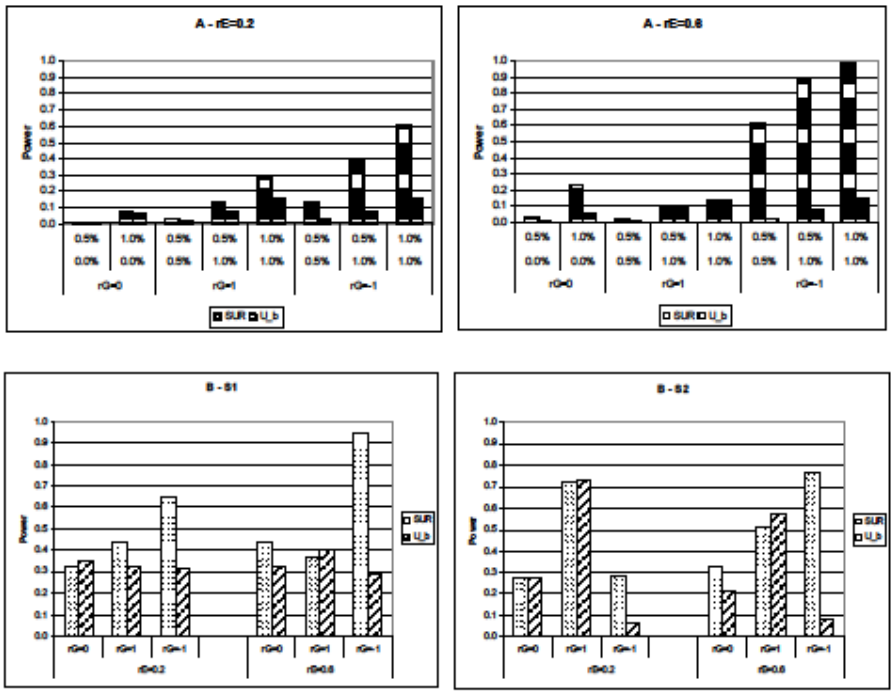
Main scenarios				Parameter values				
QTL Model	Heritability (h^2_j)	[1]effect size (aj)	rG	q	h^2_j	rE	N	Sampling Design
I. Null	$h^2_1 = h^2_2 = 0$	$a_1 = a_2 = 0$	0	0.1/0.4	$h^2_1 = 0$ & $h^2_2 = 0$	0.4 / 0.6	1000 / 300	Su / S1 / S2
II. No pleiotropic effect	$h^2_1 > 0$; $h^2_2 = 0$	$a_1 > 0$; $a_2 = 0$	0	0.1/0.4	h^2_1={0.5% / 1% / 3%} & h^2_2=0	0.4 / 0.6	1000 / 300	Su / S1 / S2
III. Pleiotropic effect								
a_j: same direction	$h^2_1 = h^2_2$	$a_1 = a_2$	+	0.1/0.4	$h^2_1 = h^2_2$= {0.5% / 1% / 3%}	0.4 / 0.6	1000 / 300	Su / S1 / S2
	$h^2_1 \neq h^2_2$	$a_1 \neq a_2$	+	0.1/0.4	h^2_1= 0.5% & h^2_2={1% / 3%) ; h^2_1=1% & h^2_2 =3%	0.4 / 0.6	1000 / 300	Su / S1 / S2
a_j: opposite direction	$h^2_1 = h^2_2$	$a_1 = -a_2$	-	0.1/0.4	$h^2_1 = h^2_2$= {0.5% / 1% / 3%}	0.4 / 0.6	1000 / 300	Su / S1 / S2
	$h^2_1 \neq h^2_2$	$a_1 \neq -a_2$	-	0.1/0.4	h^2_1= 0.5% & h^2_2={1% / 3%) ; h^2_1=1% & h^2_2 =3%	0.4 / 0.6	1000 / 300	Su / S1 / S2

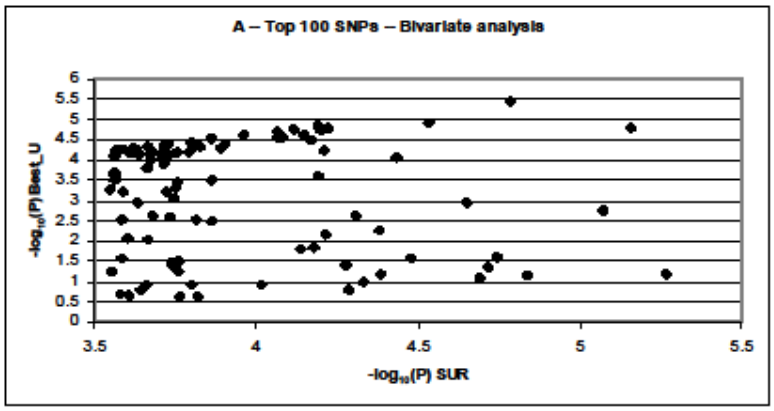
[1] $a_j = \sqrt{[h^2_j./2pq]}$, where q is the Minor Allele Frequency

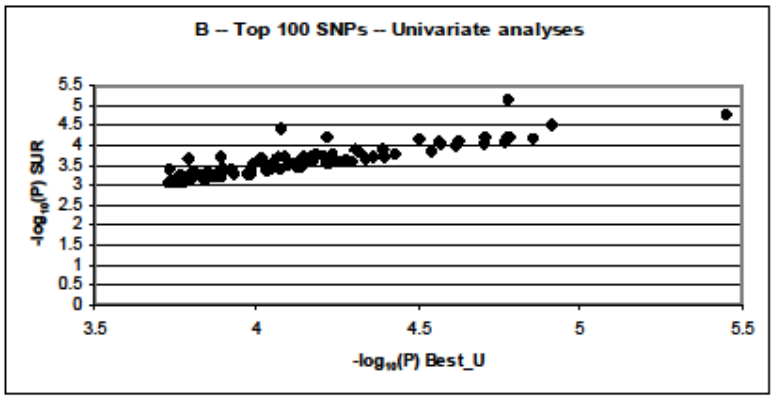
Table 4: Association results: Top 10 most associated SNPs from SUR-based bivariate or from separate univariate analysis of LS and FN BMD

	Chr. (Locus)	Closest Gene	Pos (bp)	SNP	[1]Min.	[2]MAF	SUR			Univ. LS			Univ. FN		
							P	[4]Rank	[3]R (%)	[5]P	[4]Rank	[5]R² (%)	[5]P	[4]Rank	[5]R² (%)
Top from SUR analysis	2q37.1	SP100	231 037 761	rs1649866	A	0.33	2.03E-05	8	3.41%	0.539	162064	0.12%	0.086	25794	0.95%
			231 042 007	rs1678160	G	0.33	1.45E-05	4	3.52%	0.574	172127	0.10%	0.072	21500	1.04%
	3q25	LEKR1	158 200 574	rs6799034	C	0.43	1.81E-05	6	3.45%	0.970	289752	0.00%	2.41E-02	7166	1.63%
	6q22.1	LOC643884	113 858 994	rs2046924	A	0.29	8.42E-06	3	3.69%	1.80E-03	607	3.09%	0.363	109135	0.27%
	6q25	TIAM2	155 533 083	rs998318	G	0.31	2.94E-05	10	3.30%	1.30E-05	2	5.94%	1.22E-05	1	5.98%
	12p13-p12	PZP - A2MP	9 254 198	rs1017301	C	0.34	2.24E-05	9	3.38%	0.223	67841	0.48%	1.13E-03	348	3.36%
	15q14-15	RYR3	31 680 776	rs2437143	C	0.38	6.97E-06	2	3.75%	8.41E-03	2635	2.21%	1.65E-05	3	5.80%
	19p13.11	FAM125A	17 392 450	rs2303680	G	0.41	1.92E-05	7	3.43%	0.743	222673	0.03%	4.66E-02	13886	1.27%
	22q11.2	SLC2A11	22 534 158	rs2275979	A	0.18	5.44E-06	1	3.85%	0.523	157609	0.13%	0.067	20101	1.08%
	22q13	LL22NC03	43 026 421	rs3935378	T	0.49	1.64E-05	5	3.48%	3.54E-06	1	6.69%	2.95E-05	8	5.47%
Top from Univ. of LS	1q32-41	CAMKIG	207 787 465	rs996146	T	0.11	6.39E-05	24	3.06%	1.65E-05	3	5.80%	5.19E-05	10	5.14%
	1q41	ESRRG	214 946 534	rs2813711	A	0.13	7.13E-05	29	3.02%	3.68E-05	10	5.34%	2.36E-05	5	5.59%
	2q11.2	AFF3	99 943 239	rs11887597	C	0.48	1.08E-04	36	2.90%	2.42E-05	7	5.60%	0.0005639	156	3.77%
	3q22.3	PIK3CB	139 883 989	rs531577	C	0.49	1.37E-04	40	2.85%	2.89E-05	9	5.53%	0.0004341	113	3.95%
	6q25.2	TIAM2	155 533 083	rs998318	G	0.31	2.94E-05	10	3.30%	1.30E-05	2	5.94%	1.22E-05	1	5.98%
	7q11.23	CCDC146	76 658 736	rs10252204	A	0.40	6.29E-05	23	3.07%	1.97E-05	5	5.72%	0.001236	382	3.32%
	10q21.2	LOC729184	62 158 387	rs1904418	G	0.32	7.68E-05	31	3.06%	1.71E-05	4	5.89%	0.0004438	115	3.97%
	16p13	NPM1P3	5 551 122	rs1969139	C	0.31	1.38E-04	42	2.82%	2.88E-05	8	5.48%	0.0003619	93	4.01%
	16q21	SLC38A7	57 262 362	rs9806843	G	0.43	8.55E-05	33	2.97%	1.99E-05	6	5.69%	0.0005906	169	3.73%
	22q13.31	LL22NC03	43 026 421	rs3935378	T	0.49	1.64E-05	5	3.48%	3.54E-06	1	6.69%	2.95E-05	8	5.47%
Top from Univ. Of FN	1p21.1	LOC126987	106 644 698	rs1330226	C	0.30	8.56E-05	34	2.97%	4.64E-05	18	5.20%	2.70E-05	6	5.52%
	1q32-41	CAMKIG	207 787 465	rs996146	T	0.11	6.39E-05	24	3.06%	1.65E-05	3	5.80%	5.19E-05	10	5.14%
	1q41	ESRRG	214 946 534	rs2813711	A	0.13	7.13E-05	29	3.02%	3.68E-05	10	5.34%	2.36E-05	5	5.59%
	2p21	C2orf34	44 846 991	rs11679997	T	0.15	6.73E-05	28	3.05%	2.99E-03	960	2.81%	3.14E-05	9	5.45%
	6q25.2	TIAM2	155 533 083	rs998318	G	0.31	2.94E-05	10	3.30%	1.30E-05	2	5.94%	1.22E-05	1	5.98%
	7p22.2	SDK1	3 411 837	rs6952184	C	0.09	5.97E-05	20	3.08%	4.09E-05	14	5.28%	1.68E-05	4	5.79%
	12q21.31	TSPAN19	83 921 735	rs1581563	G	0.16	8.36E-05	32	2.97%	4.31E-05	15	5.25%	2.73E-05	7	5.51%
	15q14-15	RYR3	31 680 776	rs2437143	C	0.38	6.97E-06	2	3.75%	8.41E-03	2635	2.21%	1.65E-05	3	5.80%
	20p12-p11.2	NXT1	23 219 822	rs4815192	T	0.46	6.45E-05	26	3.06%	9.81E-05	41	4.78%	1.39E-05	2	5.92%
	22q13.31	LL22NC03	43 026 421	rs3935378	T	0.49	1.64E-05	5	3.48%	3.54E-06	1	6.69%	2.95E-05	8	5.47%

[1]Minor allele; [2]Minor allele frequency; [3]r-square of the whole system taking into account the residual (co)variance matrix; [4]rank of the identified SNP; [5]r-square from linear regression; [5]Unadjusted univariate P values.


A - rE=0.2
A - rE=0.8
B - S1
B - S2
Power
SUR
U_b
rG=0
rG=1
rG=-1
rD=0.2
rD=0.8


A – Top 100 SNPs – Bivariate analysis
-log10(P) Best_U
-log10(P) SUR
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
3.5
4
4.5
5
5.5


B – Top 100 SNPs – Univariate analyses
-log10(P) SUR
-log10(P) Best_U
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
3.5
4
4.5
5
5.5

Annexe 3 : Liste des productions scientifiques

Articles publiés

Saint-Pierre A., Vitezica Z., Martinez M. (2009). *A comparative study of three methods for detecting association of quantitative traits in samples of related subjects.* BMC Proc., 7:122-128

Kaufman J.M., Ostertag A., **Saint-Pierre A**., Cohen-Solal M., Boland A., Van Pottelbergh I., Toye K., de Vernejoul M.C., Martinez M. (2008). *Genome-Wide linkage screen of bone mass density in European pedigrees ascertained through a male relative with low BMD values: Evidence for QTLs on 17q21-23, 11q12-13, 22q11 and 13q12-14*. J. Clin. Endocrinol. Metab., 93(10) :3755-62.

Saad M., Lesage S., **Saint-Pierre A**., Corvol J.C., Zelenika D., Lambert J.C., Vidailhet M., Mellick G.D., Lohmann E., Durif F., Pollak P., Damier P., Tison F., Silburn P.A., Tzourio C., Forlani S., Loriot M.A., Giroud M., Helmer C., Portet F., Amouyel P., Lathrop M., Elbaz A., Durr A., Martinez M. and Brice A (2010). *Genome-wide association study confirms BST1 and identifies a new locus on 12q24 as risk loci for Parkinson's disease in the European population*. Hum. Mol. Genet.

Articles en cours de révision

Saint-Pierre A., Kaufman J.M., Ostertag A., Toye K., Zelenika D., Cohen-Solal M., Lathrop M., de Vernejoul M.C., M. Martinez (2010). *Bivariate association analysis for quantitative traits in unrelated subjects: Performance under varying ascertainment schemes and application to a genome-wide association study of two BMD phenotypes in males with high or low BMD*. Eur. J. Hum. Genet.

Communication orale

Saint-Pierre A., Mangin B., Martinez M. (2009). *Bivariate linkage tests for quantitative traits: Assessing the null distributions in NEMO data*. European Mathematical Genetic Meeting, Munich (Allemagne), 14-15 Mai

Posters

Saint-Pierre A., Martinez M. (2009). *On the value of family data in genome-wide association studies for quantitative traits*. European Society of Human Genetics, Vienne (Autriche), 23-26 Mai; Eur. J. Hum. Genet; 17 (P08.13)

Saad M., **Saint-Pierre A**., Martinez M. (2009). *On the value of family data in genome-wide association studies for quantitative traits*. European Mathematical Genetic Meeting, Munich (Allemagne), 14-15 Mai, Annals of Human Genetics, 73, 669-669.

Saint-Pierre A., Mangin, B., Martinez, M. (2009). *Bivariate linkage analysis for mapping quantitative trait loci in pedigrees: Comparison of methods and assessment of the test statistics distributions in the NEMO study*. European Mathematical Genetic Meeting, Munich (Allemagne), 14-15 Mai, Annals of Human Genetics, 73, 658-658.

Saint-Pierre A., Martinez M. (2008). *A comparative study of three methods for detecting association of quantitative traits in samples of related subjects*. Genetic Analysis Workshop 16, Saint-Louis (Etats-Unis), 17-20 Septembre.

Saint-Pierre A., Martinez M. (2008). *On the detection of pleiotropic QTLs in non random and large pedigrees : empirical evaluation of different multitrait linkage tests.* International Genetic Epidemiology Society meeting, Saint-Louis (États-Unis), 15-16 Septembre, Genet. Epidemiol.; 32:7 (A148)

Saint-Pierre A., Martinez M. (2007). *Evaluation of different type of bivariate analysis in a genome-wide search for pleiotropic loci on two Bone Mineral Density quantitative traits.* International Genetic Epidemiology Society meeting, York (Angleterre), 7-10 Septembre, Genet. Epidemiol.; 31:6 (A130)

Bibliographie

"The R Project for Statistical Computing."
Abecasis, G. R., L. R. Cardon, et al. (2000). "A general test of association for quantitative traits in nuclear families." Am J Hum Genet **66**(1): 279-92.
Abecasis, G. R., W. O. Cookson, et al. (2001). "The power to detect linkage disequilibrium with quantitative traits in selected samples." Am J Hum Genet **68**(6): 1463-74.
Allison, D. B. (1997). "Transmission-disequilibrium tests for quantitative traits." Am J Hum Genet **60**(3): 676-90.
Allison, D. B., M. Heo, et al. (1998). "Extreme selection strategies in gene mapping studies of oligogenic quantitative traits do not always increase power." Hum Hered **48**(2): 97-107.
Allison, D. B., B. Thiel, et al. (1998). "Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages." Am J Hum Genet **63**(4): 1190-201.
Almasy, L. and J. Blangero (1998). "Multipoint quantitative-trait linkage analysis in general pedigrees." Am J Hum Genet **62**(5): 1198-211.
Almasy, L., T. D. Dyer, et al. (1997). "Bivariate quantitative trait linkage analysis: pleiotropy versus co-incident linkages." Genet Epidemiol **14**(6): 953-8.
Amos, C., M. de Andrade, et al. (2001). "Comparison of multivariate tests for genetic linkage." Hum Hered **51**(3): 133-44.
Amos, C. I. (1994). "Robust variance-components approach for assessing genetic linkage in pedigrees." Am J Hum Genet **54**(3): 535-43.
Aulchenko, Y. S., D. J. de Koning, et al. (2007). "Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis." Genetics **177**(1): 577-85.
Balemans, W., D. Foernzler, et al. (2002). "Lack of association between the SOST gene and bone mineral density in perimenopausal women: analysis of five polymorphisms." Bone **31**(4): 515-9.
Barrett, J. C., B. Fry, et al. (2005). "Haploview: analysis and visualization of LD and haplotype maps." Bioinformatics **21**(2): 263-5.
Baudoin, C., M. E. Cohen-Solal, et al. (2002). "Genetic and environmental factors affect bone density variances of families of men and women with osteoporosis." J Clin Endocrinol Metab **87**(5): 2053-9.
Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate : a practical and powerful approach to multiple testing." J R Stat Soc Ser B **57**(1): 289–300.
Boerwinkle, E., R. Chakraborty, et al. (1986). "The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods." Ann Hum Genet **50**(Pt 2): 181-94.
Bonney, G. E. (1984). "On the statistical determination of major
gene mechanisms in continuous human traits: regressive

models." Am J Med Genet **18**: 731-749.
Cardon, L. R. and J. I. Bell (2001). "Association study designs for complex diseases." Nat Rev Genet **2**(2): 91-9.
Cardon, L. R., C. Garner, et al. (2000). "Evidence for a major gene for bone mineral density in idiopathic osteoporotic families." J Bone Miner Res **15**(6): 1132-7.
Cirulli, E. T. and D. B. Goldstein (2010). "Uncovering the roles of rare variants in common disease through whole-genome sequencing." Nat Rev Genet **11**(6): 415-25.
Clerget-Darpoux, F., C. Bonaiti-Pellie, et al. (1986). "Effects of misspecifying genetic parameters in lod score analysis." Biometrics **42**(2): 393-9.
Cohen-Solal, M. and M. C. de Vernejoul (2004). "[Genetics of osteoporosis]." Rev Med Interne **25 Suppl 5**: S526-30.
Cohen-Solal, M. E., C. Baudoin, et al. (1998). "Bone mass in middle-aged osteoporotic men and their relatives: familial effect." J Bone Miner Res **13**(12): 1909-14.
Dalmasso, C., J. Pickrell, et al. (2007). "A mixture model approach to multiple testing for the genetic analysis of gene expression." BMC Proc **1 Suppl 1**: S141.
Davies, J. L., Y. Kawaguchi, et al. (1994). "A genome-wide search for human type 1 diabetes susceptibility genes." Nature **371**(6493): 130-6.
de Andrade, M., R. Gueguen, et al. (2002). "Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis." Genet Epidemiol **22**(3): 221-32.
de Andrade, M., T. J. Thiel, et al. (1997). "Assessing linkage on chromosome 5 using components of variance approach: univariate versus multivariate." Genet Epidemiol **14**(6): 773-8.
Deng, H. W., J. L. Li, et al. (1998). "Heterogeneity of bone mineral density across skeletal sites and its clinical implications." J Clin Densitom **1**(4): 339-53.
Deng, H. W., G. Livshits, et al. (2002). "Evidence for a major gene for bone mineral density/content in human pedigrees identified via probands with extreme bone mineral density." Ann Hum Genet **66**(Pt 1): 61-74.
Deng, H. W., H. Shen, et al. (2003). "Several genomic regions potentially containing QTLs for bone size variation were identified in a whole-genome linkage scan." Am J Med Genet A **119A**(2): 121-31.
Dequeker, J., J. Nijs, et al. (1987). "Genetic determinants of bone mineral content at the spine and radius: a twin study." Bone **8**(4): 207-9.
Devlin, B. and K. Roeder (1999). "Genomic control for association studies." Biometrics **55**(4): 997-1004.
Devoto, M., K. Shimoya, et al. (1998). "First-stage autosomal genome screen in extended pedigrees suggests genes predisposing to low bone mineral density on chromosomes 1p, 2p and 4q." Eur J Hum Genet **6**(2): 151-7.
Devoto, M., L. D. Spotila, et al. (2005). "Univariate and bivariate variance component linkage analysis of a whole-genome scan for loci contributing to bone mineral density." Eur J Hum Genet **13**(6): 781-8.
Dincel, E., A. Sepici-Dincel, et al. (2008). "Hip fracture risk and different gene polymorphisms in the Turkish population." Clinics (Sao Paulo) **63**(5): 645-50.
Duncan, E. L., L. R. Cardon, et al. (2003). "Site and gender specificity of inheritance of bone mineral density." J Bone Miner Res **18**(8): 1531-8.
Eisman, J. A. (1999). "Genetics of osteoporosis." Endocr Rev **20**(6): 788-804.
Elston, R. C., S. Buxbaum, et al. (2000). "Haseman and Elston revisited." Genet Epidemiol **19**(1): 1-17.

Elston, R. C. and J. Stewart (1971). "A general model for the genetic analysis of pedigree data." Hum Hered **21**(6): 523-42.
Falconer, D. S. and T. F. C. Mackay (1996). "Introduction to Quantitative Genetics." Ed 4. Longmans Green, Harlow, Essex, UK.
Ferrari, S. L., S. Deutsch, et al. (2005). "LRP5 gene polymorphisms and idiopathic osteoporosis in men." Bone **37**(6): 770-5.
Frazer, K. A., S. S. Murray, et al. (2009). "Human genetic variation and its contribution to complex traits." Nat Rev Genet **10**(4): 241-51.
Fulker, D. W., S. S. Cherny, et al. (1999). "Combined linkage and association sib-pair analysis for quantitative traits." Am J Hum Genet **64**(1): 259-67.
Grant, S. F., D. M. Reid, et al. (1996). "Reduced bone density and osteoporosis associated with a polymorphic Sp1 binding site in the collagen type I alpha 1 gene." Nat Genet **14**(2): 203-5.
Grundberg, E., E. M. Lau, et al. (2008). "Large-scale association study between two coding LRP5 gene polymorphisms and bone phenotypes and fractures in men." Osteoporos Int **19**(6): 829-37.
Gueguen, R., P. Jouanny, et al. (1995). "Segregation analysis and variance components analysis of bone mineral density in healthy families." J Bone Miner Res **10**(12): 2017-22.
Guo, Y., L. S. Zhang, et al. (2010). "IL21R and PTH may underlie variation of femoral neck bone mineral density as revealed by a genome-wide association study." J Bone Miner Res **25**(5): 1042-8.
Haseman, J. K. and R. C. Elston (1972). "The investigation of linkage between a quantitative trait and a marker locus." Behav Genet **2**(1): 3-19.
Havill, L. M., T. D. Dyer, et al. (2005). "The quantitative trait linkage disequilibrium test: a more powerful alternative to the quantitative transmission disequilibrium test for use in the absence of population stratification." BMC Genet **6 Suppl 1**: S91.
Heath, S. C. (1997). "Markov chain Monte Carlo segregation and linkage analysis for oligogenic models." Am J Hum Genet **61**(3): 748-60.
Hirschhorn, J. N. and M. J. Daly (2005). "Genome-wide association studies for common diseases and complex traits." Nat Rev Genet **6**(2): 95-108.
Hopper, J. L., R. M. Green, et al. (1998). "Genetic, common environment, and individual specific components of variance for bone mineral density in 10- to 26-year-old females: a twin study." Am J Epidemiol **147**(1): 17-29.
Hopper, J. L. and J. D. Mathews (1982). "Extensions to multivariate normal models for pedigree analysis." Ann Hum Genet **46**(Pt 4): 373-83.
Horst-Sikorska, W., M. Ignaszak-Szczepaniak, et al. (2008). "Association analysis of vitamin D receptor gene polymorphisms with bone mineral density in young women with Graves' disease." Acta Biochim Pol **55**(2): 371-80.
Hsu, Y. H., T. Niu, et al. (2006). "Variation in genes involved in the RANKL/RANK/OPG bone remodeling pathway are associated with bone mineral density at different skeletal sites in men." Hum Genet **118**(5): 568-77.
Hsu, Y. H., X. Xu, et al. (2007). "Large-scale genome-wide linkage analysis for loci linked to BMD at different skeletal sites in extreme selected sibships." J Bone Miner Res **22**(2): 184-94.
Hugot, J. P., M. Chamaillard, et al. (2001). "Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease." Nature **411**(6837): 599-603.
Jiang, C. and Z. B. Zeng (1995). "Multiple trait analysis of genetic mapping for quantitative trait loci." Genetics **140**(3): 1111-27.

Jung, J., M. Zhong, et al. (2008). "Bivariate combined linkage and association mapping of quantitative trait loci." Genet Epidemiol **32**(5): 396-412.
Kammerer, C. M., J. L. Schneider, et al. (2003). "Quantitative trait loci on chromosomes 2p, 4p, and 13q influence bone mineral density of the forearm and hip in Mexican Americans." J Bone Miner Res **18**(12): 2245-52.
Karasik, D., L. A. Cupples, et al. (2004). "Genome screen for a combined bone phenotype using principal component analysis: the Framingham study." Bone **34**(3): 547-56.
Karasik, D., J. Dupuis, et al. (2007). "Bivariate linkage study of proximal hip geometry and body size indices: the Framingham study." Calcif Tissue Int **81**(3): 162-73.
Karasik, D., R. H. Myers, et al. (2002). "Genome screen for quantitative trait loci contributing to normal variation in bone mineral density: the Framingham Study." J Bone Miner Res **17**(9): 1718-27.
Kastelan, D., Z. Grubic, et al. (2009). "The role of estrogen receptor-alpha gene TA polymorphism and aromatase gene TTTA polymorphism on peak bone mass attainment in males: is there an additive negative effect of certain allele combinations?" J Bone Miner Metab **27**(2): 198-204.
Kaufman, J. M., A. Ostertag, et al. (2008). "Genome-wide linkage screen of bone mineral density (BMD) in European pedigrees ascertained through a male relative with low BMD values: evidence for quantitative trait loci on 17q21-23, 11q12-13, 13q12-14, and 22q11." J Clin Endocrinol Metab **93**(10): 3755-62.
Koller, D. L., M. J. Econs, et al. (2000). "Genome screen for QTLs contributing to normal variation in bone mineral density and osteoporosis." J Clin Endocrinol Metab **85**(9): 3116-20.
Koller, D. L., S. Ichikawa, et al. (2010). "Genome-wide association study of bone mineral density in premenopausal European-American women and replication in African-American women." J Clin Endocrinol Metab **95**(4): 1802-9.
Koller, D. L., G. Liu, et al. (2001). "Genome screen for quantitative trait loci underlying normal variation in femoral structure." J Bone Miner Res **16**(6): 985-91.
Kong, A. and N. J. Cox (1997). "Allele-sharing models: LOD scores and accurate linkage tests." Am J Hum Genet **61**(5): 1179-88.
Kung, A. W., S. M. Xiao, et al. (2010). "Association of JAG1 with bone mineral density and osteoporotic fractures: a genome-wide association study and follow-up replication studies." Am J Hum Genet **86**(2): 229-39.
Lander, E. and L. Kruglyak (1995). "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results." Nat Genet **11**(3): 241-7.
Lander, E. S. and P. Green (1987). "Construction of multilocus genetic linkage maps in humans." Proc Natl Acad Sci U S A **84**(8): 2363-7.
Lange, C., E. K. Silverman, et al. (2003). "A multivariate family-based association test using generalized estimating equations: FBAT-GEE." Biostatistics **4**(2): 195-206.
Lange, C., K. van Steen, et al. (2004). "A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects." Stat Appl Genet Mol Biol **3**: Article17.
Lewontin, R. C. (1964). "The interaction of selection and linkage. I. General considerations; heterotic models." Genetics **49**: 49-67.
Liang, K. Y. and S. L. Zeger (1986). "Longitudinal data analysis using generalized linear models." Biometrika **73**: 13-22.
Liu, J., Y. Pei, et al. (2009). "Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations." Genet Epidemiol **33**(3): 217-27.

Liu, X. G., Y. J. Liu, et al. (2008). "A bivariate whole genome linkage study identified genomic regions influencing both BMD and bone structure." J Bone Miner Res **23**(11): 1806-14.

Liu, Y. Z., Y. F. Pei, et al. (2009). "Powerful bivariate genome-wide association analyses suggest the SOX6 gene influencing both obesity and osteoporosis phenotypes in males." PLoS One **4**(8): e6827.

Livshits, G., H. W. Deng, et al. (2004). "Genetics of bone mineral density: evidence for a major pleiotropic effect from an intercontinental study." J Bone Miner Res **19**(6): 914-23.

Mangin, B., P. Thoquet, et al. (1998). "Pleiotropic QTL analysis." Biometrics **54**: 88–99.

Mann, V. and S. H. Ralston (2003). "Meta-analysis of COL1A1 Sp1 polymorphism in relation to bone mineral density and osteoporotic fracture." Bone **32**(6): 711-7.

McElroy, M. B. (1977). "Goodness of Fit for Seemingly Unrelated Regressions." Journal of Econometrics **6**: 381-87.

McNemar, Q. (1947). " Note on the sampling error of the difference between correlated proportions or percentages." Psychometrika **12**: 153–157.

Mencej-Bedrac, S., J. Prezelj, et al. (2009). "The combinations of polymorphisms in vitamin D receptor, osteoprotegerin and tumour necrosis factor superfamily member 11 genes are associated with bone mineral density." J Mol Endocrinol **42**(3): 239-47.

Mencej, S., O. M. Albagha, et al. (2008). "Tumour necrosis factor superfamily member 11 gene promoter polymorphisms modulate promoter activity and influence bone mineral density in postmenopausal women with osteoporosis." J Mol Endocrinol **40**(6): 273-9.

Morton, N. E. (1955). "Sequential tests for the detection of linkage." Am J Hum Genet **7**(3): 277-318.

Nguyen, T. V., G. Livshits, et al. (2003). "Genetic determination of bone mineral density: evidence for a major gene." J Clin Endocrinol Metab **88**(8): 3614-20.

Niu, T., C. Chen, et al. (1999). "A genome-wide scan for loci linked to forearm bone mineral density." Hum Genet **104**(3): 226-33.

O'Connell, J. R. and D. E. Weeks (1995). "The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance." Nat Genet **11**(4): 402-8.

O'Connell, J. R. and D. E. Weeks (1998). "PedCheck: a program for identification of genotype incompatibilities in linkage analysis." Am J Hum Genet **63**(1): 259-66.

Pan, F., P. Xiao, et al. (2008). "Chromosomal regions 22q13 and 3p25 may harbor quantitative trait loci influencing both age at menarche and bone mineral density." Hum Genet **123**(4): 419-27.

Patterson, N., A. L. Price, et al. (2006). "Population structure and eigenanalysis." PLoS Genet **2**(12): e190.

Peacock, M., C. H. Turner, et al. (2002). "Genetics of osteoporosis." Endocr Rev **23**(3): 303-26.

Pei, Y. F., L. Zhang, et al. (2009). "Multivariate association test using haplotype trend regression." Ann Hum Genet **73**(Pt 4): 456-64.

Pelat, C., I. Van Pottelbergh, et al. (2007). "Complex segregation analysis accounting for GxE of bone mineral density in European pedigrees selected through a male proband with low BMD." Ann Hum Genet **71**(Pt 1): 29-42.

Pocock, N. A., J. A. Eisman, et al. (1987). "Genetic determinants of bone mass in adults. A twin study." J Clin Invest **80**(3): 706-10.

Pritchard, J. K., M. Stephens, et al. (2000). "Inference of population structure using multilocus genotype data." Genetics **155**(2): 945-59.
Purcell, S., B. Neale, et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." Am J Hum Genet **81**(3): 559-75.
Ralston, S. H., N. Galwey, et al. (2005). "Loci for regulation of bone mineral density in men and women identified by genome wide linkage scan: the FAMOS study." Hum Mol Genet **14**(7): 943-51.
Richards, J. B., F. K. Kavvoura, et al. (2009). "Collaborative meta-analysis: associations of 150 candidate genes with osteoporosis and osteoporotic fracture." Ann Intern Med **151**(8): 528-37.
Richards, J. B., F. Rivadeneira, et al. (2008). "Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study." Lancet **371**(9623): 1505-12.
Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases." Science **273**(5281): 1516-7.
Risch, N. J. (2000). "Searching for genetic determinants in the new millennium." Nature **405**(6788): 847-56.
Rivadeneira, F., U. Styrkarsdottir, et al. (2009). "Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies." Nat Genet **41**(11): 1199-206.
Saint Pierre, A., Z. Vitezica, et al. (2009). "A comparative study of three methods for detecting association of quantitative traits in samples of related subjects." BMC Proc **3 Suppl 7**: S122.
Schork, N. J., S. S. Murray, et al. (2009). "Common vs. rare allele hypotheses for complex diseases." Curr Opin Genet Dev **19**(3): 212-9.
Self, S. G. and K. Y. Liang (1987). "Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard condtitions." jasa **82**: 605-610.
Sham, P. C. and S. Purcell (2001). "Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs." Am J Hum Genet **68**(6): 1527-32.
Shen, H., Y. Liu, et al. (2005). "Nonreplication in genetic studies of complex diseases--lessons learned from studies of osteoporosis and tentative remedies." J Bone Miner Res **20**(3): 365-76.
Sigurdsson, G., B. V. Halldorsson, et al. (2008). "Impact of genetics on low bone mass in adults." J Bone Miner Res **23**(10): 1584-90.
Sims, A. M., N. Shephard, et al. (2008). "Genetic analyses in a sample of individuals with high or low BMD shows association with multiple Wnt pathway genes." J Bone Miner Res **23**(4): 499-506.
Spielman, R. S. and W. J. Ewens (1998). "A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test." Am J Hum Genet **62**(2): 450-8.
Spielman, R. S., R. E. McGinnis, et al. (1993). "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)." Am J Hum Genet **52**(3): 506-16.
Streeten, E. A., D. J. McBride, et al. (2006). "Quantitative trait loci for BMD identified by autosome-wide linkage scan to chromosomes 7q and 21q in men from the Amish Family Osteoporosis Study." J Bone Miner Res **21**(9): 1433-42.
Styrkarsdottir, U., B. V. Halldorsson, et al. (2008). "Multiple genetic loci for bone mineral density and fractures." N Engl J Med **358**(22): 2355-65.

Styrkarsdottir, U., B. V. Halldorsson, et al. (2009). "New sequence variants associated with bone mineral density." Nat Genet **41**(1): 15-7.

Tan, L. J., Y. Z. Liu, et al. (2008). "Evidence for major pleiotropic effects on bone size variation from a principal component analysis of 451 Caucasian families." Acta Pharmacol Sin **29**(6): 745-51.

Tang, Z. H., P. Xiao, et al. (2007). "A bivariate whole-genome linkage scan suggests several shared genomic regions for obesity and osteoporosis." J Clin Endocrinol Metab **92**(7): 2751-7.

Terwilliger, J. D., M. Speer, et al. (1993). "Chromosome-based method for rapid computer simulation in human genetic linkage analysis." Genet Epidemiol **10**(4): 217-24.

Uitterlinden, A. G., P. P. Arp, et al. (2004). "Polymorphisms in the sclerosteosis/van Buchem disease gene (SOST) region are associated with bone-mineral density in elderly whites." Am J Hum Genet **75**(6): 1032-45.

van Meurs, J. B., T. A. Trikalinos, et al. (2008). "Large-scale analysis of association between LRP5 and LRP6 variants and osteoporosis." Jama **299**(11): 1277-90.

Van Pottelbergh, I., S. Goemaere, et al. (2001). "Association of the type I collagen alpha1 Sp1 polymorphism, bone density and upper limb muscle strength in community-dwelling elderly men." Osteoporos Int **12**(10): 895-901.

Van Pottelbergh, I., S. Goemaere, et al. (2003). "Deficient acquisition of bone during maturation underlies idiopathic osteoporosis in men: evidence from a three-generation family study." J Bone Miner Res **18**(2): 303-11.

Videman, T., E. Levalahti, et al. (2007). "Heritability of BMD of femoral neck and lumbar spine: a multivariate twin study of Finnish men." J Bone Miner Res **22**(9): 1455-62.

Wang, K. (2003). "Mapping quantitative trait loci using multiple phenotypes in general pedigrees." Hum Hered **55**(1): 1-15.

Wang, L., Y. J. Liu, et al. (2007). "Chromosome 2q32 may harbor a QTL affecting BMD variation at different skeletal sites." J Bone Miner Res **22**(11): 1672-8.

Wang, W. Y., B. J. Barratt, et al. (2005). "Genome-wide association studies: theoretical and practical concerns." Nat Rev Genet **6**(2): 109-18.

Wang, X., C. M. Kammerer, et al. (2007). "Pleiotropy and heterogeneity in the expression of bone strength-related phenotypes in extended pedigrees." J Bone Miner Res **22**(11): 1766-72.

Wang, X. L., F. Y. Deng, et al. (2008). "Bivariate whole genome linkage analyses for total body lean mass and BMD." J Bone Miner Res **23**(3): 447-52.

Willaert, A., I. Van Pottelbergh, et al. (2008). "A genome-wide linkage scan for low spinal bone mineral density in a single extended family confirms linkage to 1p36.3." Eur J Hum Genet **16**(8): 970-6.

Wilson, S. G., P. W. Reed, et al. (2003). "Comparison of genome screens for two independent cohorts provides replication of suggestive linkage of bone mineral density to 3p21 and 1p36." Am J Hum Genet **72**(1): 144-55.

Xiong, D. H., X. G. Liu, et al. (2009). "Genome-wide association and follow-up replication studies identified ADAMTS18 and TGFBR3 as bone mass candidate genes in different ethnic groups." Am J Hum Genet **84**(3): 388-98.

Xiong, D. H., H. Shen, et al. (2006). "Robust and comprehensive analysis of 20 osteoporosis candidate genes by very high-density single-nucleotide polymorphism screen among 405 white nuclear families identified significant association and gene-gene interaction." J Bone Miner Res **21**(11): 1678-95.

Xiong, D. H., J. T. Wang, et al. (2007). "Genetic determination of osteoporosis: lessons learned from a large genome-wide linkage study." Hum Biol **79**(6): 593-608.
Xu, X. H., S. S. Dong, et al. (2010). "Molecular genetic studies of gene identification for osteoporosis: the 2009 update." Endocr Rev **31**(4): 447-505.
Yalcin, B., S. A. Willis-Owen, et al. (2004). "Genetic dissection of a behavioral quantitative trait locus shows that Rgs2 modulates anxiety in mice." Nat Genet **36**(11): 1197-202.
Yan, H., Y. J. Liu, et al. (2009). "Comparison of whole genome linkage scans in premenopausal and postmenopausal women: no bone-loss-specific QTLs were implicated." Osteoporos Int **20**(5): 771-7.
Yang, F., Z. Tang, et al. (2009). "Bivariate association analysis for quantitative traits using generalized estimation equation." J Genet Genomics **36**(12): 733-43.
Yang, T. L., L. J. Zhao, et al. (2006). "Genetic and environmental correlations of bone mineral density at different skeletal sites in females and males." Calcif Tissue Int **78**(4): 212-7.
Zellner, A. (1962). "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias." Journal of the American Statistical Association **57**(298): 348-68.
Zhang, H., K. Sol-Church, et al. (2009). "High resolution linkage and linkage disequilibrium analyses of chromosome 1p36 SNPs identify new positional candidate genes for low bone mineral density." Osteoporos Int **20**(2): 341-6.
Zhang, L., A. J. Bonham, et al. (2009). "Family-based bivariate association tests for quantitative traits." PLoS One **4**(12): e8133.
Zhang, L., Y. F. Pei, et al. (2009). "Univariate/multivariate genome-wide association scans using data from families and unrelated samples." PLoS One **4**(8): e6502.
Zhang, Z. X., S. F. Lei, et al. (2009). "Bivariate genome-wide linkage analysis for traits BMD and AAM: effect of menopause on linkage signals." Maturitas **62**(1): 16-20.