Interactions gène-gène et gène-environnement dans les études génétiques de maladies multifactorielles : application à l'asthme et l'atopie

Pierre-Emmanuel Sugier

Sorbonne Université

Ecole doctorale 393 Pierre Louis de Santé Publique à Paris :

Epidémiologie et Sciences de l'Information Biomédicale

UMR 946 – Variabilité Génétique et Maladies Humaines

Interactions gène-gène et gène-environnement dans les études génétiques de maladies multifactorielles – application à l'asthme et l'atopie

Par Pierre-Emmanuel Sugier

Thèse de doctorat d'Epidémiologie Génétique

Présentée et soutenue publiquement le 13 décembre 2018

Devant un jury composé de :

Delacourt Christophe	Professeur, Hôpital Necker-Enfants malades	Rapporteur
Bouatia-Naji Nabila	Chargée de recherche, INSERM	Rapporteur
Annesi-Maesano Isabella	Directrice de recherche, INSERM	Examinatrice
Role François	Maître de conférence, Université Paris-Descartes	Examinateur
Ahmed Ismail	Chargé de recherche, INSERM	Examinateur
Demenais Florence	Directrice de recherche, INSERM	Directrice de thèse
Bouzigon Emmanuelle	Chargée de recherche, INSERM	Co-directrice de thèse

Remerciements

Production scientifique

Articles originaux issus du travail de thèse

1. Sugier PE, Brossard M[*], Sarnowski C[*], Vaysse A, Morin A, Pain L, Margaritte-Jeannin P, Dizier M-H, Cookson W.O.C.M., Lathrop M, Moffatt MF, Laprise C[‡], Demenais F[‡], Bouzigon E[‡]. A novel role for cilia function in atopy: ADGRV1 and DNAH5 interactions. ***J Allergy Clin Immunol*** 2018; 141(5):1659-1667.e11

2. Sarnowski C, Sugier P-E, Granell R, Jarvis D, Dizier M-H, Ege M, Imboden M, Laprise C, Khusnutdinova EK, Freidin MB, Cookson WOC, Moffatt M, Lathrop M, Siroux V, Ogorodova LM, Karunas AS, James A, Probst-Hensch NM, von Mutius E, Pin I, Kogevinas M, Henderson AJ, Demenais F, Bouzigon E. Identification of a new locus at 16q12 associated with time-to-asthma onset. ***J Allergy Clin Immunol*** 2016; 138(4):1071-1080

3. Sugier PE, Sarnowski C, Granell R, Laprise C, Ege M, Margaritte-Jeannin P, Dizier MH, Minelli C, Moffatt M, Lathrop M, Cookson WOC, Henderson AJ, von Mutius E, Kogevinas M, Demenais F[*], Bouzigon E[*]. Genome-wide interaction study of early-life environmental tobacco smoke exposure on time-to-asthma onset in childhood. Soumis

Communications dans le cadre de congrès internationaux

1. Sugier PE, Sarnowski C, Granell R, Jarvis D, Ege M.J, Laprise C, von Mutius E, Dizier MH, Henderson A.J, Kogevinas M, Demenais F, Bouzigon E. Genome-wide interaction study of environmental tobacco smoke exposure in early life on time-to-asthma onset in childhood. *European Respiratory Society Meeting*, 15-19 September 2018, Paris, France - ***Eur Respir J*** 2018; sous presse

2. Sugier P-E[*], Sarnowski C[*], Granell R, Jarvis D, Ege M.J, Laprise C, von Mutius E, Dizier MH, Henderson A.J, Kogevinas M, Demenais F, Bouzigon E. Interaction of genetic variants with secondhand smoke exposure in early life on asthma time to asthma onset. *International Genetic Epidemiology Society (IGES)*, Cambridge, United-Kingdom, 9-11 September 2017 - ***Genet Epidemiol.*** 2017;41: 693-694

3. Sugier PE, Brossard M, Vaysse A, Sarnowski C, Dizier MH, Lathrop M, Laprise C, Demenais F, Bouzigon E. Combining text mining and epistasis analyses identifies new atopy genes. *International Genetic Epidemiology Society (IGES)*, Baltimore, USA, 4-6 October 2015 - ***Genet Epidemiol.*** 2015; 39: 582-583

4. Sugier PE, Brossard M, Vaysse A, Sarnowski C, Dizier MH, Lathrop M, Laprise C, Demenais F, Bouzigon E. Integration of text-mining and epistasis analyses identifies new genes underlying atopy. *65th Annual Meeting of The American Society of Human Genetics (ASHG)*, Baltimore, USA, 6-10 October 2015-– Abstract/Program 1410W through: http://www.ashg.org/2015meeting/pages/abstracts.shtml

5. Sugier PE, Vaysse A, Sarnowski C, Loucoubar C, Margaritte-Jeannin P, Dizier MH, Lathrop M, Demenais F, Bouzigon E. Integration of gene-based and text mining analyses to discover genes underlying atopy. *Capita Selecta in Complex Disease Analysis workshop*, Liège, Belgium, 24-26 November 2014. Abstract through: http://www.statgen.ulg.ac.be/CSCDA2014/abstracts.html

Sommaire

CHAPITRE I - INTRODUCTION

1. Préambule : la génétique des maladies multifactorielles

La génétique statistique est la science qui s'intéresse au développement de méthodologies statistiques pour l'analyse des données génétiques. En particulier dans le cadre de la **génétique humaine**, c'est l'étude des déterminants génétiques chez l'Homme ainsi que des facteurs environnementaux pouvant interagir avec ces déterminants, dans le but d'identifier et comprendre les effets de ces facteurs sur le risque de la maladie. Mon travail de thèse a porté en particulier sur les **maladies multifactorielles**, qui sont des maladies complexes résultant de multiples facteurs génétiques et environnementaux et d'interactions entre ces facteurs. Ces maladies sont fréquentes dans la population et représentent un enjeu majeur de santé publique. C'est le cas des maladies cardiovasculaires, des cancers, du diabète, de l'asthme, des maladies allergiques, inflammatoires, auto-immunes, infectieuses et psychiatriques notamment.

Les méthodologies utilisées pour identifier les gènes, et plus précisément, les variants génétiques impliqués dans les maladies ont évolué aux cours de ces dernières années grâce aux avancées technologiques et à l'accumulation des connaissances biologiques. Le principe général de ces méthodes est d'évaluer la corrélation entre le(s) marqueur(s) génétique(s) et le phénotype étudié (maladie ou trait quantitatif). Les marqueurs génétiques apportent deux types d'information : celle de **la liaison génétique, au niveau familial** et celle de **l'association, au niveau de la population**. Plusieurs revues de la littérature exposent plus en détail ces différentes stratégies [1,2].

Les premières études prenant en compte l'information sur les marqueurs génétiques ont été conduites à l'aide d'analyses de liaison génétique, basées sur le principe que deux marqueurs génétiques localisés physiquement proches sur un chromosome tendent à rester liés durant la méiose. L'objectif de ces analyses est de localiser les régions contenant les gènes responsables du trait ou de la maladie sur le génome dans des échantillons de familles. Les analyses de liaison ont permis d'identifier un très grand nombre de gènes impliqués dans les **maladies monogéniques** [3]. Elles se sont aussi avérées puissantes pour la découverte de marqueurs génétiques impliqués dans certains sous-types de maladies multifactorielles, comme, par exemple, les formes familiales de cancers (cancer du sein, cancer du côlon, mélanome…), ou certaines maladies neurodégénératives

comme la maladie d'Alzheimer et de Parkinson. En revanche, cette méthode semble moins adaptée pour l'étude des maladies multifactorielles dans leur ensemble, dans lesquelles les effets des facteurs génétiques impliqués sont multiples et modestes (faible corrélation trait-marqueur) et les mécanismes complexes (interactions entre gènes, interactions gène-environnement, hétérogénéité génétique…). Par ailleurs, la prévalence relativement élevée de ces pathologies suggère que les allèles à risque sont fréquents dans la population générale (l'hypothèse « Common Variants-Common Diseases » à l'origine des études d'associations pangénomiques) bien que des variants rares puissent également jouer un rôle.

En 1996, une étude théorique menée par Risch et Merikangas [4] a montré que, dans le cas où les variants ont une fréquence peu élevée et des effets relativement importants sur le trait étudié, les études de liaison génétique sont puissantes, alors que, dans le cas où les variants impliqués sont fréquents et ont des petits effets, les études d'association maladie/marqueur sont plus puissantes. Les études d'association en population représentent une alternative pertinente pour l'étude des maladies multifactorielles. Elles sont fondées sur l'existence de dépendances statistiques appelées **déséquilibre de liaison, ou DL (Encart 1)**, généralement observées entre marqueurs génétiques proches (distances inférieures à 50-100 kilobases) sur l'ADN. Pour un trait binaire (statut malade/non malade), le principe est d'identifier les marqueurs génétiques pour lesquels les fréquences alléliques diffèrent significativement entre les malades et les témoins. Pour un trait quantitatif, il s'agit d'identifier les marqueurs qui sont associés à l'augmentation ou à la diminution du trait étudié (exemple : indice de masse corporelle, taille, tension artérielle…). La première vague de ces études d'association portait sur des marqueurs proches de gènes candidats, gènes sélectionnés *a priori* selon des hypothèses biologiques. Bien qu'il existe quelques succès comme l'identification de l'association entre l'allèle e4 du gène APOE et la forme à âge de début tardif de la maladie d'Alzheimer ou des variants du gène MC1R avec le mélanome, ces études se sont avérées peu concluantes. Ceci était principalement dû à la taille souvent petite des échantillons étudiés, au petit nombre de marqueurs génotypés et à l'absence de réplication des résultats dans des échantillons indépendants. Au cours des années 2000, les avancées des **technologies de génotypage à haut débit** (« puces à ADN ») ont permis d'accéder à une grande part de la variabilité du génome entier à l'aide de centaines de milliers (et actuellement millions) de marqueurs génétiques. Ces marqueurs sont des polymorphismes d'un seul nucléotide, appelé *Single Nucleotide Polymorphism* ou **SNP** (**Figure 1.1**). Dans le cadre de ma thèse, l'analyse de SNPs

répartis sur l'**ensemble du génome** a été rendue possible par des **puces de génotypage** qui intègrent des sondes ciblant la séquence entourant le SNP et permettent de reconnaitre les deux allèles correspondants. Ces avancées technologiques et la mise en place de projets internationaux comme les projets « HapMap » et « 1 000 Génomes » ont ouvert la voie aux études d'association pangénomiques, c'est-à-dire la recherche d'association sur le génome entier (Genome Wide Association Study, que nous appellerons GWAS par la suite). Ces études ont pris un essor important à partir des années 2005.

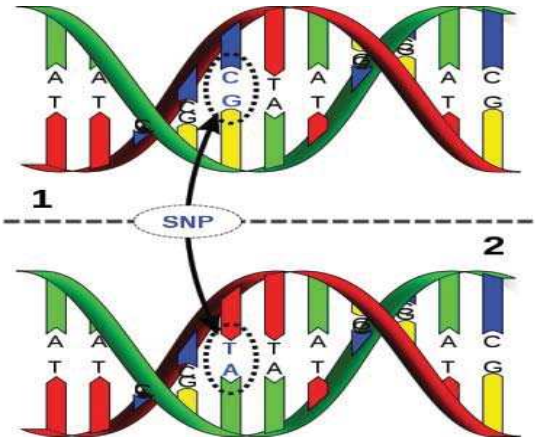


Figure 1.1 : Exemple d'un SNP. Ici, les molécules d'ADN 1 et 2 diffèrent à un locus donné d'une seule paire de base. On peut remarquer que le nucléotide C sur l'un des bruns de l'ADN 1 est un T sur le même brun de l'ADN 2 : c'est un polymorphisme C/T. Chaque individu possède deux copies ou **allèles** d'un même SNP (transmises par ses deux parents), cette combinaison des deux allèles est appelée **génotype** : il est dit **homozygote** si les deux allèles sont identiques, ou **hétérozygote** s'ils sont différents. (http://isogg.org/wiki/Single-nucleotide_polymorphism)

Encart 1 : Le déséquilibre de liaison

Le déséquilibre de liaison (noté DL) représente une corrélation entre deux SNPs dans une population. Considérons deux SNPs ayant les allèles A et a pour le premier, B et b pour le second et notons f_A, f_a ($f_a = 1 - f_A$), f_B et f_b ($f_b = 1 - f_B$) leur fréquence allélique respective. Supposons que A et B sont les allèles majeurs à chacun de ces locus. Pour un brin chromosomique donné, il existe quatre combinaisons d'allèles possibles (ou haplotypes) : AB, ab, Ab et bB, de fréquences respectives f_{AB}, f_{ab}, f_{Ab} et f_{bB}. Le niveau de DL entre les allèles A et B peut ainsi être quantifié comme suit :

$$D_{AB} = f_{AB} - f_A \times f_B$$

On retrouve facilement :

$$D_{AB} = -D_{Ab} = -D_{aB} = D_{ab}$$

Chacune de ces valeurs peut alors être utilisée pour caractériser le DL entre les allèles à ces deux loci :

- $D_{AB} = 1$, i.e. $f_{AB} = f_A \times f_B$: A et B sont dits en DL complet
- $D_{AB} \neq 0$, i.e. $f_{AB} \neq f_A \times f_B$: A et B sont en DL partiel
- $D_{AB} > 0$, i.e. $f_{AB} > f_A \times f_B$: A et B sont préférentiellement associés
- $D_{AB} < 0$, i.e. $D_{Ab} > 0$: A et b sont préférentiellement associés

Cependant, D dépend fortement des fréquences alléliques, ce qui peut rendre plus compliquée la comparaison de DL entre différentes paires de SNPs. Deux autres mesures normalisées par rapport aux fréquences alléliques ont été proposées :

- La mesure du D' [229]:

$$D' = \begin{cases} \dfrac{D_{AB}}{\min(f_A \times f_B, f_a \times f_b)}, & D_{AB} < 0 \\ \dfrac{D_{AB}}{\min(f_a \times f_B, f_A \times f_b)}, & D_{AB} > 0 \end{cases}$$

- Le coefficient de corrélation de Pearson, r^2 [230]:

$$r^2 = \frac{{D_{AB}}^2}{f_a \times f_A, f_b \times f_B}$$

Les valeurs de D' sont comprises entre – 1 et 1. Lorsque $|D'| = 1$, on parle de DL complet entre les deux SNPs : cela indique qu'au moins un des quatre haplotypes possibles est absent. Le coefficient de corrélation de Pearson r^2, est compris entre 0 et 1. Lorsque $r^2 = 1$, on parle d'un DL parfait (ou total), alors nécessairement $f_A = f_B$ et $f_{Ab} = f_{aB} = 0$ (et donc $|D'| = 1$). Dans ce cas, les deux SNPs portent exactement la même information.

2. Etudes d'associations pangénomiques simple marqueur

Les analyses d'association cherchent à identifier des corrélations entre marqueurs génétiques et un trait phénotypique par le biais de méthodes mathématiques. Ces analyses cherchent à estimer l'effet moyen d'un SNP sur le phénotype étudié et à tester cet effet par un test statistique. Parmi l'ensemble des méthodes d'analyses qui ont été proposées [1], les **méthodes de régression** sont les plus largement utilisées. Le type de régression employé varie selon le type du trait étudié : régression linéaire pour des phénotypes quantitatifs, régression logistique pour des phénotypes binaires (individu malade ou non malade), mais d'autres modèles de régression existent, par exemple ceux basés sur le modèle de Cox peuvent être utilisés pour intégrer la notion de temps écoulé avant qu'un événement ne survienne (comme la survenue de l'asthme par exemple).

L'un des atouts de ces modèles de régression est qu'ils peuvent permettre de modéliser l'effet de chaque SNP sur le trait étudié tout en ajustant pour l'effet de **cofacteurs** associés au phénotype étudié. Les connaissances accumulées guident généralement le choix des variables à considérer mais l'on peut citer parmi les plus couramment utilisées : l'âge, le sexe, des facteurs environnementaux (l'exposition au tabac, au soleil, informations géographiques, …) ou encore des cofacteurs issus d'analyses préliminaires tels que les composantes principales décrivant une part importante de la variance des données génotypiques [5], permettant de prendre en compte des **stratifications de populations**. Les stratifications génétiques dans une population sont une source d'hétérogénéité pouvant impliquer des biais statistiques et conduire à de mauvaises conclusions : une association entre un marqueur et la maladie à tort (**faux positifs**), ou une absence d'association pourtant existante (**faux négatifs**) [6].

Des effets **additifs, récessifs ou dominants** peuvent être modélisés pour le SNP. Un allèle peut-être dominant ou récessif selon la façon dont il s'exprime sur un phénotype donné pour un génotype donné. Ainsi, l'allèle A est dominant si la probabilité de manifester le phénotype en question (c'est-à-dire la **pénétrance**) est la même chez les sujets AA que chez les sujets Aa, mais supérieure à celle des sujets aa. L'allèle A est au contraire récessif si la pénétrance est supérieure chez les sujets AA par rapport aux sujets Aa et aa. Le modèle additif suppose que l'effet des génotypes sur le risque de développer la maladie est proportionnel au nombre d'allèles à risque (0, 1, 2 pour un SNP). Le **modèle général** qui ne fait aucune hypothèse sur les trois génotypes observés pour le

SNP requiert l'estimation de deux paramètres : un effet additif et un effet dominant (ou écart à l'additivité). Le modèle général requiert un degré de liberté supplémentaire pour tester l'effet du marqueur sur le trait, par rapport aux trois autres modèles. Ce modèle peut donc entraîner une perte de puissance s'il est testé directement et si le vrai modèle est additif, récessif ou dominant. Après de nombreuses discussions dans la littérature, c'est le modèle additif qui a été retenu pour la grande majorité des analyses pangénomiques simple-marqueurs (décrites plus loin) [1]. L'allèle mineur (allèle dont la fréquence est la plus petite dans l'échantillon parmi les deux allèles d'un SNP, appelé parfois MAF pour *minor allele frequency*) est alors souvent considéré comme allèle à risque, dont on cherche à estimer l'effet : cet effet peut augmenter le risque de contracter la maladie ou au contraire, le diminuer (effet protecteur). Cependant, l'allèle à risque peut être aussi pris comme l'allèle alternatif par rapport à la séquence de référence sans considération de fréquence.

Les premières études d'associations se sont focalisées sur l'analyse de **gènes candidats** dans les années 90s et première moitié des années 2000s. Ces analyses d'association ont l'avantage d'être relativement peu coûteuses, demandent peu de ressources informatiques, et sont rapides à effectuer [7]. Cependant, ces approches dépendent fortement de l'**hypothèse biologique** cible, ou du fait que le gène candidat soit dans une région de liaison préalablement définie. Les résultats issus de ces analyses dépendent donc fortement de la capacité à générer des hypothèses sous forme de gènes candidats à partir de l'information déjà connue : ce qui est particulièrement difficile dans le cas de nombreux SNPs impliqués avec des effets relativement faibles.

L'analyse pangénomique simple-marqueur communément appelée **GWAS**, pour *Genome-Wide Association Study*, permet quant à elle de s'affranchir de la nécessité de générer des hypothèses *a priori* d'association, en testant l'ensemble des SNPs le long du génome : c'est **un criblage du génome** (méthode agnostique). Cependant, ces analyses nécessitent de considérer des matrices de données très importantes où le nombre de sujets est très inférieur au nombre de SNPs à analyser. En effet, le nombre de SNPs génotypés à l'aide de puces de génotypage peut atteindre des centaines de milliers voir **plusieurs millions** grâce à l'utilisation de méthodes d'imputations permettant d'augmenter le pool de données génétiques d'un échantillon d'étude, ou encore grâce à l'évolution des technologies de séquençage. Afin de pallier ce problème, les GWAS consistent à classiquement estimer et tester l'effet un à un de chaque SNP sur le trait étudié, comme celui conduit par Klein *et al.* en 2005 [8]. Ces associations peuvent-être testées par des tests de Wald, ou par tests de rapport de vraisemblances. C'est une stratégie d'étude adaptée pour détecter des variants fréquents associés

au phénotype étudié. Cependant elle nécessite de bien contrôler les niveaux des erreurs de type I et de type II. En effet, l'analyse d'un grand nombre de SNPs implique d'effectuer de nombreux tests statistiques et nécessite d'utiliser une **correction pour les tests multiples** pour contrôler le taux d'erreur de type I (**Encart 2 – méthodes d'ajustement pour les tests multiples**). Ceci peut réduire drastiquement la puissance de détection d'une association : ces analyses nécessitent donc de travailler sur un **nombre important de sujets**. Les autres éléments importants qui permettent d'assurer le succès des GWASs sont la nécessité de prendre en compte les différentes structures de DL entre marqueurs génétiques, de considérer une couverture optimale des variations génétiques sur l'ensemble du génome, d'effectuer des contrôles de qualité stricts des données génotypées et de répliquer les résultats dans des études indépendantes.

Les données génétiques sont des **données très fortement corrélées**. La caractérisation des différents motifs de DL le long du génome, c'est-à-dire des différentes structures, ou blocs de SNPs plus ou moins corrélés entre eux qui partagent donc, au moins partiellement, la même information, est nécessaire. **La prise en compte du DL** a donc une importance cruciale dans un GWAS, que ce soit pour la sélection des SNPs nécessaires pour effectuer le GWAS, le gain d'une meilleure couverture du génome (imputations et recherche des variants causaux), le calcul du nombre de tests effectivement indépendants effectués pour la correction pour les tests multiples, ou encore pour l'interprétation des résultats. Pour une région chromosomique donnée avec une structure de DL connue, un panel de SNPs réduit peut capturer la majeur partie de l'information de l'ensemble de la structure génétique de la région [9,10] ces SNPs sont alors appelés **tag-SNPs** (**Figure 1.2**). Cela a pour conséquence de permettre de tester l'association entre un SNP et la maladie de manière indirecte à partir de l'observation d'un panel restreint de SNPs génotypés. C'est d'ailleurs le cas le plus souvent, où les SNPs trouvés associés au phénotype étudié sont le plus souvent des tag-SNPs en DL avec des variants fonctionnels non génotypés et/ou sont portés sur le même haplotype que des variants plus rares. Des études de simulation ont montré que les signaux d'association pouvaient être représentatifs d'un variant fonctionnel rare situé jusqu'à 9 Mb des SNPs rapportés par les GWAS [11]. Les structures de DL s'étendent parfois sur de longues distances (plusieurs centaines de kilobases), incluant de nombreux gènes candidats, ce qui rend parfois difficile l'interprétation biologique des résultats. Des études de **cartographie fines** (ou *fine-mapping*) consistant à génotyper et/ou imputer précisément la région chromosomique détectée sont nécessaires afin d'identifier les gènes impliqués et de caractériser les variants potentiellement

fonctionnels. Certains **projets internationaux** comme le projet HapMap [9], le projet 1000 Genomes [12], ou plus récemment le consortium HRC (the Haplotype Reference Consortium) [13] ont permis de caractériser les variants génétiques humains et leurs structures de DL le long du génome pour de nombreuses populations. Aujourd'hui, des catalogues complets de SNPs sont mis à disposition du public sans restriction dans des bases de données en ligne. Ces **panels de références** peuvent permettre notamment d'enrichir les données génotypiques issues des échantillons d'études génotypés à l'aide de puces de génotypages (dont les deux principaux fabricants sont Illumina et Affymetrix), grâce à des méthodes **d'imputation de données** basées sur des méthodes de Monte Carlo par chaînes de Markov (MCMC). Ces méthodes d'imputations permettent d'inférer les SNPs dont l'information est manquante à partir des haplotypes d'une population de référence génotypée ou séquencée sur une puce plus dense (HapMap, 1000 Genomes, HRC,…). Différents programmes d'imputation ont été proposés et comparés : MACH, BEAGLE, fastPHASE, IMPUTE2, SHAPEIT [14,15]. Les imputations de génotypes sont notamment très utiles pour uniformiser le pool d'information génétique contenue dans différents échantillons d'études génotypés avec des puces différentes, et ainsi permettre de combiner les résultats de plusieurs études pangénomiques par **méta-analyse**.

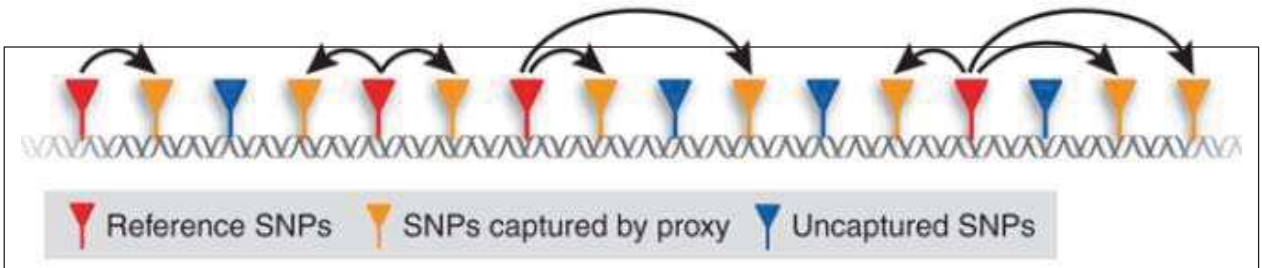


Figure 1.2: Relation entre SNPs génotypés (rouge), capturés (orange) et non capturés (bleu) par la puce de génotypage dans une région du génome (extrait de Kruglyak *et al* [16]). Les SNPs représentés en rouge sont génotypés et capturent l'information des SNPs représentés en orange avec lesquels ils sont respectivement en fort DL. Les SNPs indiqués en bleu ne sont ni génotypés, ni corrélés avec les SNPs génotypés, et de ce fait, si l'un d'eux était associé au trait étudié, alors son association au phénotype étudié pourrait ne pas être mise en évidence.

La prise en compte du DL est également un moyen de mieux corriger le seuil de significativité des tests statistiques effectués lors d'un GWAS. En effet, dû à la multiplicité des tests effectués lors d'un GWAS, il est nécessaire de fixer un seuil de signification strict, dit **seuil de significativité**

pangénomique, afin de bien corriger le taux d'erreur de type I. Ce seuil est le plus souvent estimé par la méthode de correction de Bonferroni qui consiste à diviser le seuil nominal de 5 % par le nombre de SNPs testés. Cette méthode peut s'avérer statistiquement trop conservatrice, étant donné que les tests effectués ne sont pas indépendant les uns des autres dû à la forte corrélation des SNPs. La connaissance de la structure de DL des données peut ainsi permettre de mieux estimer le seuil de significativité et de mieux contrôler le taux d'erreur de type II. Des méthodes ont été proposées pour évaluer le nombre effectivement indépendant de SNPs testés en exploitant les composantes principales de la matrice de DL [17]. Toutefois, un niveau de signification de $P = 5\times10^{-8}$ est le seuil actuellement admis pour déclarer comme significatif l'effet d'un SNP [18] dans une étude pangénomique suivant un taux de faux positifs de 5%.

Préalablement à l'interprétation des résultats d'analyses pangénomiques, le taux de faux positifs est de nouveau contrôlé à l'aide d'un **diagramme Quantile-Quantile** (ou *Q-Q plot*) qui compare la distribution observée des statistiques de tests des effets de SNPs à la distribution théorique sous l'hypothèse nulle d'absence d'association avec le phénotype (chi2 à un degré de liberté pour un modèle génétique additif). Le paramètre d'inflation génomique λ (rapport des médianes de ces deux distributions) est égal à 1 en l'absence de faux positifs. S'il s'écarte de 1 de manière trop importante, cela peut indiquer que les résultats ont pu être biaisés, possiblement à cause d'une stratification de population résiduelle. Dans ce cas, il a été proposé de corriger les statistiques des tests de SNPs pour ce paramètre d'inflation génomique [19].

Depuis la première étude pangénomique conduite en 2005 [8], plus de 2500 GWAS incluant au total plusieurs centaines de milliers d'individus ont été conduits pour plus de 280 maladies ou traits phénotypiques. Plus de 24000 associations maladie-marqueur ont été trouvées (**Figure 1.3**). Ces résultats sont regroupés dans un catalogue en ligne appelé GWAS Catalog (https://www.ebi.ac.uk/gwas/) [20,21]. Les GWASs ont permis notamment de découvrir de nouvelles régions du génome potentiellement impliquées, et de générer de nouvelles hypothèses biologiques sur l'étiologie de ces traits. Un des apports importants des GWAS a été la mise en évidence de variants génétiques (ou des régions du génome) influençant plusieurs traits (variants à effet pléiotrope). Environ 80 % des SNPs atteignant le niveau de signification pangénomique sont localisés dans des régions du génome non codantes [22,23].

Bien que ces études aient conduit avec succès à l'identification de nombreuses régions du génome associées à des traits complexes, ces résultats n'expliquent qu'une part de la composante génétique de ces maladies. Ceci peut être dû à différents facteurs discutés par Manolio *et al.* [24] et Eichler *et al.* [25]. Ces facteurs peuvent être dûs à de l'information manquante dans les échantillons d'études dû à des limites techniques, tel qu'une sous-estimation des effets des SNPs identifiés en raison du DL incomplet avec les variants causaux et/ou des variants rares (fréquence < 0,01) ou de variations structurales (CNVs, *copy number variants*). Il peut également s'agir des effets conjoints de multiples SNPs qui influencent le trait étudié, chacun ayant un effet marginal faible et donc non détectable par les études pangénomiques simple-marqueurs. Ces SNPs peuvent être détectés au moyen de méthodes appropriées qui rendent mieux compte de l'architecture génétique complexe des maladies multifactorielles, telles que les **méthodes multi-marqueurs**. Enfin, ceci peut également être expliqué par l'existence de mécanismes complexes non pris en compte, tels que des **interactions gène-environnement** ou des **interactions gène-gène** [26].

Figure 1.3 : Résultats issus de l'ensemble des GWASs publiés, représentés selon leur localisation chromosomique, et 17 catégories de maladies (code couleur). Figure récupérée sur le site du GWAS Catalog (https://www.ebi.ac.uk/gwas/diagram).

Encart 2 : Méthodes d'ajustement pour les tests multiples

La problématique des tests multiples est un point clé des études d'association à grande échelle [231]. Le nombre très important de tests réalisés sur plusieurs milliers de marqueurs implique des corrections extrêmement rigoureuses des niveaux de significativité pour contrôler l'erreur de type I (faux positifs). En effet, pour un test statistique d'association classique, on désire tester l'hypothèse d'absence d'association représentée par $H_0 : \beta = 0$ (effet nul du SNP sur la maladie étudiée). Pour chacun des tests, on considère un taux d'erreur de type I, communément fixé à 5%. Cependant, si ce taux est valable indépendamment pour chaque test, il n'est plus contrôlé globalement. Ainsi, pour deux hypothèses testées dans un même échantillon, pour deux **SNPs indépendants,** si on a individuellement 5% de chance de rejeter l'absence d'effet du SNP sur la maladie (taux marginal) à tort, globalement le risque de rejeter à tort au moins l'une des deux hypothèses est :

$$\alpha_2 = \mathrm{P}(\min(P_{SNP1}, P_{SNP2}) \leq 5\%) = 1 - P(P_{SNP1} > 5\%) \times (P_{SNP2} > 5\%) = 0.0975$$

Le taux global d'erreur de type I est presque doublé (9.75%). Plus on souhaite tester de possibles associations avec le trait étudié, plus il y a inflation de ce taux.

Une solution consiste à contrôler la probabilité d'avoir au moins un faux positif parmi l'ensemble des tests dont la matrice de corrélation est inconnue, que l'on appelle le **FWER** (*Family Wise Error Rate*). Si on considère les valeurs de p_i, $i = 1,\ldots, N$, les *P*-valeurs individuelles issues des tests de *N* SNPs, la correction de Bonferroni est sans doute la méthode la plus commune : une hypothèse est rejetée lorsque $p_i \leq \frac{\alpha}{N}$ avec α, le taux marginal d'erreur de type I. Bien que cette correction soit la plus couramment utilisée, elle est très conservatrice lorsque *N* est grand, et entraîne une perte de puissance importante. Des procédures pas à pas ont été proposées. Par exemple, dans la procédure de Hochberg [232], les valeurs p_i sont classés par ordre croissant tel que $p_{ij} \geq p_{ij+1}, j = (1,\ldots,N)$, et toutes les hypothèses telles que $p_{ij} \leq \frac{\alpha}{N-j+1}$ sont rejetées. Cette correction reste cependant également assez conservatrice.

Un autre type de correction a été proposé : le FDR pour *False Discovery Rate*. Il consiste à contrôler la proportion attendue d'hypothèses nulles rejetées à tort, c'est-à-dire un taux de faux positifs, en moyenne. Cette correction est moins conservatrice : un FDR = 5% permet à 5% des tests positifs reportés d'être des faux positifs. Ce type de correction a été pour la première fois introduit en 1995 par Benjamini et Hochberg [233] qui ont proposé de trier les *P*valeurs dans l'ordre croissant, puis de les ajuster comme ci-dessous :

$$p_i{}^{FDR} = \frac{N}{rang(p_i)} \times p_i$$

Il est à noter que ces méthodes considèrent que les tests effectués sont indépendants. Hors, les données génétiques sont corrélées le long du génome. Ces méthodes de correction pour les tests multiples réduisent d'autant plus la puissance de détection des analyses pangénomiques qu'elles ne prennent pas en compte la dépendance des tests. Certaines méthodes ont été développées pour prendre en compte les corrélations entre marqueurs dues au déséquilibre de liaison pour estimer un nombre de tests effectifs indépendants [17]. La méthode du **Meff** et ses extensions estiment un nombre de tests indépendants à partir des valeurs propres d'une matrice de corrélations entre SNPs. Cette matrice peut par exemple être la matrice de DL des SNPs. Une autre méthode propose le calcul du Meff suivant [234] :

$$Meff = M - \sum_{i=1}^{M} [I(\lambda_i)(\lambda_i - 1)]$$

Où les λ_i sont les valeurs propres d'une matrice de corrélations de statistiques de tests de M SNPs. La partie sous la somme estime le nombre de tests redondants résultant de la dépendance au SNP i. Ces méthodes ont notamment été étendues à l'analyse d'interactions gène-gène, en considérant le nombre de tests d'interactions indépendants comme le Meff appliqué au produit des paires de SNPs testées [87].

L'estimation de niveaux de significativité empirique, par simulation ou permutation de phénotypes (i.e. du phénotype entre individus) ou de génotypes (pour les méthodes multi-marqueurs) peut également être utilisée, bien que cela puisse s'avérer très couteux en temps de calcul.

3. Analyses pangénomiques prenant en compte des mécanismes complexes

Malgré de très nombreuses études, les GWASs qui considèrent uniquement les effets marginaux observés de chaque marqueur sur la maladie n'ont permis d'expliquer qu'une faible part de la composante génétique de la maladie. Les raisons qui peuvent expliquer les limites des découvertes venant des GWASs sont nombreuses, et notamment le problème de **grande dimensionnalité** qui implique une forte correction pour les tests multiples.

Des méthodes statistiques plus élaborées ont été proposées pour l**'analyse conjointe des effets de plusieurs SNPs** sur la maladie afin d'identifier de nouveaux variants génétiques à effets marginaux faibles. Les **méthodes multi-marqueurs** peuvent s'appliquer soit à l'ensemble du génome de manière agnostique soit intégrer des connaissances biologiques contenues dans des bases de données ou issues de la littérature par **fouille de texte** pour caractériser et/ou présélectionner des ensembles de gènes et des interactions entre gènes impliqués dans la maladie, et ainsi réduire le problème lié aux tests multiples. Les méthodes d'**interaction gène-gène** recherchent des relations complexes qui existent entre les loci de susceptibilité et par extension, entre les marqueurs inclus dans le jeu de données. Ces méthodes permettent d'identifier des interactions statistiques entre régions génétiques ayant des mécanismes biologiques communs, similaires ou complémentaires, et non plus des marqueurs pris individuellement. Les méthodes d'interactions gènes-environnement permettent d'identifier des déterminants génétiques qui confèrent un risque différent selon la présence ou l'absence d'exposition à un facteur environnemental.

Dans la partie qui va suivre, nous distinguerons les méthodes qui **recherchent des ensembles de SNPs** associés au trait étudié, de celles qui s'intéressent à des **interactions statistiques entre SNPs ou avec des facteurs environnementaux**.

3.1 Méthodes d'analyse multi-marqueurs

Les maladies complexes peuvent mettre en jeu un grand nombre de loci de susceptibilité avec des effets faibles ou modérés, plutôt qu'un seul locus ayant un effet majeur sur la maladie. Les méthodes multi-marqueurs permettent de tester conjointement plusieurs SNPs pour leur association avec la maladie. Ces différentes méthodes ont des contributions marginales parfois assez différentes et définir une classification de ces approches est un problème multidimensionnel. En premier lieu, nous classerons les méthodes multi-marqueurs en deux grandes catégories : 1) les méthodes qui peuvent s'appliquer soit à **l'ensemble du génome de manière agnostique**, et 2) celle qui permettent d'**intégrer des connaissances biologiques** contenues dans des bases de données (regroupant les SNPs par gènes ou ensembles de gènes) ou issues de la littérature par fouille de texte pour caractériser des ensembles de gènes et/ou des interactions entre gènes potentiellement impliqués dans la maladie. Ces méthodes peuvent également varier selon le type de données utilisé : elles peuvent s'appliquer aux **données individuelles** (génotypiques, phénotypiques et environnementales) ou aux **statistiques de tests issues des résultats d'études pangénomiques** simple-marqueur, parfois appelées *summary statistics*. Enfin, ces approches peuvent également utiliser **différentes méthodes statistiques** (méthodes de régression, méthodes combinant les statistiques de tests simples marqueurs, méthodes s'appuyant sur des connaissances biologiques, etc), et peuvent varier selon le **type de données** pour lesquelles elles ont été conçues (données cas-témoins ou familiales et la nature du trait étudié : binaire, catégoriel, quantitatif, survie) et/ou la possibilité (ou non) de prendre en compte des **cofacteurs** (âge, sexe, facteurs environnementaux, composantes principales des données génotypiques…).

3.1.1 Méthodes agnostiques

Les méthodes agnostiques regroupent des méthodes de régressions, de réduction de la dimensionnalité, ou encore des méthodes exploratoires. Ma thèse est focalisée sur les méthodes de régression. Dans ce cadre, je ne vais présenter que des méthodes d'apprentissage supervisé.

Les méthodes de **régression multiple** recherchent les effets conjoints de marqueurs sur le trait étudié en explorant l'ensemble du génome ou des régions spécifiques. Elles permettent d'estimer les effets des SNPs sur le trait étudié en tenant compte de cofacteurs (âge, sexe, composantes principales des données génotypiques…). Pour l'analyse de régions spécifiques du génome, comme l'analyse fine de régions identifiées par des études pangénomiques, les modèles de

régression multiple permettent d'estimer les effets de plusieurs SNPs (coefficients de régression). Les estimations des paramètres sont le plus souvent basées sur la théorie du maximum de vraisemblance et les effets des SNPs peuvent être testés à partir du test du rapport de vraisemblance, du test Wald ou du test du score qui sont tous équivalents asymptotiquement. Le nombre de marqueurs considérés dans les modèles peut être fixé à l'avance, ou l'on peut sélectionner les combinaisons de marqueurs les plus associés au trait étudié avec des méthodes de **régressions pas à pas** avec une procédure de sélection des SNPs ascendante (*forward*) ou descendante (*backward*), et/ou basée sur des critères d'informations (comme les critères d'information d'Akaike ou de Schwartz).

Cependant, les données génétiques impliquent la plupart du temps beaucoup plus de variables (SNPs et cofacteurs) que d'individus analysés, pouvant conduire à des problèmes de convergence des estimateurs, à des problèmes de sur-apprentissage (*overfitting*) ou encore des problèmes de colinéarité car les données génétiques sont fortement corrélées. Des méthodes de **régressions pénalisées** permettent de réduire le nombre de variables dans le modèle en sélectionnant les variables les plus pertinentes : une contrainte sur la somme des normes des coefficients dans la vraisemblance du modèle (paramètre de pénalisation ou *shrinkage*) impose la réduction à zéro des estimateurs des effets des SNPs les moins discriminants. En conséquence, l'estimation et la sélection des SNPs sont effectuées de façon simultanée. Il existe différents types de pénalisations suivant la norme utilisée : la régression **ridge** [27] utilise la norme L2 (somme des effets quadratiques des SNPs), la régression **LASSO** [28] correspond à la norme L1 (somme des valeurs absolues des effets des SNPs). Chaque méthode a ses avantages : la régression ridge aboutit à de meilleures prédictions que la régression LASSO notamment en présence de SNPs fortement corrélés [29], mais ne réduit pas les coefficients strictement à zéro, résultant en des difficultés d'interprétation du modèle. La **régression elasticnet** [30] propose un compromis entre les régressions lasso et ridge en combinant les deux critères. D'autres méthodes de type LASSO ont été développées, notamment pour corriger le problème des variables corrélées [31,32]. Notamment, la méthode **HyperLasso** [33] qui utilise un modèle bayésien, a montré de meilleures performances par rapport à d'autres méthodes de régression pénalisée dans des données simulées [34]. D'autres méthodes de type **régression PLS** (pour *Partial Least Squares*) [35] permettent d'estimer le phénotype à partir d'une dimension réduite des données génotypique. Là où les méthodes de régression classiques cherchent à réduire une erreur entre phénotypes observés et estimés, la régression PLS cherche à maximiser la covariance

entre le phénotype et des combinaisons linéaires des prédicteurs (principe similaire à celui de l'ACP, mais pour de l'apprentissage supervisé). Les axes représentés par des sous-groupes de variables peuvent donc être sélectionnés pour estimer le phénotype. Des extensions de la méthode incluent des pénalisations de type LASSO [36,37] et des regroupements de variables (SNPs d'un même gène) [38].

D'autres méthodes d'apprentissage supervisé existent, comme par exemple les méthodes de **forêts d'arbres décisionnels** (*random forest*s) [39] définissant des arbres de décision selon les valeurs des données génotypiques afin de classer et prédire le phénotype des sujets.

3.1.2 Méthodes basées sur des connaissances biologiques

Au cours de cette dernière décennie, un grand nombre de méthodes d'analyses multi-marqueurs basées sur des connaissances biologiques ont été proposées. Ces méthodes peuvent varier selon de nombreux critères comme notamment la manière d'incorporer l'information aux données (utilisée pour le regroupement des variables ou pour sélectionner des variables ou groupes de variables à partir d'information textuelle), les méthodes statistiques utilisées, ou encore le type d'information prise en compte. Notamment, dans le cas des méthodes qui intègrent des connaissances afin de regrouper les variables pour la construction d'un test, le type de méthodes peut être déterminé par l'entité biologique exploitée : les **approches basées sur les gènes**, et les approches basées sur des ensembles de gènes. Parmi ces dernières, on distingue celles contribuant à des voies biologiques connues et prédéfinies dites **méthodes de *pathways***, de celles basées sur des réseaux de gènes déterminés selon différentes sources et construits de manière adaptative, dites **méthodes de réseaux ou *Networks***.

Méthodes basées sur l'analyse au niveau des gènes

Ces méthodes ont pour objectif d'identifier des gènes associés au phénotype étudié. Elles consistent tout d'abord à regrouper les SNPs au niveau des gènes en se servant des données de positions des SNPs et des gènes sur le génome, issues de bases de données (notamment dbSNP[1] pour les SNPs, et RefSeq[2] pour les gènes). Afin de considérer des SNPs proches de gènes, en DL avec des SNPs du gène qui n'auraient pas été génotypés (ou imputés) ou de prendre en compte des SNPs pouvant

[1] http://www.ncbi.nlm.nih.gov/SNP/

[2] http://www.ncbi.nlm.nih.gov/refseq/

réguler la transcription du gène (au sein du promoteur), une fenêtre basée sur une distance de 10 à 500 kb [40] ou plus récemment sur les structures de DL [41], de part et d'autre des bornes du gène, peut être considérée. En regroupant les SNPs (plusieurs millions) au niveau des gènes (~23 000 gènes), ces méthodes réduisent le problème des tests multiples (seuils corrigés variant de $\sim 5\times10^{-8}$ au niveau des SNPs à $2{,}3\times10^{-6}$ au niveau des gènes). Ces méthodes doivent tenir compte du DL entre SNPs et de la taille variable des gènes. Elles se distinguent par le type de données utilisées (résultats de tests issus d'études pangénomiques simple-marqueur ou données génotypiques individuelles), le nombre de SNPs considérés dans le gène (tous les SNPs ou un sous-ensemble), la méthode statistique utilisée et l'évaluation de la signification statistique.

Les méthodes d'accumulation testent si une combinaison de SNPs dans un gène est associée au trait étudié. Les statistiques de tests sont souvent basées sur les produits [42–45] ou les sommes des statistiques de tests simple marqueur [40]. Certaines méthodes combinent les *P*valeurs de tous les SNPs assignés au gène, comme la méthode de Fisher [42]. D'autres approches permettent de considérer au sein du gène un sous-ensemble des SNPs les plus associés au trait étudié, dont le nombre peut être fixé arbitrairement [43,44] ou varier de manière adaptative pour chaque gène au cours de l'analyse comme pour la méthode **ARTP** [45]. Cette méthode utilise des **permutations de phénotypes** pour évaluer empiriquement la significativité des associations gène-phénotype, nécessitant des temps de calculs importants et l'accès aux données individuelles. D'autres méthodes basées sur des simulations de distributions de statistiques de tests, permettent de s'affranchir de ces contraintes, comme **VEGAS** [40], puis **VEGAS2** [41]. Ces méthodes somment les statistiques de tests d'association d'une fraction déterminée des SNPs du gène. La significativité de la statistique du test d'association au niveau du gène est déterminée empiriquement. La distribution de la statistique sous l'hypothèse nulle est générée à partir de simulations basées sur l'approche de Monte Carlo, en utilisant la matrice de DL entre SNPs (estimée dans une population de référence). La méthode **MAGMA** [46] qui nécessite les données individuelles, permet de construire une statistique similaire basée sur une régression multiple. La méthode **CGP** [47] utilise des permutations circulaires de SNPs afin de conserver la structure de DL pour le calcul de la significativité. Cette méthode a été étendue pour tester de manière analytique toutes les combinaisons possible de permutations circulaires : **FastCGP** [48]. La méthode **SKAT** (*Sequence Kernel Association Test*, [49,50]) basée sur un test de décomposition de la variance sous un modèle mixte (effets des SNPs aléatoires), permet de tester **l'effet global de tous les SNPs** assignés au

gène sur le trait étudié en tenant compte des directions des effets des SNPs parfois opposées. Cette méthode a été étendue à l'analyse de données familiales [51–53].

Méthodes basées sur l'analyse au niveau de Pathways biologiques

Les méthodes de type Pathways utilisent des connaissances issues de bases de données d'annotations pour regrouper des gènes qui contribuent à des voies biologiques communes, et tester l'association de chacun de ces groupes avec le trait étudié en utilisant des résultats issus d'études pangénomiques. Les **groupes de gènes** sont entièrement **prédéfinis et fixes**, mais ne sont pas indépendants : différents pathways peuvent avoir certains gènes en communs. Au moins 300 bases de données biologiques pour définir les pathways sont recensées dans le catalogue Pathguide [54]. Ces bases diffèrent principalement selon le type de connaissances utilisées pour définir les pathways, la couverture biologique et la qualité des annotations des gènes aux pathways (inférées électroniquement et/ou vérifiées manuellement par un expert) [55]. Ces approches peuvent être utilisées pour tester si des pathways candidats (sélectionnés ou définis sur la base d'hypothèses a priori) sont associés au trait étudié ou pour tester de manière exhaustive tous les pathways issus d'une base de données. Les méthodes de pathways peuvent être classées en deux catégories selon le type de construction des hypothèses nulles testées. Ces hypothèses diffèrent selon que l'on recherche les pathways les plus enrichis (ou surreprésentés) en gènes associés au trait étudié en comparant les pathways les uns par rapport aux autres (**hypothèse compétitive**), ou que l'on recherche les pathways associés au trait étudié indépendamment des autres pathways examinés (hypothèse autonome des pathways, dite ***self-contained***). Parmi les méthodes compétitives, on peut distinguer deux sous-types de méthodes : les méthodes de surreprésentation, et d'enrichissement. Les méthodes **de surreprésentation** qui recherchent les pathways qui sont surreprésentés en gènes (ou SNPs) associés au trait à partir d'une liste de gènes (ou SNPs) significatifs. Certaines de ces méthodes sont implémentées dans des programmes utilisables localement, comme les méthodes **ALIGATOR** [56] et **MAGENTA** [57], d'autres en revanche ne sont accessible qu'à travers une interface Web, comme la méthode **DAVID** [58]. Les **méthodes d'enrichissement** sont quant à elles basées sur les rangs des statistiques de gènes (ou SNPs) et ne sont pas limitées à ceux qui atteignent un seuil fixé comme pour les méthodes de surreprésentation. Les gènes (ou SNPs) sont ordonnés selon leur statistiques de tests, et analysés sur l'ensemble du génome selon leur degré d'association au trait étudié. Les rangs des gènes (ou SNPs) d'un pathway donné sont alors comparés aux rangs des autres gènes (ou SNPs) du génome analysés. Un exemple de méthode d'enrichissement est la

méthode **GSEA** [59], qui parcourt pour chaque pathway la liste de gènes ordonnés du plus associé au moins associé en augmentant ou diminuant une somme récursive basée sur une statistique similaire à celle de Kolmogorov-Smirnov. Les **méthodes *self-contained*** testent si un pathway donné est associé au trait étudié indépendamment des autres pathways étudiés. Elles peuvent se distinguer en deux sous-catégories selon qu'elles testent si un pathway comprend une accumulation de gènes (ou SNPs) marginalement associés au trait étudié ou qu'elles évaluent un effet global de tous les gènes (ou SNPs) assignés au pathway testé. Certaines méthodes de pathways étendent des méthodes de statistiques proposées pour les gènes aux pathways : **ARTP** [44], **VEGAS2Pathway** [60] et **MAGMA** [46], ou encore **PASCAL** [61]. Récemment, la méthode **TAD Pathway** [62] permet de tester des groupes fonctionnels de gènes définis dans des domaines topologiquement associés (*Topologically Associated Domains*), qui relient la structure du génome à sa fonction en prenant en compte les mécanismes de régulation des gènes.

Méthodes de Networks

Il existe d'autres méthodes basées sur des réseaux d'interaction entre gènes. Ces approches se concentrent sur les relations entre composants moléculaires ou protéines codées par les gènes et représentent les relations entre molécules par des liens entre gènes pour former un réseau [63]. Ces méthodes utilisent des réseaux de gènes issus de bases de données d'annotation en ligne et peuvent partager des étapes communes aux méthodes de pathways. Les méthodes de « Networks » associent des statistiques issues d'un GWAS à ces réseaux de gènes afin de **reconstituer des groupes de gènes de manière adaptative** (considérer ou non certains liens entre gènes), ou sous-réseaux, qui maximisent les associations réseau-maladie. Les modules identifiés incluant des gènes liés entre eux directement ou indirectement (par l'intermédiaire d'autres gènes) permettent de générer de **nouvelles hypothèses biologiques**. De nouvelles méthodes ont récemment été développées dans le laboratoire où j'ai effectué ma thèse, et appliquées à l'asthme [48,64].

En assignant les SNPs (plusieurs millions) aux gènes et les gènes en groupes de gènes, les méthodes de networks aussi bien que celles de pathways permettent de **réduire le problème des tests multiples** par rapport aux SNPs et peuvent faciliter l'interprétation biologique des résultats.

3.2 Méthodes d'analyse d'interactions gène-gène

Les méthodes d'interaction gène-gène (GxG) recherchent si l'effet joint de deux ou plusieurs loci sur le génome diffère de l'effet prédit par chacun des locus pris individuellement. La plupart des méthodes utilisent des tests basés sur les SNPs, et recherchent des interactions entre paires ou un nombre supérieur de SNPs dans les données pangénomiques, soit par une recherche exhaustive de toutes les combinaisons de SNP, soit par le test d'un ensemble présélectionné réduit. Cette présélection peut être d'origine statistique, ou faire appel à des connaissances *a priori* extérieures au jeu de données.

3.2.1 Méthodes agnostiques

Il existe différents types de méthodologies qui permettent de rechercher des interactions gène-gène de manière agnostique. Ces méthodes se placent au niveau des SNPs et recherchent des interactions pour chaque paire de SNPs soit en les testant de façon exhaustive au niveau du génome ou au sein d'un sous-ensemble de SNPs présélectionnés à partir de critères statistiques. Il existe différents types d'approches, fréquentistes et/ou bayésiennes [65]. Dans le cadre de ma thèse, je me suis focalisé sur les approches fréquentistes, et en particulier sur les méthodes basées sur des modèles de régression, qui sont les plus couramment utilisées. Les approches basées sur la régression ont l'avantage d'utiliser des modèles facilement interprétables où l'effet de l'interaction gène-gène est identifiable par un ou plusieurs paramètres qui relient les génotypes au phénotype. Ces modèles permettent d'introduire des termes d'interactions. L'interaction est testée selon un test du rapport de vraisemblance qui compare les vraisemblances des modèles avec les termes d'interactions (modèle saturé) à un modèle restreint (sans terme d'interaction). Ils permettent d'explorer différents modèles d'interactions, avec des effets additifs (test à 1 degré de liberté) ou avec des combinaisons d'effets additifs et dominants (test à 4 degrés de liberté) dans le cadre d'un modèle général.

Pour une recherche sur l'ensemble du génome, l'objectif principal de ces méthodes est d'identifier des interactions entre paires de SNPs parmi l'ensemble des paires possibles le long du génome, dont le nombre croit exponentiellement avec le nombre de SNPs (n SNPs testés impliquent n (n – 1)/2 interactions possibles). Tester l'ensemble des interactions possibles implique une énorme **charge de calcul**, et une perte de puissance extrêmement importante dû à la **correction pour le nombre de tests effectués**. Plusieurs algorithmes ont été proposés pour réduire les temps de calcul

d'une recherche exhaustive sur le génome. D'une part, ces algorithmes tirent parti des nouvelles technologies informatiques (clusters informatiques et/ou processeurs graphiques GPU (Graphics Processing Unit) pour les parallélisassions des calculs. D'autre part, ces algorithmes s'appuient sur des optimisations numériques avec des approches en deux étapes, incluant une sélection de paires de SNPs basée sur une approximation des tests, suivie de tests formels d'interaction permettant d'accroître la vitesse de calcul. La méthode EPIBLASTER [66], implémentée pour utiliser des processeurs graphiques afin de réduire les temps de calcul, sélectionne les paires les plus prometteuses sur la base de comparaison de corrélation entre SNPs, entre cas et témoins. Les interactions entre paires de SNPs sont ensuite testées avec un test classique d'interaction. La méthode BOOST [67] est un exemple de méthode développée pour l'analyse de trait qualitatif [68], basée sur l'approximation de superposition de Kirkwood du test du rapport de vraisemblance. Bien que les progrès dans l'optimisation informatique et numérique rendent possible les analyses d'interaction à l'échelle du génome avec des temps de calcul raisonnables [66,69] le problème des tests multiples reste une limitation majeure qui nécessite des effectifs importants pour atteindre une puissance suffisante [70].

Afin de restreindre le nombre de tests effectués, d'autres approches ont été proposées afin de rechercher des interactions au sein d'un sous-ensemble de SNPs sélectionnés à partir de critères statistiques, ou d'algorithmes efficaces. Les approches basées sur des filtres statistiques sont guidées par les données [71] et dépendent le plus souvent de critères de sélection pour conserver chacun des SNPs les plus informatifs à tester en interaction. Marchini *et al.*[72] ont proposé de tester en interactions seulement un sous ensemble de SNPs montrant des effets marginaux au test simple-marqueur (*P*-valeur sous un certain seuil). D'autres méthodes sélectionnent les SNPs sur des critères d'hétérogénéité des variances phénotypiques entre les différents génotypes du SNP [73,74], sur des approches basées sur une matrice d'information mutuelle [75] ou utilisant d'autres fonctions [76]. Cependant, il a été montré que la méthode basée sur l'hétérogénéité des variances phénotypiques pouvait manquer de puissance pour les traits autres que quantitatifs [77].

Des méthodes d'apprentissage automatique ont également été développées [78–84]. Ces méthodes permettent de réduire la dimension des données et/ou de sélectionner des SNPs afin de réduire le temps de calcul d'une recherche exhaustive. Elles ont l'avantage sur les modèles de régression de ne pas être dépendantes d'un modèle définit *a priori* et de permettre de détecter des interactions non linéaires et/ou entre un très grand nombre de SNPs. Cependant, ces méthodes ne fournissent

pas d'estimation des effets des interactions sur le trait étudié, et l'interprétation biologique des résultats est d'autant plus difficile que le nombre de combinaisons de SNPs identifiées est important.

3.2.2 Méthodes basées sur des connaissances extérieures

D'autres approches permettent de réduire le nombre d'interactions testées en utilisant des connaissances extérieures stockées dans des bases de données en ligne. Ces méthodes utilisent des connaissances scientifiques soit pour regrouper les SNPs dans des gènes ou des groupes de gènes, soit comme filtre pour restreindre le nombre d'interactions GxG testées à des sous-ensembles spécifiques qui ont plus de sens [85].

Certaines approches proposent de réduire le nombre de tests en regroupant les paires de SNPs testées au sein d'un gène ou de plusieurs paires de gènes. Ma *et al*. [86] ont étendu plusieurs méthodes basées sur l'analyse au niveau des gènes pour l'analyse d'interactions entre paires de gènes dans le cadre d'un trait quantitatif. Ces méthodes testent si chaque paire de gènes comprend une accumulation d'interactions entre paires de SNPs, en combinant les *P*-valeurs issues des tests d'interaction entre paire de SNPs (SNPxSNP). Comme pour les méthodes basées sur les gènes, la signification au niveau de chaque paire de gènes est évaluée à partir de simulations de distributions normales multivariées, en prenant en compte la dépendance entre paires de SNPs et le nombre variable de paires de SNPs testées par l'intermédiaire de la matrice de corrélation entre statistiques de tests d'interactions. Ces méthodes ont également l'avantage de permettre de répliquer au niveau de la paire de gènes plutôt qu'au niveau d'une paire de SNPs [86].

Une approche alternative au fait d'effectuer des recherches exhaustives, consiste à surmonter le problème de l'imposition d'un seuil de signification très strict en limitant la recherche à quelques loci candidats choisis parce qu'ils sont soupçonnés avoir un rôle biologique. Des stratégies en deux étapes ont été proposées pour réduire le nombre de tests en recherchant des interactions gène-gène au sein de groupes de gènes précédemment identifiés grâce à des analyses de pathways (ou de networks) réalisées dans les données. Dans un premier temps, des analyses de pathways biologiques sont conduites dans les données, puis seules les interactions entre toutes les paires de SNPs (ou de gènes) appartenant aux paires de gènes des pathways sélectionnés sont testées [87,88]. Afin de mieux corriger pour les tests multiples, ces méthodes utilisent des extensions de la méthode du Meff pour estimer le nombre de paires de SNPs effectives indépendantes dans un pathway.

D'autres approches permettent également de sélectionner des paires de gènes impliquées dans des pathways candidats ou dans des réseaux d'interactions entre gènes pour tester des interactions à partir de bases de données en ligne. La méthode Biofilter [89] génère des modèles de paires de SNPs à tester en interaction basés sur des interactions biologiques entre gènes et protéines répertoriées dans des bases de données disponibles en lignes : Reactome, KEGG, GO, DIP, Pfam, Ensembl, NetPath.

Méthodes basées sur la fouille de texte

L'information textuelle contenue dans des bases de données biomédicales en ligne est également une vaste source d'information pouvant être utilisée comme processus de sélection pour des interactions gène-gène, et les méthodes de fouille de textes (**Encart 3**) permettant d'extraire ces informations sont de plus en plus attractives pour les chercheurs afin de révéler des informations nouvelles [90,91]. Parmi les méthodes de fouille de textes permettant la **découverte de nouvelles connaissances** (*Knowledge discovery*), je me suis particulièrement intéressé durant ma thèse aux méthodes pouvant être **combinées à des résultats issus des analyses pangénomiques**, pour **sélectionner des gènes ou groupes de gènes** [92] possiblement candidats pour le phénotype étudié, et/ou particulièrement reliés entre eux suivant une métrique textuelle. Le but était de rechercher des liens entre différents gènes identifiés lors de l'analyse pangénomique simple marqueur. Ces méthodes peuvent varier selon différents critères tels que : 1) les bases de données fouillées, 2) le type de données exploitées pour décrire les gènes (mots clés, ensemble des mots associés, structure ontologique,…), 3) le type de similarité (co-citations de gènes ou de mots clés associés, ou cooccurrence de mots), 4) l'utilisation directe, ou indirecte des données issues d'analyses pangénomiques, 5) un corpus de textes agnostique ou non au phénotype d'intérêt. Je vais ici présenter uniquement des méthodes qui travaillent sur des listes de gènes (ou SNPs) pouvant être présélectionnées à partir des résultats d'analyses pangénomiques. On peut distinguer notamment les méthodes qui évaluent des liens entre gènes, et celles qui recherchent plutôt des liens gènes-maladies.

Certaines méthodes se basent sur la co-citation pour représenter des relations entre gènes. Les méthodes de la suite **ToppGene** [93,94] calculent le nombre de fois où deux gènes sont cités ensemble dans un corpus de texte (ex : PubMed, Bibliome) ou qu'ils apparaissent ensemble dans un même pathway biologique [95], codent pour des protéines retrouvées associées dans des réseaux

d'interactions protéine-protéine [96,97], ou partagent des profils issus de données d'expression [98]. Les gènes les plus reliés aux autres peuvent ensuite être sélectionnés. La méthode **G2D** [99–101] sélectionne des gènes candidats à une maladie dans une région en se basant sur des catégories d'ontologies. Les gènes sont annotés aux catégories d'ontologies qui sont ensuite associées aux termes MeSH correspondant au phénotype. Cette méthode peut également s'appuyer sur des interactions protéine-protéine issues de la base de données STRING dans le cas où une liste de référence de gènes associés à la maladie est précisée en entrée. La méthode **SNPs3D** [102] utilise les résumés de PubMed pour déterminer des gènes candidats à partir des nombres bruts de mots clés (noms et d'adjectifs) partagés par des gènes.

D'autres méthodes de fouille de textes se basent sur des vecteurs de mots utilisés pour décrire les gènes dans différentes bases de données textuelles, afin d'évaluer des similarités entre gènes (**Table 1.4**). La méthode **CAESAR** [103] qui requiert un corpus de textes en entrée comportant des noms de gènes ou identifiants OMIM (*Mendelian inheritance in man*) [104], puis une représentation de chaque gène est construite à partir de plusieurs bases de données d'ontologies biomédicales [105–107] suivant une représentation vectorielle de mots. CAESAR évalue ensuite un niveau de similarité entre chaque gène et également avec l'ensemble du corpus grâce à une similarité cosinus (produit scalaire des vecteurs normalisés). Les auteurs préconisent d'utiliser une source contrôlée et complète composée de connaissances biologiques sur le phénotype d'intérêt : dans ce cas la méthode n'est pas agnostique au phénotype étudié. Cette méthode est limitée par le fait de devoir spécifier un corpus sur lequel travailler, bien que cela puisse donner la liberté à l'utilisateur de sélectionner son corpus de textes à partir d'un groupe de gènes s'il a les moyens techniques de le faire. La méthode **GRAIL** [108] s'attache à hiérarchiser des gènes situés dans des régions possiblement associées à une maladie. Afin de comparer les gènes entre eux, la méthode utilise une métrique de cooccurrence de mots appliquée à l'ensemble des résumés issus de la littérature contenue dans PubMed. Ici la représentation de chaque gène se fait dans un espace vectoriel constitué des mots utilisés dans les articles dans lesquels le gène est cité : le poids de chaque mot et chaque résumé est construit selon la méthode TF-IDF (pour *term frequency-inverse document frequency* ; **Encart 4**), puis le poids de chaque résumé est également corrigé de manière inversement proportionnelle au nombre de gènes cités par le celui-ci. Une similarité entre paires de gènes est ensuite calculée à partir de la similarité cosinus. Cette méthode permet de déduire les **gènes les plus similaires ou représentatifs du groupe**. La méthode va ensuite plus loin et permet le calcul d'une *P*valeur d'association textuelle

entre chaque paire de gènes en comparant les rangs des similarités entre les gènes de la paire avec ceux des similarités de chacun des gènes de la paire avec les autres gènes du génome. La méthode GRAIL permet de prendre en compte des biais de sélection et représentation en regroupant les gènes voisins par région. Une *P*valeur d'association textuelle de chaque région avec les autres régions du corpus est calculée. La méthode permet également de fournir une liste de **gènes de références** avec lesquels seront calculées les similarités des autres gènes. Cette méthode n'utilise **aucune information sur le phénotype étudié**, et est donc agnostique à la maladie. Au contraire, une méthode similaire développée par Ailem *et al.* [109] permet de **prendre en compte le contexte phénotypique** en ne sélectionnant dans le corpus que les résumés de PubMed citant à la fois un gène et la maladie. La méthode calcule aussi un **score global de connectivité entre les gènes du groupe,** mais ne permet pas de comparer une liste de gènes avec une liste de référence. La méthode **ENDEAVOUR** [110,111] permet de hiérarchiser une liste de gènes suivant leur similarité avec des gènes de référence basée sur des fonctions géniques et protéiniques, des informations chimiques et des pathways biomoléculaires. Cette liste peut inclure l'ensemble du génome (sans *a priori*). D'autres méthodes recherchent des liens entre gènes et maladies plutôt qu'entre gènes. La méthode développée par Zhou *et al.* [112] permet de prédire des associations gène-maladie à travers la similarité cosinus entre le vecteur du gène et celui de la maladie. Ces vecteurs sont construits à partir des mots des titres, des résumés et des mots-clés MeSH d'indexation des articles de PubMed, et des poids différents sont données aux mots selon leur localisation dans l'article (titres, résumés, mots-clés) et la co-citation ou non avec la maladie.

A côté des méthodes qui travaillent sur des listes de gènes, il existe d'autres méthodes qui intègrent directement des informations contenues dans des bases de données aux résultats d'association pangénomiques pour le calcul d'un score d'association gène-maladie. Notamment, la méthode **AdAPT** [113] utilise les résumés de PubMed pour récupérer des mots clés et utilise cette information pour assigner des probabilités d'associations *a priori* aux SNPs.

Table 1.4. Exemples de méthodes basées sur des vecteurs de mots pouvant être utilisées pour sélectionner des gènes dans les études d'association pangénomiques

Exemple de méthodes	**Catégorie d'utilisation**	**Données d'entrée issues d'association pangénomiques**	**Métrique textuelle**	**Ressources utilisées**
CAESAR [103]	Evaluation d'une similarité entre gènes	liste de gènes	Dépendante d'un corpus de texte spécifié en entrée	OMIM, GO, eVOC, MPO
GRAIL [108]	Hiérarchisation d'une liste de gènes en fonction des similarités entre gènes de la liste	liste de SNPs ou de gènes	Agnostique au phénotype	PubMed, GO
Ailem *et al* [109]	Evaluation de la force d'association entre gènes d'un ensemble de gènes	liste de gènes	Dépendante du phénotype	PubMed
ENDEAVOUR [110,111]	Hiérarchisation d'une liste de gènes en fonction des similarités avec des gènes de référence	liste de gènes	Dépendante du phénotype	GO, SwissProt, Blast CisRegModule
Zhou *et al.* [112]	Evaluation d'un lien gène-maladie	Une maladie et une liste de gènes	Dépendante de la maladie	PubMed, MeSH

Encart 3 : La fouille de textes

La base de données PubMed est la principale source de dépôt pour la **littérature biomédicale** et contient plus de 26 millions d'articles, et le nombre d'articles publiés augmente exponentiellement. Compte-tenu de l'importance croissante de l'information textuelle contenue dans des bases de données biomédicales en ligne, et avec elle de la difficulté croissante de retrouver et identifier cette information, les méthodes de fouille de textes sont de plus en plus attractives pour les chercheurs pour révéler des informations nouvelles [90,91]. La fouille de texte est le fait d'acquérir une information en analysant les corrélations et les structures statistiques à partir de texte non structuré [235]. Le but principal de l'exploration de la littérature biomédicale à l'aide de la fouille de texte est de retrouver et transformer le savoir caché dans la multitude de papiers scientifiques publiés en **information plus structurée**, cohérente et identifiable afin de **mettre en évidence** des relations non suspectées, de permettre et/ou de générer la découverte de nouvelles hypothèses [236]. C'est donc un moyen de **révéler des informations nouvelles**, ou **générer des hypothèses** à partir d'information comprises dans le pool de connaissances déjà publiées, mais indiscernable manuellement (à cause de contrainte de temps et des volumes importants), à l'aide de **processus automatisés**. Notamment, les approches qui **intègrent les résultats des données génomiques** (aussi appelées « *Omics* ») et des informations générées à partir de l'**exploration textuelle** pour découvrir de nouvelles informations biomédicales ont connues d'important progrès ce dernières années [237]. La fouille de texte est un domaine très large avec de vastes possibilités d'applications. Il existe quatre grandes catégories d'étapes pour ces méthodes : 1) les méthodes de récupération d'information (ou IR pour *information retrieval*), 2) la classification de documents (DC pour *Document Classification* ou *Document Prioritization*), 3) l'extraction de l'information (IE pour *information extraction*), et 4) la découverte de nouvelles connaissances (KD pour *Knowledge discovery*) qui peut elle-même intégrer plusieurs aspects différents.

La récupération d'information, se concentre sur la recherche dans d'importantes collections de données pour trouver et extraire des documents qui sont pertinents pour une question, ou requête, donnée [238]. Ils sont évalués sur leur capacité à retrouver tous les documents pertinents (sensibilité), et sur la proportion de documents pertinents dans l'ensemble des documents retrouvés (spécificité). C'est par exemple le cas des moteurs de recherches des explorateurs internet, ou de PubMed (http://www.ncbi.nlm.nih.gov/pubmed).

La classification de documents s'intéresse à trier les documents pour les classer en groupes de documents similaires et/ou les prioriser à partir de différents facteurs. Ces facteurs peuvent être des paramètres liés au journal comme son facteur d'impact (*impact factor*) ou son nombre de citations [239], ou des concepts issus de répertoires structurés de termes servant de références (thésaurus), comme les termes MeSH (*Medical Subject Headings*) dans le domaine biomédical [240,241].

L'extraction d'information a pour but la récupération d'information structurée à l'intérieur des documents, comme l'extraction de concepts ou de relations. Ces méthodes recherchent de manière systématique un nombre important de publications pour extraire des informations spécifiques issus des données textuelles non structurées et les ordonner dans des bases de données [242]. C'est par exemple le cas de la méthode PPInterFinder [243] permettant d'extraire des relations entre protéines et constituer des réseaux d'interaction de protéines pouvant être ensuite utilisées dans des analyses de type Networks.

La découverte de nouvelles connaissances est le dernier objectif des méthodes de fouille de textes et peut découler au moins en partie des trois autres. Ce domaine a pour but d'induire de nouvelles hypothèses de manière automatique en parcourant et traitant les publications déjà existantes.

Encart 4 : La méthode de pondération TF-IDF

La mesure statistique TF-IDF est une méthode de **pondération** permettant de définir le poids d'**un terme** (ou mot) **dans un document**, relativement à l'ensemble des documents étudiés, définissant **le corpus**. Dans un modèle vectoriel ou chaque document est décrit par les mots qui le composent, cette statistique permet de pondérer l'influence d'un terme selon son degré de présence, et de pertinence. Cette statistique peut notamment être très utile pour affiner le calcul d'une similarité entre deux ensembles définis par des mots (par exemple des gènes définis par l'ensemble des mots co-occurrents dans des articles scientifiques). Dans le cas d'un ensemble faisant intervenir plusieurs documents, il est à noter que le calcul de cette statistique ne s'effectue pas globalement sur l'ensemble des termes, mais nécessite des informations sur l'entité document. Cette statistique est définie par **deux éléments**, le TF pour *term frequency* et l'IDF pour *inverse document frequency*.

Le **TF** est la partie relative à l'importance d'un terme dans un document. Il existe plusieurs types de pondérations TF, la plus basique étant la représentation binaire qui représente la présence ou l'absence du terme dans le document. Le nombre d'occurrences du terme dans le document est également une mesure simple et intuitive qui permet en sus d'affecter un poids différents à chaque terme, et donc d'estimer un **poids relatifs de chaque terme** à l'intérieur d'un document. De nombreuses variations dépendantes de la **fréquence** pouvant impliquer différentes **normalisations** existent, comme notamment la fréquence normalisée par la fréquence du terme le plus représenté, ou le changement d'échelle logarithmique, permettant d'affecter un moins grand poids à des termes très communs qui sont assez systématiques, et donc en pratique très peu pertinent.

L'**IDF** est une mesure de l'importance du terme relativement à l'ensemble du corpus, et permet justement de donner plus de poids aux termes qui sont présents dans peu de documents. L'idée de cette mesure est que moins un terme est retrouvé dans différents documents, et plus son sens est **spécifique** : le terme doit donc être statistiquement plus discriminant. L'évaluation de cette quantité consiste donc généralement à calculer **l'inverse de la proportion de documents** du corpus dans lequel le terme est présent. Encore une fois, différentes variantes existent.

La mesure statistique TF-IDF consiste à multiplier les deux quantités TF et IDF pour chaque terme afin d'obtenir le **vecteur des coordonnées** de l'unité d'intérêt (document, gène, etc) dans l'espace vectoriel des termes du corpus. Une **similarité** entre deux éléments peut-être représentée par le produit vectoriel des deux vecteurs normalisés correspondants.

3.3 Méthodes d'analyse d'interactions gène-environnement

Les maladies multifactorielles résultent à la fois de facteurs génétiques et de facteurs environnementaux. Un même polymorphisme génétique peut-être un facteur de risque pour une maladie ou au contraire de protection selon selon l'exposition ou non à un facteur environnemental donné. Dans le cas de la présence d'une interaction gène-environnement (GxE) ayant un effet sur le risque de la maladie, un modèle génétique simple peut ne pas permettre de décerner l'effet génétique sous-jacent. L'utilisation de méthodes permettant **l'analyse des effets et d'interactions statistiques des différents facteurs génétiques et environnementaux** peut conduire à la découverte de nouveaux loci de susceptibilité génétique et ainsi permettre de mieux comprendre les mécanismes à l'origine de ces maladies.

Différentes méthodologies ont été proposées pour caractériser les interactions entre facteurs génétiques et environnementaux dans le cadre d'études pangénomiques. L'interprétation de l'interaction dépend de l'échelle de mesure sous-jacente sur laquelle est modélisé l'effet GxE. Dans le cadre de ma thèse, je me suis intéressé aux méthodes GxE basées sur des modèles de régression utilisant un **modèle multiplicatif de l'interaction**, qui est communément considéré.

Ces approches peuvent être classées selon différents critères : le type de données analysées (données familiales ou non) ainsi que la nature du trait étudié (qualitatif, quantitatif, survie), le type de stratégie utilisée (en une étape, appliquée à l'ensemble du génome ou en deux étapes, incluant une étape de présélection des régions d'intérêt sur le génome suivie par une étape d'analyse des régions sélectionnées), ou encore le niveau de l'objet sur lequel l'analyse se place (au niveau des SNPs, des gènes ou groupes de SNPs, ou de groupes de gènes). Ces méthodes peuvent être réparties en trois catégories : **1) les méthodes en une étape** qui se focalisent sur l'estimation de l'interaction GxE au niveau simple marqueur sur l'ensemble du génome, **2) les méthodes de tests en deux étapes**, qui cherchent à réduire la perte de puissance dû aux tests multiples en ajoutant une étape de sélection / ou filtrage des SNPs qui seront testés en interaction avec un facteur environnemental dans un second temps, **3) les méthodes qui utilisent la connaissance biologique** pour regrouper les SNPs en unités biologiquement pertinentes (en gènes ou groupes de gènes) dont l'interaction avec un facteur environnemental peut être testée conjointement. Nous allons exposer ces méthodes selon ces trois catégories (**Table 1.6**).

3.3.1 Méthodes en une étape

Une interaction GxE est caractérisée par un effet différent du facteur génétique (**G**) sur la maladie (**M**) selon l'état d'un facteur environnemental (**E**) qui peut représenter une variable environnementale exogène (ex : pollution de l'air), un niveau d'exposition d'un individu à un facteur environnemental ou le mode de vie (le fait de fumer ou non) ou une caractéristique endogène (ex : sexe). Ce facteur peut être quantitatif (le nombre de paquets de cigarettes…) ou qualitatif (l'exposition au tabac, passive ou active…).

Dans le cadre de **modèles de régression avec interaction GxE**, la déviation par rapport à un **modèle d'indépendance des effets de G et E sur M** se mesure *via* l'introduction d'un terme d'interaction GxE dans le modèle statistique d'association. Dans les études cas-témoins, pour un phénotype binaire représentant le statut de la maladie (malade ou non), on utilise le modèle de régression logistique qui estime les effets des différents cofacteurs sur la maladie suivant le lien $Logit(P) = \ln(\frac{P}{1-P})$:

$$Logit\ P(M|G,E) = \beta_0 + \beta_E \times E + \beta_G \times G + \beta_{GxE} \times G.E + \beta_C \times C \quad (M1)$$

où β_E, β_G, β_{GxE}, et β_C représentent respectivement les effets estimés du facteur environnemental, du facteur génétique, de l'interaction gène-environnement (GxE), et d'éventuels cofacteurs sur la maladie (qui ne seront plus cités dorénavant). Pour plus de simplicité, les modèles présentés par la suite considèreront un phénotype de maladie (M) et un facteur environnemental (E) binaire. Certaines méthodes peuvent être étendues à l'étude de phénotypes quantitatifs, catégoriels ou modélisés par analyses de survie et à l'étude de facteurs environnementaux quantitatifs. Plus récemment, de nouvelles méthodes d'interactions GxE ont été développées pour prendre en compte des expositions variant dans le temps (variables longitudinales) [114].

Dans la majorité des cas, l'hypothèse est faite que les deux facteurs G et E sont **indépendants dans la population générale**. En conséquence, la corrélation entre G et E (θ_{GE}) est considérée comme nulle. Cependant, ces facteurs peuvent être corrélés à cause de la stratification de population, si G influence un comportement qui influence à son tour E, ou encore s'ils sont influencés par un facteur de confusion non mesuré. Cette corrélation G-E peut entraîner une inflation de l'erreur de type I et la détection de faux positifs [115].

Dans le cas d'un facteur environnemental binaire et d'un modèle génétique additif, une interaction GxE peut être facilement visualisée graphiquement (**Figure 1.5**). On peut alors distinguer plusieurs types d'interactions :

- Effet de G chez les individus exposés à E uniquement (A) : le risque de maladie augmente en fonction du nombre de copies de l'allèle à risque (B) porté par les individus exposés à E alors que le génotype à ce SNP n'a aucun effet chez les individus non exposés à E

- Effet de G chez les individus non exposés à E uniquement (B) : le risque de maladie augmente en fonction du nombre de copies de l'allèle à risque (B) porté par les individus non exposés à E alors que le génotype à ce SNP n'a aucun effet chez les individus exposés à E

- **Effet flip-flop (C)** : le génotype étudié (BB ou AA) a un effet inverse sur la maladie en fonction de la présence ou l'absence d'exposition à E

- **Effets synergiques (D)** : les trois cas précédents représentent des cas extrêmes d'interaction, mais des cas intermédiaires existent avec par exemple un effet de G chez les exposés et les non exposés mais avec des intensités différentes entre les deux groupes

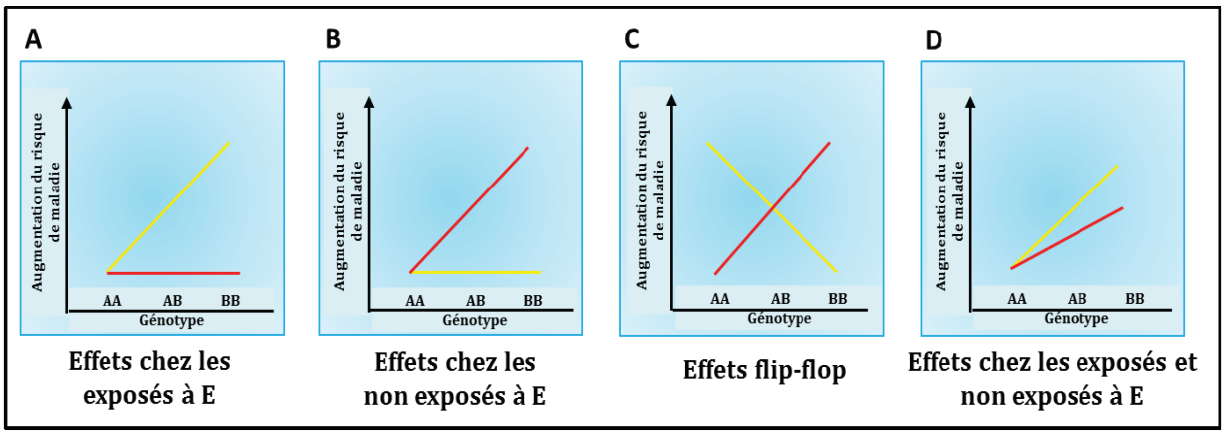


Figure 1.5. Types d'interactions gène-environnement. Le risque de maladie, en fonction du génotype à un SNP, est représenté en jaune pour les individus exposés au facteur environnemental E et en rouge pour les individus non exposés à E (adapté de Ober *et al*, *Trend Genet*, 2011) [116].

Les méthodes les plus classiques testent directement et une à une l'ensemble des interactions possibles entre l'exposition E et tous les SNPs du génome. Le modèle classique (M1) est communément utilisé dans les études cas-témoins (CC) quand E est connu chez les cas et les témoins. Les effets estimés par ce modèle ont l'avantage d'être toujours valides, bien que ce test

puisse être conservateur en cas d'indépendance G-E. Le modèle *Case-Only* (CO)[117] qui se base uniquement sur un échantillon de cas est plus puissant mais suppose l'hypothèse d'indépendance G-E : une corrélation G-E peut entraîner une inflation de l'erreur de type I et la détection de faux positifs. Il mesure l'association entre G et E chez les cas uniquement. Le modèle s'écrit sous la forme :

$$\text{Logit } P(E|G) = \beta_0 + \beta_{CO} \times G \text{ avec } (H0) : \beta_{CO} = 0$$

où β_{CO} représente l'effet estimé du facteur génétique sur l'environnement, nommé estimateur *Case-Only*. Afin de profiter de l'avantage de modèle CO tout en contrôlant son biais, un modèle Bayésien (*Empirical Bayes* (EB)) a été développé par Mukherjee et Chatterjee [118]. L'estimateur est une moyenne pondérée des estimateurs CO et CC qui dépend de la corrélation entre G et E (θ_{GE}). Si la corrélation tend vers zéro, l'estimateur bayésien coïncide avec l'estimateur CO ; mais en cas de corrélation entre G et E, un poids plus important est assigné à l'estimateur CC. L'estimateur s'écrit sous la forme :

$$\beta_{EB} = \frac{\sigma^2{}_{CC}}{\theta^2{}_{GE} + \sigma^2{}_{CC}} \beta_{CO} + \frac{\theta^2{}_{GE}}{\theta^2{}_{GE} + \sigma^2{}_{CC}} \beta_{CC}$$

où $\sigma^2{}_{CC}$ représente la variance de l'estimateur *Case-Control*, β_{CO} et β_{CC} sont les estimateurs de l'interaction issus des modèles *Case-Control* et *Case-Only* respectivement et θ_{GE} est la corrélation entre G et E. Une version plus générale de cette méthode permettant la sélection par pénalisation des estimateurs a été développée par Chen *et al.* [119] et implémentée dans le package CGEN. Li et Conti [120] ont également développés une variante de cette méthode qui définit les poids attribués aux estimateurs par des probabilités *a posteriori*.

Pour tous les estimateurs précédents, le test d'interaction est basé sur l'hypothèse nulle $H_0 : \beta_{GxE} = 0$ et suit une loi du χ^2 à un degré de liberté (**ddl**). D'autres méthodes peuvent être utilisées pour tester de manière simultanée l'effet de G et de l'interaction GxE en effectuant des **tests joints** à 2ddl pour augmenter la puissance de détection des interactions dans certaines situations. Le test proposé par Kraft *et al* [121] (aussi appelé Kraft 2-df) se base sur le modèle (M1) pour tester l'hypothèse $H_0 : \beta_G = \beta_{GxE} = 0$. Dai *et al* [122] ont proposé une méthode (Dai-2df) testant, dans un premier modèle, l'effet marginal de G sur la maladie puis dans un deuxième modèle, l'effet du terme d'interaction GxE. Les deux modèles étant indépendants [122,123], la statistique combinée suit une loi du χ^2 à 2ddl. La méthode **multinomial-GI** [124] a été proposée dans le cas où E n'est pas

disponible chez les témoins. Le phénotype est alors catégorisé en trois classes : 0 (les témoins), 1 (les cas non exposés) et 2 (les cas exposés). Ce test consiste à tester l'effet de G sur le phénotype séparément pour les cas exposés et les cas non exposés et à combiner les deux tests de rapport de vraisemblance (χ^2 à 2ddl).

3.3.2 Méthodes en deux étapes

Les méthodes directes testent chaque SNP indépendamment pour l'interaction et nécessitent d'appliquer une correction pour le nombre de tests effectués. Afin de réduire le problème de la perte de puissance lié à cette correction, ont été développées des méthodes d'interaction en deux étapes qui se décomposent en : 1) une première étape de **criblage** qui comporte un test T_1 (variable selon les méthodes) et une *P*-valeur P_1 correspondante pour tous les SNPs, utilisée comme critère de **sélection** ; 2) une seconde étape T_2 de test d'interaction GxE pour les SNPs sélectionnés, suivant les valeurs de P_1. L'ajustement de P_2 est effectué pour le nombre de tests effectué à l'étape 2 suivant une correction de Bonferroni classique, ou un test d'hypothèse pondéré [125–127]. Ces méthodes nécessitent l'indépendance des tests T_1 et T_2, et se différencient par l'étape 1 de criblage tandis que le test à l'étape 2 est le même.

La méthode DG (« *Disease-Gene Two-Step approach* ») [128] utilise les résultats de *P*-valeurs du GWAS pour le phénotype étudié comme **filtre** pour ensuite tester en interaction avec E uniquement les SNPs passant un certain seuil en GWAS en utilisant le modèle (M1). Le test suit une loi du χ^2 à 1ddl. Le test d'association de G avec la maladie (G-M) est indépendant du test d'interaction [123]. **La méthode EG** (« *Environment-Gene Two-Step approach*) [129] teste l'association de G avec E à l'étape 1 (G-E).

Les SNPs passant un certain seuil sont testés en interaction. Les deux étapes ne sont pas indépendantes si l'estimateur CO est utilisé pour le test d'interaction [123], ce qui peut biaiser la distribution des *P*-valeurs du test d'interaction. **Une approche hybride H2 (« *Hybrid Two-Step approach* »**) combinant ces deux méthodes a été proposée par Murcray *et al* [130]. Cette méthode alloue une fraction de l'erreur de type I à chaque criblage (ρ pour EG et (1-ρ) pour DG) [130]. Une autre méthode développée par Ege *et al* [131] combine les tests G-E et G-M en une statistique à 2ddls lors du test de criblage.

D'autres **méthodes dites « cocktails »** [123] combinent les tests G-E et G-M lors du test de criblage selon la méthode H2, pour la **sélection** des SNPs pour l'étape 2 de test de l'interaction. Ces

méthodes testent l'interaction GxE pour l'ensemble des SNPs avec une correction pour les tests multiples pondérée selon le niveau de significativité des SNPs à l'étape 1 [125–127]. Deux variantes existent suivant l'hypothèse d'indépendance ou non des deux tests de corrélation et d'association marginale. 1) La méthode **Cocktail-I** qui suppose **l'indépendance de ces tests** utilise un seuil arbitraire pour déterminer le test à utiliser pour la sélection des SNPs à l'étape 1 ($P_{G\text{-}M}$ si $P_{G\text{-}M} \leq s$, $P_{G\text{-}E}$ sinon). 2) La méthode **Cocktail-II** qui suppose **la dépendance de ces tests**, utilise comme seuil de significativité la *P*valeur minimum des deux tests.

3.3.3 Méthodes multi-marqueurs

Afin d'augmenter la puissance de détection des interactions GxE, des méthodes combinant les signaux d'interactions par groupe de SNPs (ou gènes) biologiquement pertinents ont été développées. L'idée de ces méthodes est d'augmenter la puissance de détection en **regroupant des signaux d'interactions** possiblement trop faibles pour être détectés individuellement, en **entités génétiques biologiques pertinentes** (comme des gènes ou des pathways) testées en interaction avec l'environnement. Un unique score de risque GxE global est calculé pour l'interaction entre chaque groupe de SNP et l'environnement plutôt que pour chaque SNPs un par un, réduisant le poids de la correction pour les tests multiples. Ces méthodes peuvent être réparties en trois catégories : 1) les méthodes de pondérations des coefficients d'interaction, dites « burden-type » (**BT**), 2) les méthodes basées sur la composante de la variance des tests d'interaction, dites variance composent (**VC**), 3) et les méthodes combinant les deux.

Les méthodes BT proposent un score GxE par gène, construit à partir des interactions SNPxE appartenant au gène, selon le modèle suivant :

$$Logit\ P(M|G,E) = \beta_0 + \beta_E \times E + G.\beta_G + \mu \times E.\mathbf{G}.\mathbf{W} + \beta_C.C$$

où G est le vecteur des SNPs appartenant au gène, W le vecteur des poids pour chaque SNP. L'hypothèse testée est (H0) : $\mu = 0$. La méthode **SBERIA** [132] utilise les statistiques de test G-M pour attribuer des poids aux interactions SNPxE. Liu *et al* [133] ont développé une autre méthode qui utilise la corrélation G-E pour définir les poids utilisés. Ces méthodes nécessitent le choix d'un seuil, et ne permettent pas de prendre en compte les directions des effets des interactions SNPxE. La méthode **GRS-interaction-training** [134] propose d'utiliser les effets des interactions SNPxE eux même comme poids à attribuer à la statistique GxE. Cependant ces effets doivent être estimés sur un échantillon indépendant afin de ne pas biaiser les estimations. **Les méthodes VC** supposent que

les effets SNPxE sont des effets aléatoires suivant une distribution d'espérance nulle et de variance ρ en étendant la méthode SKAT [50] aux interactions GxE. Ces méthodes testent l'hypothèse (H0) : ρ = 0, mais peuvent varier selon les stratégies utilisées pour modéliser les effets marginaux, ou le type de traits analysés [83,135,136]. La méthode **GESAT** [136] utilise notamment une régression pénalisée (de type ridge) pour estimer les effets marginaux. Les méthodes BT sont plus performantes quand de nombreux SNPs dans le gène sont causaux et ont des effets dans la même direction. A l'inverse, les méthodes VC sont plus puissantes en présence d'hétérogénéité dans la direction et l'importance des effets. Pour prendre avantage de ces deux types de méthodes en fonction du modèle sous-jacent, des méthodes hybrides ont été développées et implémentées dans le programme MiST-I [137]. Notamment, les méthodes **eSBERIA** [138] et **coSBERIA** [138] combinent la méthode SBERIA et une méthode développée par Sun *et al* [139]. Des méthodes analogues ont également été développées pour les tests joints [140]. Il existe d'autres méthodes de ce type spécifiques à l'analyse des variants rares et qui ne sont pas discutées ici [141].

Synthèse sur les méthodes d'interaction GxE

Dans le cadre de ma thèse (chapitre IV), j'ai effectué une méta-analyse des facteurs d'interaction de l'exposition au tabagisme passif durant la petite enfance sur le délai de survenue de l'asthme dans l'enfance. Cette analyse faisait suite à une précédente méta-analyse pangénomique simple-marqueur du délai de survenue de l'asthme à laquelle j'ai contribué. L'asthme, puis l'asthme dans l'enfance, ont été modélisés par des analyses de survies. Peu de méthodes d'interaction permettent d'analyser des phénotypes modélisés ainsi.

Si les modèles à deux degrés de liberté ont montré, d'une manière générale, posséder plus de puissance que ceux testant l'effet marginal ou d'interaction seul [122], la pénalisation de la statistique de test par un degré de liberté supplémentaire peut provoquer une importante perte de puissance dans le cas d'un effet marginal relativement faible (notamment pour les effets flip-flop), par rapport à au test de l'effet d'interaction GxE (1ddl) [122]. Les méthodes en deux étapes ont été montrées comme plus **puissantes** dans certains cas que les méthodes en une étape testant l'effet de l'interaction seul [142] mais dépendent très fortement de l'effet marginal et/ou de la corrélation G-E. Les méthodes de regroupements nécessitent de choisir un seuil arbitraire de sélection des SNPs à tester pour l'interaction GxE. La plupart de ces méthodes nécessitent de réserver une partie de l'échantillon d'analyse pour l'évaluation des poids à attribuer aux interactions, ce qui peut être

problématique compte tenu des problèmes de puissances inhérents aux analyses pangénomiques d'interaction. Dans le cas contraire, ces méthodes dépendent également fortement de l'effet marginal et/ou de la corrélation G-E. Globalement, **la puissance des méthodes dépend très fortement du modèle d'interaction sous-jacent**, qui est inconnu avant analyse.

Dans la mesure où les effets marginaux ont déjà été étudiés à travers un GWAS dans l'asthme, suivant la même modélisation, j'ai choisi de me focaliser sur le test d'interaction GxE (1ddl) non dépendant des effets marginaux.

Table 1.6. Synthèse des trois catégories de méthodes d'analyses d'interactions multiplicatives

Catégories	Modélisation de l'interaction	Etape de sélection ou pondération	Spécificité des données	Exemple de méthodes
Méthodes pangénomiques directes	Interaction "pure"(chi2 à 1 ddl)	/	cas-témoin	Régression CC, EB [118]
		/	cas uniquement	Régression CO
	Effet joint de l'interaction et de l'effet marginal (chi2 à 2 ddls)	/	cas-témoin	Kraft-2df [121], Dai-2df [122]
		/	Information de E non disponible chez les témoins	multinomial-GI [124]
Méthodes en deux étapes : sélection / association	Interaction "pure"(chi2 à 1 ddl)	Association SNP-M	cas-témoin	DG [128]
		Corrélation SNP-E	cas-témoin	EG [129]
		Sélection hybride : association SNP-M / corrélation SNP-E	cas-témoin	H2 [130], méthodes « Cocktails » [123], Ege *et al* [131]
Méthodes de regroupements	BT : Score basé sur les effets pondérés des interactions à l'intérieur d'un gène	Variables indicatrices suivant l'association du SNP-M	cas-témoin	SBERIA [132]
		Variables indicatrices suivant la corrélation du SNP- E	cas-témoin	Liu *et al* [133]
		Coefficients SNPxE	cas-témoin	GRS-interaction-training [134]
	VC : Test d'un effet aléatoire global GxE (score test)	/	cas-témoin	GESAT [136]
	Combinaison de statistiques de type BT et VC suivant la méthode de Fisher	Pondération hybride : association SNP-M / corrélation SNP-E	cas-témoin	eSBERIA [138]
			cas uniquement	coSBERIA [138]

3.4 Méta-analyses d'études d'associations pangénomiques

Dans le cadre de mon travail de thèse, j'avais accès à différents échantillons de données. Afin de pouvoir les analyser conjointement, les résultats des analyses de ces échantillons ont été combinés par méta-analyse. Une méta-analyse est une analyse statistique qui permet de combiner les résultats de plusieurs études de même type et indépendantes. Son objectif est l'estimation d'un paramètre global (ou combiné), dans l'ensemble des échantillons, à partir des estimations de ce paramètre dans chacun des échantillons (une méta-analyse peut également être appliquée à l'estimation de plusieurs paramètres). L'un des avantages des méta-analyses est d'augmenter la puissance statistique de l'étude par la prise en considération d'un nombre de sujets plus important qu'une étude unique. Un autre avantage des méta-analyses est de permettre de tester l'homogénéité des estimations des paramètres selon les études et, si ce test d'homogénéité est significatif, d'évaluer la variabilité des estimations entre études.

3.4.1 Méta-analyses d'études d'associations pangénomiques maladie-marqueur génétique

Plusieurs modèles existent pour estimer l'effet global d'un marqueur génétique sur la maladie dans l'ensemble des échantillons, dont le choix va influencer la méthode d'estimation des paramètres. Le **modèle à effet fixe** suppose que le paramètre qui est estimé, l'effet du SNP sur la maladie, a la même valeur dans toutes les populations représentées par les échantillons inclus dans la méta-analyse et donc que les différentes estimations du paramètre dans chaque échantillon sont des estimations du même paramètre, et suivent une loi normale centrée sur cet effet commun. Dans ce cas, l'estimation de l'effet global est une moyenne pondérée des estimations de l'effet du SNP dans chaque échantillon en utilisant l'inverse de la variance de l'estimation dans chaque échantillon comme poids. Ceci permet de donner plus d'importance aux études comportant un grand nombre de sujets. Dans une méta-analyse de N études, si l'estimation de l'effet du SNP sur la maladie et sa variance dans l'étude i sont respectivement β_i et s_i^2, l'estimation combinée β_{fix} et sa variance $\mathrm{Var}(\beta_{fix})$ sur l'ensemble des études est :

$$\beta_{fix} = \frac{\sum_{i=1}^{N} w_i \beta_i}{\sum_{i=1}^{N} w_i} \quad , \quad Var(\beta_{fix}) = \frac{1}{\sum_{i=1}^{N} w_i} \quad avec\ w_i = \frac{1}{s_i^{\,2}}$$

Il convient ensuite de **tester l'homogénéité** des β_i entre études à l'aide du **Q-test de Cochran**, qui correspond à la somme pondérée des carrés des différences entre l'effet combiné estimé et les effets estimés dans chacune des N études. Si $w_i = \frac{1}{s_i^2}$ est le poids de l'étude i dans le modèle à effet fixe, la statistique Q définie ci-dessous suit une loi du χ^2 à N-1 degrés de liberté :

$$Q = \sum_{i=1}^{N} w_i(\beta_{fix} - \beta_i)^2$$

Le Q-test peut manquer de puissance lorsque le nombre d'études est très faible, et peut-être trop sensible si le nombre d'études est grand [143]. **La statistique I^2**, qui décrit le pourcentage de variation entre études due à l'hétérogénéité (plutôt que due au hasard), permet de prendre en compte de manière simple et intuitive le nombre d'études. Soit df le degré de liberté (df = N-1), la statistique I^2 est définie comme suit :

$$I^2 = \begin{cases} 100 * \dfrac{Q - (N-1)}{Q}, & si\ Q \geq N-1 \\ 0, & si\ Q < N-1 \end{cases}$$

Dans le cas particulier où la variance σ^2 dans tous les échantillons est égale et connue ($\sigma^2 = s_i^2$ pour tout i), le I^2 peut s'écrire comme le pourcentage de variance inter-étude τ^2 dans la variance totale, c'est-à-dire la somme des variances inter-études et intra-études : $I^2 = \tau^2/(\sigma^2 + \tau^2)$.

$$o\grave{u}\ \tau^2 = \max\left\{0\ ;\ \frac{Q - (N-1)}{\sum_{i=1}^{N} w_i - \frac{\sum_{i=1}^{N} w_i^2}{\sum_{i=1}^{N} w_i}}\right\}$$

S'il y a très peu de variation entre les échantillons (τ^2 nul ou faible), alors la statistique I^2 sera faible et un modèle à effets fixes peut être approprié. Dans le cas d'une hétérogénéité importante, le modèle à effet fixe peut s'avérer inadapté [144]. En effet, l'hétérogénéité des estimations de l'effet du SNP entre études peut être dû au fait que cet effet diffère eff

ectivement selon les études. Il est alors préférable d'utiliser un **modèle à effets aléatoires**, qui ne fait pas l'hypothèse que l'effet estimé est le même dans toutes les populations représentées par les différents échantillons, et permet d'intégrer une variabilité inter-études τ^2 au modèle à effets fixes.

L'estimateur β_{ran} de l'effet du SNP dans le modèle à effets aléatoires et sa variance V (β_{ran}) sont de la forme :

$$\beta_{ran} = \frac{\sum_{i=1}^{N} v_i \beta_i}{\sum_{i=1}^{N} v_i} \quad , \qquad V(\beta_{ran}) = \frac{1}{\sum_{i=1}^{N} v_i} \qquad avec \;\; v_i = \frac{1}{s_i^2 + \tau^2}$$

3.4.2 Méta-analyses d'études pangénomiques d'interaction gène-environnement

Peu de méthodes ont été proposées dans le cadre spécifique des méta-analyses d'interaction GxE. Pour la plupart des **méthodes d'interaction GxE en une étape**, des méta-analyses classiques (à effets fixes ou à effets aléatoires) peuvent être effectuées en combinant l'interaction GxE estimée dans chaque échantillon, comme vu dans le paragraphe précédent pour l'effet d'un SNP sur la maladie. Des extensions des méthodes classiques ont été proposées, en particulier lorsque les effets du SNP et de l'interaction SNP-environnement sont conjointement testés. Aschard *et al* [145] ont proposé d'effectuer des analyses d'association simple marqueur chez les exposés et non exposés séparément. Les résultats des GWAS sont combinés dans chaque strate par méta-analyse puis les statistiques du test joint (G, GxE) et du test d'interaction GxE peuvent être calculés facilement à partir des statistiques des deux méta-analyses. La statistique du test joint (2ddl) est égale à la somme χ^2 des effets dans chaque strate (méthode de Fisher). La statistique d'interaction GxE est basée sur la différence des effets chez les exposés et non exposés, et suit une loi du χ^2 à 1 degré de liberté :

$$\chi^2_{1ddl} = \frac{(\widehat{\beta_{Exp}} - \widehat{\beta_{Non-exp}})^2}{Var(\widehat{\beta_{Exp}}) + Var(\widehat{\beta_{Non-exp}}) - 2Cov(\widehat{\beta_{Exp}}, \widehat{\beta_{Non-exp}})}$$

La valeur de la covariance entre les estimations peut notamment être approximée à partir de la corrélation entre les vecteurs de chaque strate de l'ensemble des estimations des effets des SNPs sur la maladie. Manning *et al* [146] ont proposé une autre méthode de méta-analyse dans un contexte de test joint. Les effets de G, de l'interaction GxE et leur matrice de covariance sont estimés dans chaque échantillon. La méthode estime ensuite les effets G et GxE et une matrice de covariance pour l'ensemble des études et test l'hypothèse (H0) : $\beta_G = \beta_{GxE} = 0$. Ce test a été proposé dans le cas d'un facteur environnemental binaire ou quantitatif et a montré des performances équivalentes à la méthode de méta-analyse stratifiée sur le statut d'exposition. Il est à noter que la méthode Dai-2df qui teste dans une première étape l'effet du SNP sur la maladie, puis dans un deuxième temps

l'effet d'interaction GxE, nécessite de combiner par méta-analyse séparément les effets des SNPs, puis ceux du terme d'interaction GxE, afin de reconstruire le test joint.

Les méthodes d'interaction GxE en deux étapes nécessitent des procédures de méta-analyses plus conséquentes à mettre en oeuvre. Ces méthodes nécessitent de combiner les résultats d'association à chaque étape. La plupart de ces méthodes nécessitent donc d'effectuer trois méta-analyses d'associations dans l'ensemble des échantillons d'analyse. Pour les **méthodes de regroupement des SNPs au niveau d'entités biologiques** (gènes, pathways), les processus de méta-analyses peuvent s'avérer encore plus complexes et peuvent poser d'importants problèmes d'interprétabilité. Les poids attribués à chaque interaction SNPxE calculés séparément dans chaque échantillon peuvent conduire à des statistiques de gènes basées sur des interactions SNPxE différentes, ce qui rend les méthodes de méta-analyses non adaptées pour la plupart de ces méthodes.

4. L'asthme et l'atopie

L'objectif de cette thèse est de développer des stratégies d'analyses d'interactions gène-gène et gène environnement, pour identifier de nouveaux variant génétiques impliqués dans l'asthme et l'atopie (sensibilisation aux allergènes). Dans la partie qui va suivre, je vais présenter deux traits - définition, épidémiologie et physiopathologie - ainsi que décrire les facteurs de risques génétiques et environnementaux associés à ces phénotypes.

4.1 Définition

L'asthme est une maladie inflammatoire chronique des voies respiratoires caractérisée par des épisodes récurrents d'essoufflements (dyspnée) accompagnées de sifflements, de sensation d'oppression thoracique, et de toux. Ces épisodes sont souvent associés à une obstruction ventilatoire de degré variable, réversible spontanément ou sous l'effet d'un traitement. L'asthme est associé à une **hyperréactivité bronchique** (HRB) correspondant à une obstruction bronchique excessive en réponse à des agents physiques, chimiques ou pharmacologiques. La fréquence et l'ampleur des crises peuvent être variables, pouvant se produire jusqu'à plusieurs fois par jour, et de manière plus violente la nuit ou après un exercice physique, pouvant créer insomnies, fatigue durant la journée, et absentéisme scolaire et professionnel [147].

Il est reconnu que l'asthme est une maladie hétérogène représentant plus une collection de différentes entités qu'une maladie unique. Certains facteurs permettent de différencier les différentes entités d'asthme comme l'âge d'apparition de l'asthme (asthme de l'enfant, asthme de l'adulte), la sévérité de la maladie, les facteurs déclenchants des crises d'asthme (comme par exemple les expositions professionnelles) et des variations de réponses au traitement [148]. C'est une maladie complexe associée à différentes mesures comme les mesures de la fonction ventilatoire (comme la diminution du volume expiratoire maximal à la première seconde (VEMS)), ou les mesures de l'inflammation (pente de réactivité bronchique, nombre d'éosinophiles sanguins circulants). L'asthme est également très souvent associé à l'atopie et à d'autres **maladies allergiques**, comme la rhinite allergique et la dermatite atopique. **L'atopie** un facteur de risque associé à l'asthme. Elle est définie *stricto sensu* par une réponse positive aux tests cutanés à des allergènes communs, mais peut aussi selon les définitions, regrouper une élévation du taux sérique des Immunoglobuline E (IgE) totales et spécifiques, et l'éosinophilie. L'asthme apparaît souvent

dans l'enfance, où il est principalement d'origine allergique est souvent associé à une histoire d'asthme dans la famille. Il peut rester stable tout au long de la vie, disparaître pendant plusieurs années, réapparaître, et s'aggraver à tout âge. Chez l'adulte, il faut distinguer les sujets qui sont malades depuis leur enfance et ceux qui ont développé la maladie à l'âge adulte, dans ce dernier cas, l'asthme est plus rarement d'origine allergique.

4.2 Epidémiologie

L'asthme est l'une des pathologies chroniques les plus fréquentes dans le monde, et la plus fréquente chez l'enfant et l'adolescent. Les décès dus à l'asthme ont diminué dans les pays industrialisés avec l'utilisation régulière de glucocorticoïdes inhalés mais l'impact global de l'asthme reste élevé. Les prévalences de l'asthme et de l'allergie n'ont cessé d'augmenter au cours des dernières décennies. On estime à 340 millions le nombre de personnes souffrant actuellement de l'asthme à travers le monde, et à environ 1000 morts chaque jour [149]. Deux études multicentriques internationales, l'étude ISAAC (*International Study of Allergy and Asthma in Childhood*) [150] pour les enfants et les adolescents, et l'étude ECRHS (*European Community Respiratory Health Survey*) [151] pour les adultes ont permis de mettre en évidence de fortes disparités des prévalences de l'asthme et de l'allergie en fonction de l'âge (enfants/adultes) et entre les différents pays du globe : en général la prévalence est plus importante dans les pays industrialisés. Pour les enfants et les adolescents, la prévalence de l'asthme varie entre 2 et 30%. La prévalence de l'asthme est plus faible chez l'adulte et est estimée entre 4 et 5 % dans le monde [149] (**Figure 1.7**). La prévalence de l'allergie, est estimée entre 25 et 30%.

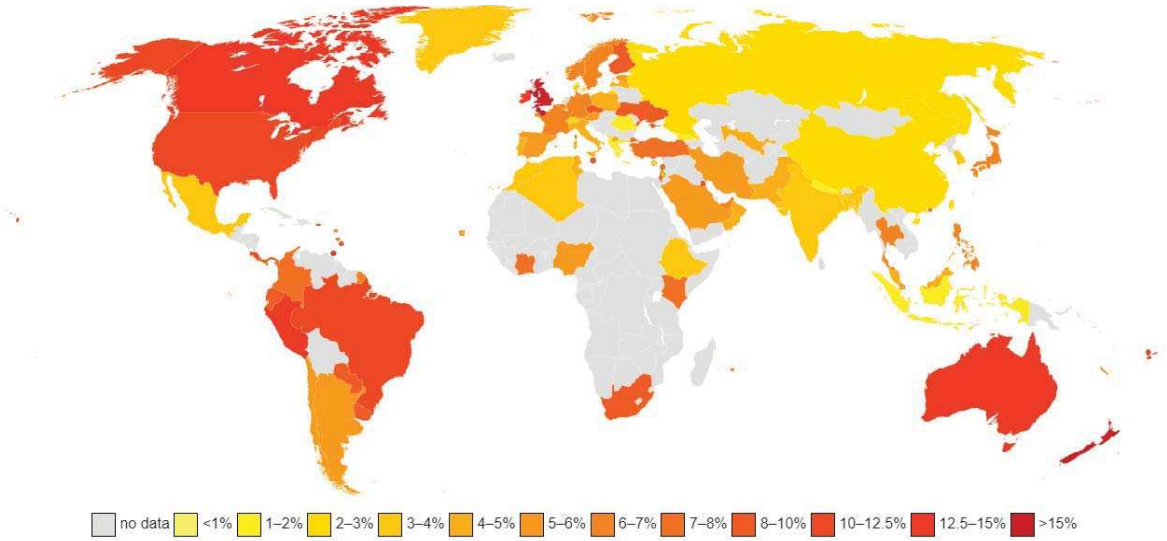


Figure 1.7 : Prévalence des symptômes d'asthme à travers le monde en 2004 (https://en.wikipedia.org/wiki/Asthma)

4.3 Physiopathologie

L'asthme est un processus dynamique impliquant des mécanismes immunitaires, une inflammation chronique et un remodelage de l'épithélium des voies respiratoires qui surviennent de façon concomitante ou de manière successive. **L'inflammation** joue un rôle central dans la physiopathologie de l'asthme. L'inflammation des voies aériennes implique une interaction de nombreux types de cellules et de médiateurs dans les voies respiratoires. Il donne lieu aux principales caractéristiques de la maladie : l'inflammation bronchique et la limitation du débit d'air qui provoquent des épisodes récurrents de toux, de respiration sifflante et d'essoufflement. Les processus détaillés par lesquels ces événements interactifs se produisent et mènent à l'asthme clinique ne sont pas encore entièrement connus. Cependant, malgré l'existence de sous-types d'asthme distincts (par exemple, intermittent, persistant léger, persistant modéré ou persistant sévère), l'inflammation des voies aériennes reste un mécanisme omniprésent.

Parmi les différents types d'asthme, **l'asthme allergique** est le plus fréquent, notamment chez l'enfant. Il s'agit de la forme d'asthme la plus grave sur le court terme, le degré de réaction bronchique pouvant être particulièrement important et parfois mortel. L'asthme allergique est en général caractérisé par la survenue d'une ou de plusieurs crises causées par une réaction excessive des bronches à un agent allergène extérieur. La stimulation allergénique induit une forte

production d'IgE. La crise d'asthme allergique se manifeste par une obstruction soudaine et de progression rapide des voies bronchiques, créant suffocation, puis étouffement. Cette forme d'asthme peut évoluer en asthme chronique, notamment si l'exposition à l'allergène est récurrente. **Les facteurs déclenchants** de cette forme commune d'asthme sont en général des **aéroallergènes** : acariens, poils d'animaux, pollens, spores de moisissures. L'asthme allergique peut également souvent être associé à la rhinite allergique.

4.4 Facteurs de risque

L'étiologie de l'asthme est complexe, mettant en jeu de multiples mécanismes physiopathologiques, causes de différentes et nombreuses manifestations cliniques. Les causes de l'asthme n'ont pas encore été complètement élucidées. Les nombreux processus biologiques impliqués dans l'asthme et les phénotypes associés, comme l'atopie, suggèrent que les facteurs génétiques peuvent être spécifiques à un phénotype, ou commun à plusieurs traits, selon qu'ils jouent un rôle dans une voie physiologique donnée ou qu'ils se placent à la jonction de différentes voies. En résulte une maladie multifactorielle, mettant en jeu de nombreux facteurs génétiques et environnementaux et des processus plus complexes d'interaction entre ces facteurs (**Figure 1.8**).

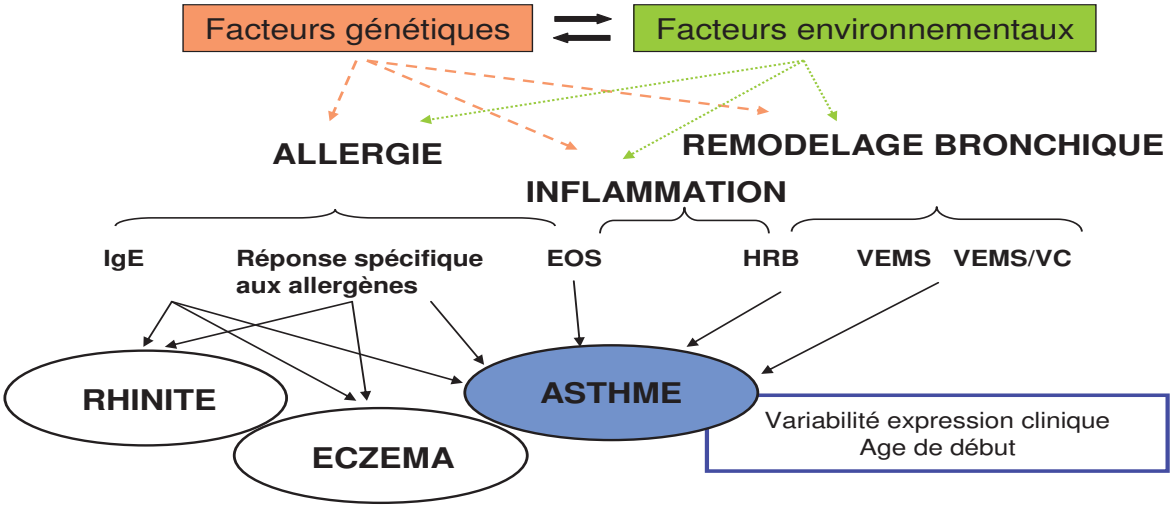


Figure 1.8 : L'asthme et l'allergie sont des maladies complexes multifactorielles, résultant de multiples facteurs génétiques et environnementaux et de leurs interactions

Facteurs environnementaux

L'augmentation de la prévalence de l'asthme et de l'allergie au cours des dernières décennies est vraisemblablement due en grande partie à la modification de **facteurs environnementaux** [152]. Les changements de mode de vie et une augmentation des expositions aux allergènes ont été suggérés comme pouvant être des facteurs de risque de l'asthme et/ou de l'allergie dans les pays développés. Parmi ces facteurs, l'exposition à des substances déclenchant les crises d'asthme comme les allergènes d'intérieurs tels que les acariens dans la literie, les tapis; la pollution et les squames d'animaux domestiques; les allergènes d'extérieurs tels que les pollens ; les moisissures; l'exposition à la fumée de tabac (active ou passive) et les irritants chimiques dans le milieu professionnel [153]. Le moment d'exposition au cours de la vie est aussi un facteur déterminant du risque d'apparition de la maladie (**Figure 1.9**), qui peut varier pour l'asthme de l'enfant et celui de l'adulte : par exemple, la présence d'un chat dans la maison avant l'âge de trois ans est protecteur pour l'asthme et la sensibilisation allergique, mais à risque plus tard au cours de la vie [116]. Des études ont montrées que l'asthme maternel est un risque très important d'asthme et d'allergie de l'enfant, suggérant l'importance majeure de l'environnement sur le risque ultérieur d'asthme et d'allergie [154–156]. D'autres études ont également montrées que **l'exposition au tabagisme passif *in utero* ou dans l'enfance** est un facteur de risque d'asthme et de sa sévérité [157,158]. L'exposition *in utero* au tabagisme passif est associée à une augmentation du risque de développer de l'asthme dans l'enfance [159], et l'exposition au tabagisme passif paternel et maternel durant l'enfance est également associée à une augmentation du risque d'asthme à l'âge adulte [160]. Des études ont également montré une association entre tabagisme actif et risque d'asthme à l'âge adulte [161–163].

D'autres facteurs environnementaux sont connus comme les infections virales respiratoires précoces [164], l'obésité [165,166], la pollution de l'air [167], ou encore la diminution de l'activité physique, les habitudes alimentaires, et les facteurs hormonaux. En dehors de ces facteurs de risque délétères, la recherche s'est portée sur des facteurs protecteurs qui pourraient avoir diminué dans les dernières décennies comme les contacts avec les agents infectieux dans la petite enfance [168], ne permettant pas le développement normal de la réponse immunitaire, le rôle protecteur des grandes fratries ou des modes de vie traditionnels avec un contact avec des animaux de ferme [116].

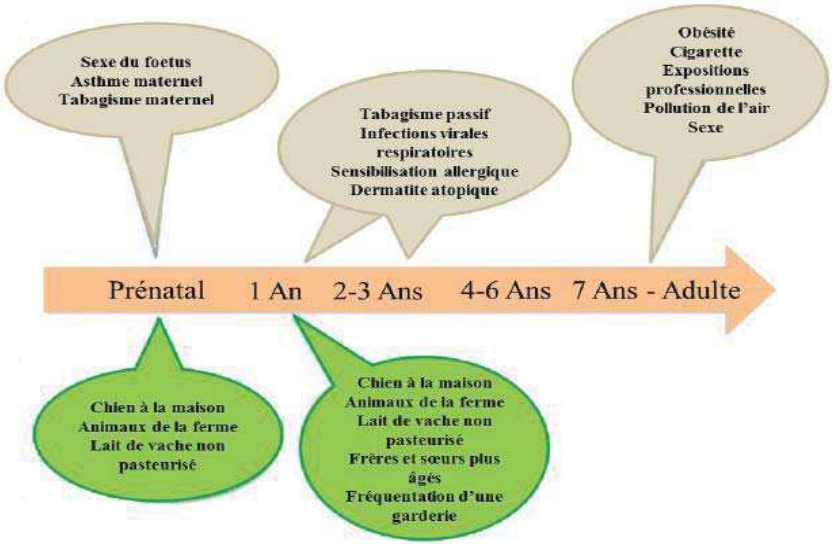


Figure 1.9 : Facteurs environnementaux protecteurs (vert) ou délétères (gris) influençant l'asthme à des périodes spécifiques de la vie (adapté de Ober *et al*, *Trend Genet*, 2011) [116].

La composante génétique

Les facteurs génétiques jouent également un rôle important dans l'asthme et l'atopie [169]. Des études familiales ont montrées une importante contribution génétique, allant de 25 à 80% pour le risque de l'asthme [170] et de 40 à 85% pour celui de l'atopie [171,172], avec un risque relatif de développer de l'asthme 2,5 à 3 fois plus élevé chez les frères et sœurs de sujets asthmatiques qu'en population générale [173].

Afin de caractériser les facteurs génétiques impliqués dans l'asthme et les phénotypes liés à l'asthme (comme l'atopie, les niveaux d'IgE et la fonction ventilatoire), de nombreuses études ont été réalisées, incluant les analyses de liaison génétique, les études gènes candidats et les études d'association pangénomiques. Des analyses de liaison génétique ont révélées plus de 70 régions liées à l'asthme et l'atopie, bien que toutes ces régions n'aient pas toujours été répliquées dans des échantillons indépendants. Ces régions incluaient parfois des gènes candidats, mais aussi des nouveaux gènes identifiés par clonage positionnel [174] (analyse de liaison suivi d'une analyse d'association dans la région identifiée), mais le rôle fonctionnel de ces gènes est encore mal connu [175,176]. Le nombre de gènes candidats possibles dans l'asthme et l'allergie est très important, étant donné les mécanismes physiopathologiques multiples impliqués dans ces maladies. C'est pourquoi les premières études d'associations se sont focalisées sur des gènes candidats. A ce jour, plus 1000

études d'association avec des gènes candidats ont été publiées caractérisant plus de 200 loci associés avec l'asthme et les phénotypes liés à l'asthme. Cependant, seulement 32 gènes ont été retrouvés associés aux phénotypes de l'asthme de manière constante dans au moins cinq études indépendantes (**Table 1.10**). Ces gènes peuvent être classés en quatre grandes catégories : (1) les gènes impliqués dans l'immunité innée et la régulation immunitaire (*CD14*, *TLR2*, *TLR4*); (2) les gènes impliqués dans la réponse immunitaire Th2 (*IL4, IL13*, *IL4RA*, *FCER1B*); (3) les gènes impliqués dans la biologie de l'épithélium des voies aériennes et l'immunité au niveau des muqueuses (*CCL5*, *CCL11, SPINK5*); les gènes impliquées dans la fonction ventilatoire et le remodelage de l'épithélium des voies respiratoires (*ADRB2*, *TNF*, *NOS1*, *ADAM33*) [177–179].

Table 1.10 : Loci trouvés associés à l'asthme (asthme, hyperréactivité bronchique et niveaux d'IgE) et répliqués dans au moins cinq études gènes candidats indépendantes. Adapté de March *et al* (2013) [176].

Gène	Région chromosomique	Fonction
GSTM1	1p13.3	Détoxification, élimination des produits du stress oxydatif
FLG	1q21.3	Intégrité épithéliale et fonction barrière de l'épiderme
IL10	1q31-q32	Cytokine - Régulation immunitaire
CTLA4	2q33	Contrôle/inhibition des réponses des cellules T / Régulation immunitaire
IL13	5q31	Induit les fonctions effectrices immunitaires Th2
IL4	5q31.1	Différentiation Th2
CD14	5q31.1	Détection de microbes – Reconnaît les modèles moléculaires associés aux agents pathogènes
ADRB2	5q31-q32	Relaxation des muscles lisses
SPINK5	5q32	Inhibiteur épithélial de sérine-protéase
HAVCR1	5q33.2	Réponse des cellules T - Récepteur du virus de l'hépatite A
LTC4S	5q35	Synthèse des leucotriènes - Médiateur inflammatoire
LTA	6p21.3	Médiateur inflammatoire
TNF	6p21.3	Médiateur inflammatoire
HLA-DRB1	6p21	Complexe majeur d'histocompatibilité de classe II – présentation antigénique
GPRA	7p14.3	Régulation de l'expression des métallo-protéases, effets neuronaux
NAT2	8p22	Détoxification
GSTP1	11q13	Détoxification, élimination des produits du stress oxydatif
FCER1B	11q13	Récepteurs des IgE - Atopie
IL18	11q22.2-q22.3	Inflammation
CC16	11q12.3-q13.1	Fonction immune-régulatrice potentielle - expression épithéliale
STAT6	12q13	Signalisation IL-4 et IL-13
NOS1	12q24.2-q24.31	Oxyde nitrique synthase — communication cellulaire
CMA1	14q11.2	Chymase – protéase exprimée par les mastocytes
IL4R	16p12.1-p12.2	Chaîne alpha de récepteurs pour IL-4 et IL-13
CCL11	17q21.1-q21.2	Eoxtaxin-1 - composé chimio-attirant des éosinophiles

Gène	Région chromosomique	Fonction
CCL5	17q11.2-q12	RANTES — chimio-attirant pour cellules T, éosinophiles et basophiles
ACE	17q23.3	Régulation de l'inflammation
TBXA2R	19p13.3	Agrégation plaquettaire
TGFB1	19q13.1	Influence la croissance cellulaire, la différenciation, la prolifération, l'apoptose
ADAM33	20p13	Interactions cellule-cellule et cellule-matrice
GSTT1	22q11.23	Détoxification, élimination des produits du stress oxydatif

Abréviations : IgE, immunoglobuline E; IL, interleukine; RANTES, Cellules T normales et régulées, exprimées et sécrétées; T_H, lymphocytes T auxiliaires.

L'évolution des technologies de génotypages a ensuite conduit à l'ère des **études d'association pangénomiques**, permettant une analyse systématique de centaines de milliers puis de millions de variants génétiques tout le long du génome. La première étude d'association pangénomique simple marqueur (GWAS) de l'asthme a été réalisée en 2007 dans des populations d'origine européenne [180]. Cette étude avait analysée plus de 300 000 SNPs génotypés chez 3 237 enfants (994 asthmatiques et 1 243 non-asthmatiques) britanniques (MRC-UK) et allemands (MAGICS) et comportait deux études de réplication portant sur 2 320 enfants allemands (ISAAC Phase II) et 3 301 adultes britanniques (BC58 : *1958 British Birth Cohort*). Cette GWAS a permis d'identifier **le locus 17q12-q21** associé à l'asthme (comprenant notamment les gènes *ORMDL3*, *GSDMB* et *GSDMA*). Faisant suite à ce GWAS, l'étude de cette région dans 372 familles de l'étude française EGEA (1 511 sujets) a confirmé ces résultats et a démontré que l'effet des polymorphismes de la région 17q12-21 était spécifique de l'asthme apparaissant à un âge précoce, avant l'âge de 4 ans dans EGEA [181], qui correspond à la période de respiration sifflante dans la petite enfance [182].

Le locus17q12-21 est une région majeur de l'asthme qui a par la suite été confirmé par de nombreux GWAS dans des populations de divers origines, et notamment dans deux méta-analyses appliquées à un grand nombre d'études. La première conduite dans le cadre du **consortium européen GABRIEL** [183] regroupait 23 études et plus de 26 000 individus (10 365 asthmatiques et 16 110 non asthmatiques) génotypés avec un panel d'environ 600 000 SNPs. Cette méta-analyse a permis d'identifier six régions associées à l'asthme : 2q12.1 (*IL1RL1, IL18R1*), 6p21 (*HLA-DQ*), 9p24.1 (*IL33*), 15q22.33 (*SMAD3*), 17q12-21 (*GSDMB/ORMDL3*), 22q12.3 (*IL2RB*). Cette étude qui considérait différents types d'asthme (asthmes de l'enfant et de l'adulte, professionnel, sévère) a

permis de confirmer que la région 17q12-21 était spécifique de l'asthme dans l'enfance (asthme avant 16 ans), et a également identifié les régions 2q12 et 9p24.1 comme plus associées à l'asthme dans l'enfance, et la région 6p21 plus associée à l'asthme de l'adulte.

Une seconde méta-analyse a été réalisée dans le cadre du consortium américain EVE. Elle regroupait 9 études incluant près de 13 000 individus (3 601 asthmatiques, 3 853 témoins, 1 702 trios cas-parents) de différentes origines ethniques (américains d'origine européenne, afro-américains/caribéens, et latino-américains), et un panel de 2 à 3 millions de SNPs imputatés (HapMap 2, release 21), selon l'étude et le groupe ethnique [184]. Cette étude a identifié cinq régions incluant quatre loci précédemment rapportés (2q12.1, 9p24.1, 5q22.1 et 17q12-21) mais identifiées pour la première fois associés à l'asthme dans trois groupes d'origine ethnique différente, et une nouvelle région 1q23 (*PYHIN*1) spécifique aux sujets d'origine afro-américaine. Egalement, une région spécifique aux latino-américains (3q27) avait été suggérée sans être répliquée. Au total, 20 GWAS ont permis de mettre en évidence un grand nombre de **régions chromosomiques associées à l'asthme** (**Table 1.11**). La majorité de ces études a été menées dans des populations d'origine européenne. Cependant, on dénombre quelques études ayant identifié au moins un locus associés au risque d'asthme dans des populations japonaises, latino-américaines ou d'ascendances multiples (certaines ont d'abord été conduites dans des populations d'origines européennes, puis les résultats de ces études ont été répliqués dans des populations d'origine ethnique différentes). Parmi ces GWAS, sept études se sont intéressées uniquement à l'asthme apparaissant dans l'enfance, et trois à l'asthme de l'adulte.

Récemment, la méta-analyse à grande échelle TAGC (*Trans-National Asthma Genetic Consortium*, [185]) a porté sur plus de 140 000 sujets (23 948 asthmatiques et 118 538 non asthmatiques) de différentes origines à travers le monde : européenne (19 954 cas, 107 715 témoins), africaine-américaine (2 149 cas, 6 055 témoins), japonaise (1 239 cas, 3 976 témoins), et latino-américaine (606 cas 792 témoins) ; et un panel de presque 3 millions de SNPs (imputations HapMap 2, release 21). Cette étude a permis d'identifier un total de 878 SNPs au niveau de 18 loci qui atteignaient le niveau de signification génome entier. Ces 18 loci incluaient : cinq nouveaux loci associés au risque d'asthme (les gènes candidats les plus vraisemblables au sein de chaque locus sont indiqués entre parenthèse) : 5q31.3 (*NDFIP1*), 6p22.1 (*ZSCAN12* et *ZSCAN31*), 6q15 (*BACH2*), 12q13.3 (*STAT6*), et 17q21.33 (*GNGT2*) ; deux nouvelles associations aux loci 6p21.33 et 10p14 qui étaient indépendantes des signaux précédemment rapportés dans des populations

latino-américaines et japonaies ; deux régions 8q21.13 et 16p13.13, précédemment rapportées pour le phénotype combiné « asthme et rhinite allergiques » mais pas pour l'asthme lui-même ; neuf loci précédemment identifiés par plusieurs GWAS. Plusieurs des régions caractérisées par cette étude étaient des régions de grande taille (pouvant couvrir plus de deux Mégabases). En particulier la région 17q12-21 s'étendait sur plus de 800kb et incluait un nouveau signal dans le gène *ERBB2* (à 200kbs en amont du locus connu *GSDMB/ORMDL3*) : le signal dans le gène *ERBB2* était le plus significatif dans l'ensemble des études et ne montrait pas d'hétérogénéité selon l'âge de début de l'asthme tandis que le locus *GSDMB/ORMDL3* était le plus significatif dans le sous-groupe d'études d'asthme pédiatrique (comme précédemment observé). Cette étude a également permis de mettre en évidence l'enrichissement des loci identifiés en marques épigénétiques caractéristiques d'éléments de régulation appelés « enhancers » dans les cellules immunitaires, indiquant que les variants génétiques identifiés jouent un rôle dans la régulation de la réponse immunitaire et que des des mécanismes épigénétiques ont un rôle clef dans l'étiologie de l'asthme.

De manière très intéressante, cette étude a aussi montré que l'asthme partageait des loci avec d'autres maladies ayant une composante immunitaire ou inflammatoire (non seulement les maladies autoimmunes mais également les maladies cardiovasculaire, neuro-psychiatriques et les cancers). Plus récemment, une méta-analyse de GWAS de l'asthme incluant des données de l'étude UK-Biobank a porté sur 45 425 individus (7 041 cas, 38 384 témoins), renseignés pour un panel d'environ 7,5 millions (1000G/UK10K), avec réplication dans les données du consortium GABRIEL [186]. Cette étude a détecté 17 loci associés à l'asthme, dont six nouveaux : 2p25.1, 2q37.3, 4p14, 4q27, 7p21.1. Au total, 36 loci ont été associés avec le risque d'asthme *per se*.

Les études génétiques s'intéressant au risque d'atopie sont moins nombreuses. Plusieurs études gènes candidats ont été effectuées mais ont conduit à des résultats inconsistants [187]. La première GWAS de l'atopie, réalisée dans les données de la *British 1958 Birth Cohort* (1 083 cas et 2 770 témoins) avec étude de réplication dans trois cohortes britanniques indépendantes, n'avaient pas permis de détecter de résultat significatif au niveau génome-entier [188]. Par la suite, trois méta-analyses de GWAS ont permis de détecter 19 régions associées à l'atopie (**Table 1.12**). La première méta-analyse a étudié le risque de sensibilisation allergique définie par un taux sanguin élevé d'IgE spécifiques à des allergènes, ou par une réaction cutanée positive à des allergènes. Cette étude incluait plus de 15 000 individus d'origine européenne (5 789 cas et 10 056 témoins) issus de 16 études des Consortiums EAGLE (*EArly Genetics and Lifecourse Epidemiology*) et AAGC

(*Australian Asthma Genetics Consortium*) [189]. La seconde méta-analyse a porté sur 53 862 individus (22 012 cas et 31 850 témoins) de populations d'origine européenne issues des cohortes 23andMe et ALSPAC (*The Avon Longitudinal Study of Parents and Children*) et sur 2,5 millions de SNPs communs aux deux échantillons. Dans ces deux populations, l'atopie avait été définie de à partir de réponses à un auto-questionnaire : auto-déclaration d'allergie aux chats, aux pollens ou aux poussières [190]. La troisième méta-analyse incluait 24 481 individus d'origine européenne (8 040 cas, 16 441 témoins) issus de 13 études [191]. Cette méta-analyse a portée sur le risque de sensibilisation allergique définie par un taux sanguin élevé d'IgE spécifiques et/ou une réaction cutanée positive à des allergènes (très variables selon les études). Les SNPs analysés avaient été imputés avec 1000Genome Phase 1. Parmi les 19 loci identifiés par ces méta-analyses, cinq ont été trouvés associés à l'atopie dans les trois méta-analyses : 2q12.1, 3q28, 4p14, 5q22.1, 11q13.5 ; et parmi ces cinq régions, on retrouvait trois loci associés à l'asthme : 2q12.1, 4p14 et 5q22.1. Ainsi, la région 2q12.1 (*IL1RL1, IL1RL2, IL18R1*) avait été identifiée par quatre méta-analyses de GWAS de l'asthme (GABRIEL, EVE, TAGC et UK-Biobank)., la région 4p14 dans UK-Biobank, et la région 5q22.1 (*TSLP*) par trois méta-analyses (EVE, TAGC et UK-Biobank) dans des populations d'ascendance multiple, et également chez des japonais [192]. De plus, la grande région des antigènes leucocytaires humains (HLA) située en 6p21 a été rapportée par deux des trois méta-analyses de l'atopie, et de nombreuses fois dans l'asthme par des GWASs dans des populations de différentes origines. Cette région inclut également des signaux d'association retrouvés pour à la fois l'asthme de l'enfant et l'asthme de l'adulte. Une étude approfondie de cette région dans des populations d'origine ethnique différente s'avère nécessaire pour élucider quels sont les différents gènes impliquées dans différentes formes d'asthmes [193].

Des méta-analyses d'études pangénomiques à grande échelle ont également été conduites sur des **combinaisons de phénotypes** se rapportant à l'asthme et/ou à des maladies allergiques [186,194,195].

Interaction entre facteurs génétiques et environnementaux

La première étude d'interaction pangénomique de **l'asthme de l'enfant et de l'atopie** a été conduite chez 1 708 enfants de régions d'Europe centrale (population GABRIELA issue du consortium GABRIEL) [131], n'a pas permis de montrer d'interaction entre facteur génétique et **l'exposition au mode de vie de la ferme** au seuil génome entier. Seuls deux études pangénomiques d'interactions GxE avec le tabac ont été publiées sur l'asthme, dans des populations d'origine

européenne. Une étude sur l'**asthme dans l'enfance** a rapporté des interactions entre **l'exposition passive *in utero* au tabagisme** maternel et le locus 18p11, et entre l'exposition au tabagisme parental durant l'enfance et le locus 6q26 [196]. Une autre analyse similaire conduite pour l'**asthme de l'adulte** a rapporté des interactions entre le **tabagisme actif** et les locus 9p23 et 12p12 [197]. Une autre étude d'interaction pangénomique conduite pour l'**asthme de l'enfant** a rapporté des interactions entre le locus 5p15 et la **pollution de l'air extérieur** (représentée par le niveau d'exposition au NO_2) dans des populations d'origine européenne [198]. Cependant, aucune de ces études pangénomiques d'interaction gène-environnement n'atteignait le seuil de significativité génome entier. A ma connaissance, aucune analyse pangénomique d'interaction GxG n'a été conduite dans l'asthme ou l'atopie.

Les locus génétiques identifiés par ces analyses n'expliquent qu'une partie du risque génétique, ce qui pourrait-être en partie dû à un rôle important de l'exposition à des facteurs environnementaux et l'hétérogénéité phénotypique caractéristique de l'asthme. Des approches complémentaires permettant de prendre en compte des mécanismes complexes, comme les analyses d'interaction gène-gène et gène-environnement telles qu'effectuées dans cette thèse, peuvent s'avérer utiles pour identifier de nouveaux gènes impliqués dans l'asthme et l'allergie.

Table 1.11 : Les 30 loci retrouvés associés à l'asthme (asthme, niveaux d'IgE) à travers des études d'associations pangénomiques

Région	Population de découverte (N)	Population de réplication (N)	Phénotype	Gènes reportés	Meilleur signal (*P*)	Réf
1q21.3	Ascendance européenne (32 442)	Ascendance européenne (25 358)	Asthme	*IL6R*	$2,3x10^{-8}$	Ferreira et al, 2011 [199]
	Ascendance européenne (45 425)	Ascendance européenne (26 475)	Asthme	*C1orf68, CRCT1, …*	$3,1x10^{-8}$	Zhu et al, 2018 [186]
1q23.1-2	Ascendance européenne (6 819)	Ascendance européenne (7 809)	Niveaux d'IgE	*FCER1A, DARC, OR10J3*	$4,5x10^{-26}$	Granada et al, 2011 [200]
	Ascendance européenne (1 530)	Ascendance européenne (9 769)	Niveaux d'IgE	*FCER1A*	$1,9x10^{-20}$	Weidinger et al, 2008 [201]
	Ascendance multiple (5 388 cas)	Ascendance multiple (7 173 cas)	Asthme	*PYHIN1*	$4,0x10^{-9}$	Torgerson et al, 2011 [184]
1q31.3	Ascendance européenne (2 781)	Ascendance multiple (6 175)	Asthme dans l'enfance (3-12 ans)	*DENND1B, CRB1*	$1,7x10^{-13}$	Sleiman et al, 2010 [202]
2p25.1	Ascendance européenne (45 425)	Ascendance européenne (26 475)	Asthme	*LINC00299*	$6,8x10^{-10}$	Zhu et al, 2018 [186]
2q12.1	Ascendance européenne (27 375) et multiple (142 486)	-	Asthme	*IL1RL1, IL1RL2, IL18R1*	$3,9x10^{-21}$	Demenais et al, 2018 [185]
	Ascendance européenne (45 425)	Ascendance européenne (26 475)	Asthme	*IL1R1, IL1RL1, IL1RL2, …*	$7,4x10^{-18}$	Zhu et al, 2018 [186]
	Ascendance multiple (5 388 cas)	Ascendance multiple (7 173 cas)	Asthme	*IL1RL1*	$2,0x10^{-15}$	Torgerson et al, 2011 [184]
	Ascendance européenne (18 604)	Ascendance européenne (15 286)	Asthme de l'adulte	*IL1RL1, IL18R1*	$1.1x10^{-9}$	Ramasamy et al, 2012
	Ascendance européenne (26 475)	-	Asthme	*IL18R1*	$3.4x10^{-9}$	Moffat et al, 2010 [183]
2q37.3	Ascendance européenne (45 425)	Ascendance européenne (26 475)	Asthme	*D2HGDH, GAL3ST2*	$3,9x10^{-11}$	Zhu et al, 2018 [186]
3p26.2	Ascendance européenne (573)	Ascendance européenne (931)	Asthme dans l'enfance	*IL5RA*	$2,3x10^{-8}$	Forno et al, 2012 [203]
4p14	Ascendance européenne (45 425)	Ascendance européenne (26 475)	Asthme	*FAM114A1, MIR574, …*	$2,6x10^{-10}$	Zhu et al, 2018 [186]
4q12	Ascendance européenne	Ascendance multiple	Asthme dans l'enfance (5-12 ans)	*SRIP1, MIR548AG1*	$2,0x10^{-8}$	Ding et al, 2013 [204]

Région	Population de découverte (N)	Population de réplication (N)	Phénotype	Gènes reportés	Meilleur signal (*P*)	Réf
4q27	Ascendance européenne (45 425)	Ascendance européenne (26 475)	Asthme	*ADAD1, IL2, IL21, …*	$3,3x10^{-10}$	Zhu et al, 2018 [186]
4q31.21	Japonaise (4 836)	Japonaise (30 247)	Asthme de l'adulte	*USP38, GAB1*	$1,9x10^{-12}$	Hirota et al, 2011 [192]
5q12.1	Ascendance européenne (1 955)	Ascendance multiple (22 971)	Asthme dans l'enfance	*PDE4D*	$3,0x10^{-8}$	Himes et al, 2009 [205]
5q22.1	Ascendance européenne (27 375) et multiple (142 486)	-	Asthme	*SLC25A46, TSLP*	$9,4x10^{-26}$	Demenais et al, 2018 [185]
	Japonaise (4 836)	Japonaise (30 247)	Asthme de l'adulte	*TSLP, WDR36*	$1,2x10^{-16}$	Hirota et al, 2011 [192]
	Ascendance multiple (5 388 cas)	Ascendance multiple (7 173 cas)	Asthme	*TSLP*	$1,0x10^{-14}$	Torgerson et al, 2011 [184]
	Ascendance européenne (45 425)	Ascendance européenne (26 475)	Asthme	*SLC25A46, TMEM232, TSLP*	$4,6x10^{-10}$	Zhu et al, 2018 [186]
5q31.1	Ascendance européenne (6 819)	Ascendance européenne (7 809)	Niveaux d'IgE	*IL13*	$3,4x10^{-18}$	Granada et al, 2011 [200]
	Ascendance européenne (27 375) et multiple (142 486)	-	Asthme	*IL13, RAD50, IL4*	$5,0x10^{-16}$	Demenais et al, 2018 [185]
	Ascendance européenne (45 425)	Ascendance européenne (26 475)	Asthme	*IL13, RAD50, IL4, …*	$1,3x10^{-11}$	Zhu et al, 2018 [186]
	Ascendance européenne (26 475)	-	Asthme	*IL13*	$1,4x10^{-8}$	Moffat et al, 2010 [183]
	Ascendance européenne (1 530)	Ascendance européenne (9 769)	Niveaux d'IgE	*RAD50*	$4,5x10^{-8}$	Weidinger et al, 2008 [201]
5q31.3	Ascendance multiple (142 486)	-	Asthme	*NDFIP1, GNDPA1, SPRY4*	$7,9x10^{-9}$	Demenais et al, 2018 [185]
6p22.1	Ascendance européenne (6 819)	Ascendance européenne (7 809)	Niveaux d'IgE	*HLA-G, HLA-A*	$3,6x10^{-9}$	Granada et al, 2011 [200]
	Ascendance européenne (27 375)	-	Asthme	*GPX5, TRIM27*	$5,3x10^{-9}$	Demenais et al, 2018 [185]
6p21.32-33	Ascendance européenne (45 425)	Ascendance européenne (26 475)	Asthme	*HLA-DOB, HLA-DQB1, HLA-DQA1, HLA-DQA2*	$1,31x10^{-35}$	Zhu et al, 2018 [186]
	Ascendance européenne (27 375) et multiple (142 486)	-	Asthme	*HLA-DQB1, HLA-DQA1, MICB, HCP5, MCCD1*	$4,8x10^{-28}$	Demenais et al, 2018 [185]

Région	Population de découverte (N)	Population de réplication (N)	Phénotype	Gènes reportés	Meilleur signal (*P*)	Réf
	Japonaise (4 836)	Japonaise (30 247)	Asthme de l'adulte	*NOTCH4, BTNL2, C6orf10, HLA-DOA, HLA-DRA, HLA-DQA2, HLA-DQB1, PBX2,*	$4{,}1\times10^{-23}$	Hirota et al, 2011 [192]
	Ascendance européenne (26 475)	Ascendance européenne (meta-analyse)	Asthme	*HLA-DQB1*	$7{,}0\times10^{-14}$	Moffat et al, 2010 [183]
	Japonaise (1 180)	Japonaise (2 474)	Niveaux d'IgE	*HLA-C, HCG27*	$1{,}1\times10^{-10}$	Yatagai et al, 2013 [206]
	Japonaise (3 314)	Japonaise (3 106)	Asthme dans l'enfance	*HLA-DPA1, HLA- DPB1*	$2{,}3\times10^{-10}$	Noguchi et al, 2011 [207]
	Ascendance européenne (18 604)	Ascendance européenne (15 286)	Asthme de l'adulte	*BTNL2, HLA-DRA*	$1{,}1\times10^{-8}$	Ramasamy et al, 2012 [208]
	Ascendance européenne (6 819)	Ascendance européenne (7 809)	Niveaux d'IgE	*HLA-DQA2*	$1{,}4\times10^{-8}$	Granada et al, 2011 [200]
	Ascendance européenne (3 855)	Ascendance multiple (>11 000)	Asthme de l'adulte	*HLA-DQA1*	$2{,}0\times10^{-8}$	Lasky-Su et al, 2012 [209]
	Latino-américaine (3774)	Latino-américaine	Asthme dans l'enfance	*MUC22*	$P<5\times10^{-6}$	Galanter et al, 2014 [210]
6q15	Ascendance européenne (27 375) et multiple (142 486)	-	Asthme	*BACH2, GJA10, MAP3K7*	$8{,}6\times10^{-13}$	Demenais et al, 2018 [185]
7p21.1	Ascendance européenne (45 425)	Ascendance européenne (26 475)	Asthme	*ITGB8*	$2{,}1\times10^{-9}$	Zhu et al, 2018 [186]
7q22.3	Ascendance européenne (3 684)	Ascendance multiple (21 229)	Asthme sévère dans la petite enfance (2-6 ans)	*CDHR3*	$3{,}0\times10^{-14}$ ($P_{\text{hétérogénéité}}=2{,}7\times10^{-7}$)	Bonnelykke et al, 2014 [211]
8q21.13	Ascendance européenne (27 375) et multiple (142 486)	-	Asthme	*TPD52, ZBTB10*	$1{,}1\times10^{-10}$	Demenais et al, 2018 [185]
	Ascendance européenne (45 425)	Ascendance européenne (26 475)	Asthme	*ZBTB10*	$3{,}7\times10^{-10}$	Zhu et al, 2018 [186]
8q24.11	Japonaise (3 314)	Japonaise et coréenne (3 106)	Asthme dans l'enfance	*SLC30A8*	$5{,}0\times10^{-13}$	Noguchi et al, 2011 [207]
9p23	Ascendance européenne	Ascendance multiple	Asthme dans l'enfance (1-18 ans)	*JKAMPP1, TYRP1*	$8{,}0\times10^{-9}$	Ding et al, 2013 [204]

Région	Population de découverte (N)	Population de réplication (N)	Phénotype	Gènes reportés	Meilleur signal (*P*)	Réf
9p24.1	Ascendance européenne (45 425)	Ascendance européenne (26 475)	Asthme	*GLDC, IL33, TPDS2L3, UHRF2*	$7,4x10^{-23}$	Zhu et al, 2018 [186]
9p24.1	Ascendance européenne (27 375) et multiple (142 486)	-	Asthme	*RANBP6, IL33*	$7,2x10^{-20}$	Demenais et al, 2018 [185]
	Ascendance européenne (3 684)	Ascendance multiple (21 229)	Asthme sévère dans la petite enfance	*IL33*	$9,0x10^{-13}$	Bonnelykke et al, 2014 [211]
	Ascendances multiples (5 388 cas)	Ascendances multiples (7 173 cas)	Asthme	*IL33*	$2,0x10^{-12}$	Torgerson et al, 2011 [184]
	Ascendance européenne (26 475)	-	Asthme	*IL33*	$9,2x10^{-10}$	Moffat et al, 2010 [183]
10p14	Japonaise (4 836)	Japonaise (30 247)	Asthme de l'adulte	*LOC338591*	$1,8x10^{-15}$	Hirota et al, 2011 [192]
	Ascendance européenne (27 375) et multiple (142 486)	-	Asthme		$3,5x10^{-9}$	Demenais et al, 2018 [185]
10q24.2	Ascendance européenne	Ascendance multiple	Asthme dans l'enfance (1-18 ans)	*HPSE2*	$5,0x10^{-8}$	Ding et al, 2013 [204]
11q13.5	Ascendance européenne (45 425)	Ascendance européenne (26 475)	Asthme	*C11orf30, PRKRIR*	$2,2x10^{-21}$	Zhu et al, 2018 [186]
	Ascendance européenne (27 375) et multiple (142 486)	-	Asthme	*EMSY, LRRC32*	$2,2x10^{-14}$	Demenais et al, 2018 [185]
	Ascendance européenne (32 442)	Ascendance européenne (25 358)	Asthme	*LRRC32*	$1,8x10^{-8}$	Ferreira et al, 2011 [199]
11q24.2	Ascendance européenne (573)	Ascendance européenne (931)	Asthme dans l'enfance	-	$6,5x10^{-9}$	Forno et al, 2012 [203]
12q13.2-3	Japonaise (4 836)	Japonaise (30 247)	Asthme de l'adulte	*IKZF4, CDK2*	$2,3x10^{-13}$	Hirota et al, 2011 [192]
	Ascendance européenne (6 819)	Ascendance européenne (7 809)	Niveaux d'IgE	*STAT6, NAB2*	$2,0x10^{-12}$	Granada et al, 2011 [200]
	Ascendance européenne (45 425)	Ascendance européenne (26 475)	Asthme	*GPR182, MYO1A, NAB2, …*	$1,1x10^{-11}$	Zhu et al, 2018 [186]
	Ascendance multiple (142 486)	-	Asthme	*STAT6, NAB2, LRP1*	$3,9x10^{-9}$	Demenais et al, 2018 [185]

Région	Population de découverte (N)	Population de réplication (N)	Phénotype	Gènes reportés	Meilleur signal (*P*)	Réf
15q22.2	Ascendance européenne (27 375) et multiple (142 486)	-	Asthme	*RORA, NARG2, VPS13C*	$1,9x10^{-10}$	Demenais et al, 2018 [185]
	Ascendance européenne (18 604)	Ascendance européenne	Asthme	*RORA*	$2,4x10^{-9}$	Ramasamy et al, 2012 [208]
15q22.33	Ascendance européenne (45 425)	Ascendance européenne (26 475)	Asthme	*SMAD3*	$2,6x10^{-23}$	Zhu et al, 2018 [186]
	Ascendance européenne (27 375) et multiple (142 486)	-	Asthme	*SMAD3, SMAD6, AAGAB*	$7,4x10^{-15}$	Demenais et al, 2018 [185]
	Ascendance européenne (26 475)	-	Asthme	*SMAD3*	$3,9x10^{-9}$	Moffat et al, 2010 [183]
16p13.13	Ascendance européenne (45 425)	Ascendance européenne (26 475)	Asthme	*CLEC16A, DEXI*	$9,1x10^{-17}$	Zhu et al, 2018 [186]
	Ascendance européenne (27 375) et multiple (142 486)	-	Asthme	*CLEC16A, DEXI, SOCS1*	$2,1x10^{-10}$	Demenais et al, 2018 [185]
17q12-21	Ascendance européenne (27 375) et multiple (142 486)	-	Asthme	*ERBB2, PGAP3, MIEN1*	$2,2x10^{-30}$	Demenais et al, 2018 [185]
	Ascendance européenne (3 684)	Ascendance multiple (21 229)	Asthme sévère dans la petite enfance (2-6 ans)	*GSDMB*	$6,4x10^{-23}$	Bonnelykke et al, 2014 [211]
	Ascendance européenne (3 684)	Ascendance multiple (21 229)	Asthme sévère dans la petite enfance	*GSDMA*	$3,0x10^{-21}$	Bonnelykke et al, 2014 [211]
	Ascendance européenne (45 425)	Ascendance européenne (26 475)	Asthme	*ERBB2, GRB7, GSDMA, …*	$7,0x10^{-18}$	Zhu et al, 2018 [186]
	Ascendance européenne (26 475)	-	Asthme dans l'enfance	*GSDMA*	$3,0x10^{-17}$	Moffat et al, 2010 [183]
	Ascendance multiple (5 388 cas)	Ascendance multiple (7 173 cas)	Asthme	*GSDMB*	$2,0x10^{-16}$	Torgerson et al, 2011 [184]
	Latino-américaine (3774)	Latino-américaine	Asthme dans l'enfance	*IKZF3*	$5,7x10^{-13}$	Galanter et al, 2014 [210]
	Ascendance européenne (2 237)	Ascendance européenne (5 621)	Asthme dans l'enfance	*ORMDL3*	$9,0x10^{-11}$	Moffat et al, 2007 [180]

Région	Population de découverte (N)	Population de réplication (N)	Phénotype	Gènes reportés	Meilleur signal (*P*)	Réf
	Ascendance européenne (4 279)	Ascendance européenne (1 576)	Asthme sévère	*ORMDL3*	$1,0x10^{-8}$	Wan et al, 2012 [212]
17q21.33	Ascendance européenne (27 375) et multiple (142 486)	-	Asthme	*ZNF652, PHB*	$3,3x10^{-9}$	Demenais et al, 2018 [185]
22q12.3	Ascendance européenne (26 475)	-	Asthme	*IL2RB*	$1,2x10^{-8}$	Moffat et al, 2010 [183]

Table 1.12 : Les 18 loci retrouvés associés à l'atopie à travers des études d'associations pangénomiques

Région	Population de découverte (N)	Population de réplication (N)	Phénotype	Gènes reportés	Meilleur signal (*P*)	Réf
1q24	Ascendance européenne (24 481)	-	Atopie (IgE spécifiques)	*FASLG*	$9,2x10^{-10}$	Waage et al, 2018 [191]
2p25.1	Ascendance européenne (24 481)	-	Atopie (IgE spécifiques)	*LINC00299*	$1,3x10^{-8}$	Waage et al, 2018 [191]
	Ascendance européenne (46 646)	Ascendance européenne (7 216)	Atopie (chat, pollens, poussières)	*IL1RL1, IL1RL2*	$1,6x10^{-19}$	Hinds et al, 2013 [190]
	Ascendance européenne (15 845)	Ascendance européenne (16 034)	Atopie (IgE spécifiques)	*IL1RL1*	$4,9x10^{-11}$	Bonnelykke et al, 2013 [189]
	Ascendance européenne (24 481)	-	Atopie (IgE spécifiques)	*IL1RL1*	$3,4x10^{-9}$	Waage et al, 2018 [191]
2q33.1	Ascendance européenne (46 646)	Ascendance européenne (7 216)	Atopie (chat, pollens, poussières)	*PLCL1*	$6,1x10^{-10}$	Hinds et al, 2013 [190]
3q28	Ascendance européenne (15 845)	Ascendance européenne (16 034)	Atopie (IgE spécifiques)	*LPP*	$2,7x10^{-10}$	Bonnelykke et al, 2013 [189]
	Ascendance européenne (46 646)	Ascendance européenne (7 216)	Atopie (chat, pollens, poussières)	*LPP*	$1,2x10^{-9}$	Hinds et al, 2013 [190]
	Ascendance européenne (24 481)	-	Atopie (IgE spécifiques)	*LPP*	$3,8x10^{-9}$	Waage et al, 2018 [191]
4p14	Ascendance européenne (46 646)	Ascendance européenne (7 216)	Atopie (chat, pollens, poussières)	*TLR1, TLR6*	$5,3x10^{-21}$	Hinds et al, 2013 [190]
	Ascendance européenne (24 481)	-	Atopie (IgE spécifiques)	*TLR1, TLR6*	$2,2x10^{-13}$	Waage et al, 2018 [191]

Région	Population de découverte (N)	Population de réplication (N)	Phénotype	Gènes reportés	Meilleur signal (*P*)	Réf
	Ascendance européenne (15 845)	Ascendance européenne (16 034)	Atopie (IgE spécifiques)	*TLR1-TLR6-TLR10*	$5{,}2x10^{-11}$	Bonnelykke et al, 2013 [189]
4q24	Ascendance européenne (24 481)	-	Atopie (IgE spécifiques)	*NFKB1*	$2{,}0x10^{-8}$	Waage et al, 2018 [191]
4q27	Ascendance européenne (15 845)	Ascendance européenne (16 034)	Atopie (IgE spécifiques)	*IL2- ADAD1*	$5{,}5x10^{-10}$	Bonnelykke et al, 2013 [189]
	Ascendance européenne (24 481)	-	Atopie (IgE spécifiques)	*IL2*	$1{,}4x10^{-8}$	Waage et al, 2018 [191]
5p13.1	Ascendance européenne (46 646)	Ascendance européenne (7 216)	Atopie (chat, pollens, poussières)	*PTGER4*	$8{,}2x10^{-11}$	Hinds et al, 2013 [190]
5q22.1	Ascendance européenne (46 646)	Ascendance européenne (7 216)	Atopie (chat, pollens, poussières)	*WDR36 - CAMK4*	$2{,}3x10^{-20}$	Hinds et al, 2013 [190]
	Ascendance européenne (15 845)	Ascendance européenne (16 034)	Atopie (IgE spécifiques)	*SLC25A46*	$5{,}2x10^{-14}$	Bonnelykke et al, 2013 [189]
	Ascendance européenne (24 481)	-	Atopie (IgE spécifiques)	*SLC25A38P1, TMEM232*	$6{,}9x10^{-11}$	Waage et al, 2018 [191]
	Ascendance européenne (12 347)	-	IgE spécifiques (herbe)	*TMEM232, SLCA25A46*	$1{,}2x10^{-8}$	Ramasamy et al, 2011 [213]
6p21.32-33	Ascendance européenne (46 646)	Ascendance européenne (7 216)	Atopie (chat, pollens, poussières)	*HLA-DQA1- HLA-DQB1, HLA-C-MICA*	$7{,}1x10^{-15}$	Hinds et al, 2013 [190]
	Ascendance européenne (15 845)	Ascendance européenne (16 034)	Atopie (IgE spécifiques)	*HLA-DQA1, HLA-B-MICA*	$2{,}2x10^{-12}$	Bonnelykke et al, 2013 [189]
	Ascendance européenne (12 347)	-	IgE spécifiques (herbe)	*HLA region*	$1{,}6x10^{-9}$	Ramasamy et al, 2011 [213]
8q24.21	Ascendance européenne (15 845)	Ascendance européenne (16 034)	Atopie (IgE spécifiques)	*MYC-PVT1*	$5{,}4x10^{-10}$	Bonnelykke et al, 2013 [189]

Région	Population de découverte (N)	Population de réplication (N)	Phénotype	Gènes reportés	Meilleur signal (*P*)	Réf
9p24.1	Ascendance européenne (46 646)	Ascendance européenne (7 216)	Atopie (chat, pollens, poussières)	*RANBP6- IL33*	$1,7x10^{-9}$	Hinds et al, 2013 [190]
11q13.5	Ascendance européenne (46 646)	Ascendance européenne (7 216)	Atopie (chat, pollens, poussières)	*C11orf30 - LRRC32*	$1,6x10^{-19}$	Hinds et al, 2013 [190]
	Ascendance européenne (15 845)	Ascendance européenne (16 034)	Atopie (IgE spécifiques)	*C11orf30*	$1,4x10^{-18}$	Bonnelykke et al, 2013 [189]
	Ascendance européenne (24 481)	-	Atopie (IgE spécifiques)	*C11orf30*	$2,1x10^{-14}$	Waage et al, 2018 [191]
	Ascendance européenne (12 347)	-	IgE spécifiques (herbe)	*C11orf30, LRRC32*	$9,4x10^{-9}$	Ramasamy et al, 2011 [213]
12q13.2-3	Ascendance européenne (15 845)	Ascendance européenne (16 034)	Atopie (IgE spécifiques)	*STAT6*	$1,0x10^{-14}$	Bonnelykke et al, 2013 [189]
	Ascendance européenne (24 481)	-	Atopie (IgE spécifiques)	*STAT6*	$1,1x10^{-9}$	Waage et al, 2018 [191]
14q21.1	Ascendance européenne (46 646)	Ascendance européenne (7 216)	Atopie (chat, pollens, poussières)	*FOXA1 - TTC6*	$4,8x10^{-8}$	Hinds et al, 2013 [190]
15q22.33	Ascendance européenne (46 646)	Ascendance européenne (7 216)	Atopie (chat, pollens, poussières)	*SMAD3*	$1,2x10^{-8}$	Hinds et al, 2013 [190]
17q12-21	Ascendance européenne (46 646)	Ascendance européenne (7 216)	Atopie (chat, pollens, poussières)	*GSDMB*	$8,9x10^{-9}$	Hinds et al, 2013 [190]
20q13.2	Ascendance européenne (46 646)	Ascendance européenne (7 216)	Atopie (chat, pollens, poussières)	*NFATC2*	$6,9x10^{-9}$	Hinds et al, 2013 [190]

5. Plan du travail de thèse

Les études d'association génétique de l'asthme et de l'atopie ont identifié des variants fréquents avec des effets relativement modestes qui n'expliquent qu'une part de la composante génétique de ces maladies. Ces analyses manquent de puissance pour détecter des variants ayant un effet individuel faible ou interagissant avec d'autres variants ou des facteurs de l'environnement. Considérer des mécanismes complexes comme des interactions gène-gène et gène-environnement peut permettre d'identifier de nouveaux gènes associés à l'asthme et à l'atopie et d'expliquer ainsi une part plus importante du risque génétique de ces maladies. De plus, l'hétérogénéité phénotypique est caractéristique de l'asthme. Considérer des types d'asthme (comme l'asthme survenant dans l'enfance) ou des phénotypes intermédiaires associés à l'asthme (comme l'atopie) peut accroitre la puissance de détection de facteurs génétiques. L'objectif général de mon travail de thèse était de proposer des stratégies d'analyses d'interactions gène-gène et gène environnement pour mettre en évidence de nouveaux facteurs génétiques associés à l'asthme et à l'atopie.

L'ensemble de ce travail a été effectué dans le cadre de collaborations internationales, impliquant des études européennes et une étude canadienne qui font partie du **consortium GABRIEL** et dont les données pour l'asthme sont hébergées au sein du laboratoire UMR-946 (http://genestat.inserm.fr/fr/). Les données utilisées lors de cette thèse sont décrites dans le chapitre II.

Le **premier volet de ma thèse,** décrit dans le chapitre III, avait pour objectif la mise en œuvre d'une stratégie d'analyse d'interaction gène-gène pour identifier de nouveaux gènes impliqués dans l'atopie. L'atopie est définie comme une réponse cutanée positive à au moins un aéroallergène. Au lieu d'effectuer une analyse d'interaction gène-gène sur l'ensemble du génome qui, le plus souvent, manque de puissance, j'ai proposé une stratégie d'analyse visant à identifier des interactions gène-gène et intégrant un filtrage statistique des résultats d'une étude d'association pan-génomique suivi par un filtrage basé sur les connaissances contenues dans la littérature en utilisant la fouille de textes. Le détail de cette stratégie, les résultats des analyses et leur interprétation biologique sont résumés dans le chapitre III et présentés dans l'article publié dans *Journal of Allergy and Clinical Immunology.*

Le deuxième volet de ma thèse, décrit dans le chapitre IV, avait pour objectif d'identifier des gènes influençant le délai de survenue de l'asthme et en particulier, les gènes pouvant interagir avec l'exposition au tabagisme passif pendant la petite enfance dans le délai de survenue de l'asthme dans l'enfance. Ce projet fait suite àune méta-analyse d'associations pangénomiques du délai de survenue de l'asthme, réalisée dans neuf études du consortium GABRIEL (5 462 asthmatiques et 8 424 témoins). Cette étude a permis d'identifier une nouvelle région du génome influençant le délai de survenue de l'asthme et de confirmer quatre autres régions précédemment trouvées associées à l'asthme. Les principaux résultats sont résumés dans le chapitre IV. Cette étude à laquelle j'ai contribué a été publiée dans le *Journal of Allergy and Clinical Immunology* (disponible en annexes).

Parmi les neuf études de l'étude précédente, cinq d'entre elles comportaient des données sur l'exposition au tabagisme passif pendant la petite enfance (ELTS). Mon travail a consisté à mettre en évidence des gènes interagissant avec l'ELTS dans le délai de survenue de l'asthme dans l'enfance par une méta-analyse des cinq études d'interactions gène-environnement sur l'ensemble du génome en utilisant des modèles de survie. Des annotations fonctionnelles des SNPs identifiés et une recherche extensive de la littérature ont été conduites. Ceci a été complété par une analyse de pathways biologiques par la méthode GSEA [59] Le détail de la stratégie utilisée, les résultats de ces analyses et leur interprétation biologique sont résumés dans le chapitre IV et présentés dans l'article soumis à *Clinical & Experimental Allergy*

La discussion et les perspectives de ce travail de thèse sont présentées dans le chapitre V.

CHAPITRE II – DONNEES PHENOTYPIQUES ET GENOTYPIQUES DES ETUDES PANGENOMIQUES ANALYSEES

Les données utilisées dans le cadre de cette thèse sont issues du **consortium international GABRIEL** [183]. C'est un consortium multidisciplinaire ayant pour but d'identifier les différentes causes génétiques et environnementales de l'asthme, impliquant plus de 150 scientifiques de 14 pays européens. L'ensemble des données issues de l'analyse pangénomique GABRIEL regroupe 23 études (études de cohortes, cas-témoins et familiales) incluant 26 475 personnes génotypées dont 10 365 asthmatiques de **différents types d'asthme** (asthme de l'enfant, de l'adulte, asthme sévère et professionnel) et 16 110 non asthmatiques après contrôle qualité des données génotypiques. Tous les participants étaient d'origine européenne. L'asthme de l'enfant a été défini comme la présence de la maladie chez les individus avant l'âge de 16 ans, l'asthme de l'adulte comme l'apparition des premiers symptômes après 16 ans. Tous les participants ou leurs parents ont fourni un consentement éclairé signé pour leur participation à l'étude, conformément aux règles des comités d'éthique locaux.

Les données utilisées dans le cadre de ma thèse, correspondent aux individus des études qui comportaient les informations sur les phénotypes étudiés : au total, six études incluses dans le consortium GABRIEL ont contribué aux différents volets de ma thèse. Voici une description des données utilisées pour les deux volets de ma thèse.

1. Description des études pour identifier des interactions gène-gène dans l'atopie

Le premier volet de ma thèse portait sur l'analyse de facteurs de risque génétiques de l'atopie, définie *stricto sensus* comme la réponse cutanée positive à au moins un aéroallergène, dans des familles recensées par un ou deux asthmatiques. Trois études familiales de GABRIEL pour

lesquelles était disponible l'information de l'atopie selon cette définition, ont pu être utilisées pour ces analyses. Les données de l'étude EGEA (L'étude Epidémiologique des facteurs Génétiques et Environnementaux de l'Asthme, l'hyperréactivité bronchique et l'atopie) ont été utilisées comme échantillon de découverte. Les données des études SLSJ (La collection familiale asthmatique du Saguenay-Lac-Saint-Jean) et MRC (*The Medical Research Council UK National family collection*) ont été utilisées en tant qu'échantillons de réplication. Après le QC des données génotypiques, 1 660 individus de l'échantillon EGEA ont été analysés, dont 925 sujets atopiques et 735 non atopiques. Dans SLSJ, l'échantillon analysé incluait 1 138 individus dont 641 atopiques et 497 non atopiques. L'échantillon MRC comprenait 446 sujets dont 106 atopiques et 340 non atopiques.

EGEA

L'étude EGEA [214] est une étude française multicentrique et multidisciplinaire, regroupant cliniciens, épidémiologistes et généticiens. Cette étude combine une étude cas-témoins et une étude familiale, avec au total 2120 sujets. La première enquête transversale a été réalisée entre 1991 et 1995 (EGEA1), suivie de deux enquêtes à douze ans (2003 -2007 ; EGEA2) et à 20 ans (2011-2013 ; EGEA3) (voir https://egeanet.vjf.inserm.fr/index.php/fr/ pour plus de détails).

Le recrutement des sujets participants à l'enquête EGEA a été effectué en deux phases notées EGEAI et EGEAII. Les proposants (asthmatiques et témoins), âgés de 7 à 70 ans, ont été recrutés par un protocole standardisé. Tous les proposants devaient être nés en France ainsi que leurs deux parents et habiter dans une zone prédéfinie des villes participant à l'étude. Les cas asthmatiques ont été recrutés à l'aide d'auto-questionnaires distribués dans les consultations de pneumologie et d'allergologie de différents hôpitaux. Les cas (sujets asthmatiques) de l'étude cas-témoins ont été recrutés dans les services de pneumologie adulte et pédiatrique de six centres cliniques situés dans cinq villes : Paris, Lyon, Marseille, Grenoble et Montpellier. Les témoins étaient pour l'essentiel recrutés en population générale (donc pouvant être asthmatiques) sur la base de listes électorales et de centre de sécurité sociale pour les adultes, et dans des services de chirurgie pour les enfants. Les critères globaux d'appariement entre cas et témoins étaient le centre, le mois d'enquête (en raison de l'exposition aux allergènes), l'âge et le sexe. Dans la première partie de l'enquête (EGEAI), 348 familles recrutées à partir d'au moins un proposant asthmatique (enfants ou adultes) et 415 témoins ont été inclus dans l'étude. Afin d'enrichir la population EGEA en asthmatiques,

40 familles supplémentaires avec au moins deux germains asthmatiques (189 sujets) ont été recensées (EGEAII) (**Figure 2.1**). Le nombre total de sujets était de 2 047 à EGEA1.

L'asthme chez les proposants a été défini par une réponse positive aux quatre questions suivantes : 1) « Avez-vous déjà eu des crises d'essoufflement au repos avec des sifflements dans la poitrine ? », 2) « Avez-vous déjà eu des crises d'asthme ? » et si oui, 2a) « Ce diagnostic a-t-il été confirmé par un médecin ? », et 2b) « Avez-vous eu une crise d'asthme dans les douze derniers mois ? ». La définition de **l'asthme pour les sujets autres que les proposants** était une réponse positive à au moins une des deux questions suivantes : « Avez-vous déjà eu des crises d'essoufflement au repos avec des sifflements dans la poitrine ? », et « Avez-vous déjà eu des crises d'asthme ? ». Cette définition est moins spécifique que celle des proposants mais correspond à la définition épidémiologique internationale de l'asthme, selon les questionnaires du BMRC (*British Medical Research Council)* / CECA (Communauté Européenne du Charbon et de l'Acier).Lors de la deuxième enquête (EGEA2), de nouveaux membres des familles asthmatiques ont été inclus. Au total, 2 120 sujets ont été examinés dans le cadre d'EGEA. Après inclusion, l'ensemble des sujets a été examiné à l'hôpital selon un protocole standardisé basé sur des outils internationaux. Les sujets ont répondu à un questionnaire détaillé sur les symptômes respiratoires et allergiques, les facteurs environnementaux personnels, familiaux et professionnels durant l'enfance et la vie adulte. Différents examens ont été réalisés à EGEA1 et EGEA2 : une exploration fonctionnelle respiratoire (avec épreuve d'hyperréactivité bronchique à la métacholine ou test de bronchodilatation), des tests cutanés à des allergènes, une prise de sang pour le dosage des IgE totales, un multiRAST et une numération formule sanguine.

L'atopie a été évaluée en utilisant des tests cutanés. Un test cutané (**SPT** pour *skin prick test*) positif a été défini comme un diamètre de la papule égal ou supérieur à 3mm par rapport au contrôle négatif (diluant), pour au moins un des onze aéroallergènes testés à EGEA1 : acariens, chat, cafard, fléole des près, olivier, bouleau, *Parietaria judaica*, ambroisie, *Cladosporium herbarum*, *Aspergillus* et *Alternaria tenuis*. Les sujets n'ayant aucun test positif ont été considérés comme non atopiques.

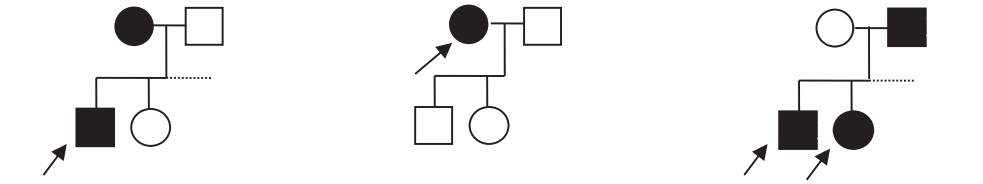


Figure 2.1 : Représentation du mode de recensement des familles de l'étude EGEA (proposant indiqué par une flèche, rond ou carré noir correspondant au statut asthmatique).

SLSJ

La collection familiale asthmatique du Saguenay-Lac-Saint-Jean (SLSJ) regroupe 253 familles franco-canadiennes multigénérationnelles provenant de la région du Québec du Saguenay-Lac-Saint-Jean. Ces familles ont été recrutées par deux proposants asthmatiques entre 1997 et 2002 [215]. Les proposants étaient inclus s'ils remplissaient au moins deux des critères suivants : 1) un minimum de trois visites à la clinique pour de l'asthme aigu dans un délai d'un an ; 2) deux ou plusieurs hospitalisations liées à l'asthme dans un délai d'un an ; ou 3) une dépendance aux stéroïdes, définie par six mois d'utilisation orale ou par une année d'inhalation de corticostéroïdes. Les familles ont été incluses dans l'étude si au moins un parent était disponible pour une évaluation phénotypique, si au moins un parent n'était pas atteint d'asthme, et si les quatre grands-parents étaient d'origine franco-canadienne. Les membres de la famille ont été considérés comme asthmatiques si une histoire d'asthme auto-déclarée et une histoire d'asthme diagnostiquée étaient disponibles, ou par évaluation clinique après un test positif de provocation à la méthacholine (inférieur ou égal à 8 mg/ml de méthacholine) au recrutement.

Les individus étaient considérés comme atopiques s'ils présentaient au moins un test cutané positif à un des 24 aéroallergènes testés, défini comme un diamètre de la papule égal ou supérieur à 3mms. Les allergènes testés étaient répartis en six catégories : 1) chat, chien, squames de cheval et de bœuf, et plumes d'oiseaux ; 2) poussières ; 3) *Dermatophagoides farinae* et *Dermatophagoides pterionisus* (acariens); 4) herbes, mauvaises herbes, ambroisie, fléole des près, ivraie ; 5) mélange d'arbres, bouleau, érable, chêne, orme ; 6) *Chladosporium, Hormodendrum, Alternaria alternata, Alternaria tenuis, Aspergillus*, et *Penicillium* (moisissures). Les sujets négatifs à tous les tests étaient considérés comme non atopiques.

MRC-UK

L'échantillon MRC utilisé comme échantillon de réplication était issu de collections financées par *The Medical Research Council* (MRC) *UK National family collection*. La première collection (MRCA - *The Medical Research Council on Asthma*) incluait 207 familles nucléaires recrutées à partir d'au moins un enfant avec asthmatique avec de l'asthme sévère [180]. Les frères et sœurs des proposants ainsi que leurs parents ont été inclus indépendamment de leur statut asthmatique. Les enfants et leurs parents ont répondu à un questionnaire standardisé (basé sur les questionnaires de l'ATS et l'étude internationale de l'asthme et des allergies dans l'enfance (ISAAC)) au cours d'un entretien réalisé par une infirmière ou un médecin.

L'atopie a été définie par des tests cutanés à cinq aéroallergènes communs : acariens, chat, fléole des près, *Cladosporium herbarum* et *Aspergillus*. Les sujets négatifs à tous les tests ont été considérés comme non atopiques.

La collection MRCE qui incluait des familles nucléaires recrutées à partir de proposants avec de l'eczéma, a également été utilisée afin d'enrichir l'échantillon MRCA en sujets non atopiques. Cette collection ne comportait pas l'information sur l'atopie. Toutefois, seuls les individus MRCE non asthmatiques, sans eczéma, et avec un niveau faible d'Immunoglobulines E ont été inclus dans le projet. Il n'y avait pas de différence en fonction de l'âge et du sexe entre les échantillons MRCA et MRCE.

2. Description des études pour identifier des interactions gène-exposition au tabac pendant la petite enfance dans le délai de survenue de l'asthme

Le deuxième volet de ma thèse avait pour objectif d'identifier des gènes susceptibles d'interagir avec l'exposition au tabagisme passif durant la petite enfance sur le risque de survenue de l'asthme de l'enfant. Cinq études indépendantes de GABRIEL ont été incluses dans ce projet : deux études familiales – EGEA et SLSJ ; deux études en population générale – *the European Community Respiratory Health Survey* (ECRHS) et *The GABRIEL Advanced surveys* (GABRIELA) ; et une cohorte de naissance – *The Avon Longitudinal Study of Parents and Children* (ALSPAC). Toutes ces études comprenaient l'information sur l'âge de début de l'asthme, l'âge au dernier examen, et l'exposition au tabagisme passif pendant l'enfance et/ou *in utero*. Les analyses dans le cadre de ce

projet ont été effectuées en utilisant des modèles de survies : l'asthme a été défini en prenant en compte son âge de début. Dans le cadre de ce projet, je me suis focalisé sur l'asthme apparaissant avant l'âge de 16 ans. Deux études ne comportaient que des asthmatiques avec un asthme de l'enfant : ALSPAC et GABRIELA. Les individus asthmatiques des trois autres études, avec un âge de début de l'asthme après 16 ans, ont été exclus des analyses. Après les étapes de contrôle qualité, 8 273 sujets ont été analysés, dont 2 874 asthmatiques et 5 399 non-asthmatiques. La répartition du nombre de sujets par étude est indiquée le tableau 2.2.

J'ai effectué les analyses d'association pangénomiques de manière centralisée à l'unité INSERM UMR-946 à partir des données individuelles disponibles pour quatre études tandis que les analyses sur les données d'ALSPAC ont été effectuées à Bistrol (UK) et les résultats de ces analyse nous ont ensuite été fournis.

ALSPAC

The Avon Longitudinal Study of Parents and Children (ALSPAC) (www.bristol.ac.uk/alspac) est une cohorte de naissance prospective, longitudinale, basée sur la population générale, comprenant à l'origine 14 541 mères et leurs enfants recrutés dans le comté d'Avon (Royaume-Uni) [216].

Les femmes ont été recrutées dans des cliniques prénatales pendant leur grossesse, avec des dates estimées d'accouchement entre le 1er avril 1991 et le 31 décembre 1992 [217]. Parmi les 14 541 grossesses recrutées, il y a eu 14 072 naissances effectives et 13 988 enfants étaient en vie à l'âge d'un an. Les enfants ont été suivis à partir de la naissance à l'aide d'une combinaison d'auto-questionnaires envoyés à intervalles réguliers à leurs mères et d'évaluations cliniques annuelles réalisées à partir de l'âge de sept ans dans des cliniques de recherche spécialisées. Un total de 5 231 enfants avait des données sur l'asthme et de l'ADN disponible. **Les asthmatiques** ont été définis par au moins une réponse positive aux questions : « Est-ce que un docteur vous a déjà dit que votre enfant a de l'asthme ? » dans un questionnaire envoyé à leurs mères 91 mois après la naissance, et « Est-ce que votre enfant a eu une crise d'asthme lors des douze derniers mois ? » posée à 103, 128, 157 et 166 mois. Les non-asthmatiques étaient ceux qui avaient toujours répondu non à ces enquêtes. Chez les asthmatiques, **l'âge de début de l'asthme** a été défini par la première déclaration de sifflements ; les sifflements ayant été définis par une réponse positive à la question : « Votre enfant a-t-il eu des sifflements, des difficultés respiratoires ou des épisodes d'essoufflements dans les 12 derniers mois ou depuis son âge au dernier questionnaire ? ». Chez

les non-asthmatiques, nous avons considéré l'âge au dernier examen sans aucune visite manquante au préalable. Ainsi, pour les non-asthmatiques avec des rapports négatifs complets, nous avons considéré l'âge au dernier examen et pour les non-asthmatiques avec un rapport négatif incomplet ou discontinu, nous avons considéré l'âge à la dernière visite avant la première visite manquante. Nous n'avons pas inclus dans notre analyse les non-asthmatiques ayant rapporté des épisodes de sifflements avant l'âge de six ans.

L'exposition au tabagisme passif dans la petite enfance a été définie par : 1) un tabagisme actif maternel pendant la grossesse (variable dérivée de questions posées à chaque trimestre de la grossesse) ou 2) une exposition post-natale à la fumée, dans un questionnaire posé huit mois après la grossesse. Avec cette définition, l'exposition avait une forte probabilité d'avoir eu lieu avant l'apparition de l'asthme (qui était en moyenne à 3,1 ans).

ECRHS

L'étude ECRHS est une étude multicentrique, pan-européenne, basée sur la population générale et incluant de jeunes adultes avec un suivi à 8 ans et à 11 ans (ECRHS I : 1991-1993, ECRHS II : 1999-2002 et ECRHS III : 2010-2015) [218,219]. Dans chaque centre (N=56, 25 pays), un échantillon représentatif d'environ 150 000 adultes âgés de 20 à 44 ans a été invité à remplir un bref questionnaire postal sur les symptômes respiratoires (ECRHS I – Etape 1) entre 1991-1993. Un échantillon aléatoire (600 sujets par centre dont 300 hommes, 26 000 individus) a bénéficié d'un complément d'enquête intensif (ECRHS I – Étape 2 – échantillon aléatoire). Les participants qui avaient des symptômes très évocateurs d'asthme, mais qui n'avaient pas été choisis au hasard pour participer à l'étape 2 ont également été invités à bénéficier de recherches approfondies (ECRHS I – Phase 2 – échantillon enrichi, environ 150 adultes par centre). Huit ans plus tard, tous les adultes qui avaient pris part à l'étape 2 ont été contactés (ECRHS II) et ont de nouveau été questionnés sur les symptômes respiratoires (plus de 10 000 individus).

L'échantillon inclus dans le consortium GABRIEL est un sous-échantillon d'ECRHS. **Les cas asthmatiques** étaient les participants de l'échantillon aléatoire ou enrichi qui avaient répondu positivement à la question : « Avez-vous déjà eu de l'asthme » à ECRHS I Etape 2 ou à ECRHS II. Les témoins étaient un échantillon aléatoire (de l'échantillon aléatoire) qui avaient répondu négativement à la même question aux deux enquêtes. Au total, 16 centres (8 pays, 2 210 individus) ont contribué à la GWAS de l'asthme de GABRIEL [218,219].

L'échantillon analysé pour notre projet était basé sur les deux premières enquêtes (ECRHS I et II). Les participants inclus dans cette analyse provenaient de l'échantillon de cas-témoins nichés dans la cohorte et génotypés dans GABRIEL. Pour les individus ayant développé un asthme, l'information sur **l'âge de début de l'asthme** a été obtenue à partir de l'âge à la première crise d'asthme rapporté à ECRHS I ou II. Afin de considérer uniquement **l'asthme dans l'enfance**, les asthmatiques avec un âge de début de l'asthme après 16 ans ont été exclus. Pour les individus n'ayant pas développé un asthme, nous avons considéré l'âge au dernier suivi. **L'exposition au tabagisme passif dans l'enfance** a été définie à partir des réponses aux questions suivantes : « Votre père a-t-il déjà fumé régulièrement pendant votre enfance ? », «Votre mère a-t-elle déjà fumé régulièrement pendant votre enfance ou avant votre naissance ? » et/ou basé sur l'exposition *in utero* à la fumée de tabac, définie comme le tabagisme actif de la mère à n'importe quel moment de la grossesse.

EGEA

L'étude EGEA a été décrite dans le cadre du projet 1. Dans le second projet de thèse, les données EGEA1, EGEA2 et EGEA3 ont été utilisées. Pour les individus ayant développé un asthme, l'information sur **l'âge de début de l'asthme** a été obtenue à partir des réponses des adultes asthmatiques ou des parents d'enfants asthmatiques aux questions suivantes « Quel âge aviez-vous quand vous avez eu votre première crise d'asthme ? » ou « Quel âge avez votre enfant quand il a eu sa première crise d'asthme ? ». Pour les individus non asthmatiques, nous avons considéré l'âge au dernier suivi.

L'exposition au tabagisme passif dans la petite enfance a été définie pour les sujets adultes au moment de l'étude par la réponse positive à la question : « Est-ce que votre mère ou votre père fumait pendant votre petite enfance ? », et/ou « Est-ce que votre mère fumait quand elle était enceinte de vous ? ». Chez un enfant, la définition de l'exposition a été définie à partir de la question posée aux parents : « Est-ce que vous ou votre conjoint fumait quand votre enfant avait moins de deux ans ? » et/ou « Avez-vous fumé quand vous étiez enceinte ? ».

GABRIELA

The GABRIEL Advanced surveys (GABRIELA) est une enquête démographique transversale menée dans des zones rurales d'Allemagne, d'Autriche, de Suisse et de Pologne à l'automne-hiver 2006 et au printemps-été 2007 [220]. Le but de cette étude était de rechercher les causes

environnementales de l'asthme et de l'atopie [221]. Au total, 135 359 enfants âgés de 6 à 12 ans ont été recrutés dans des écoles. Dans une première étape à l'automne-hiver 2006, l'asthme, les maladies allergiques, et l'exposition à des environnements agricoles ont été évalués à l'aide d'un questionnaire court parental (N=79 888). Dans une deuxième étape au printemps-été 2007, 9 668 enfants ont été sélectionnés parmi les familles ayant consenti à des prélèvements sanguins, des tests génétiques et des recueils d'échantillons environnementaux, par échantillonnage aléatoire stratifié afin d'assurer la représentation d'enfants fortement exposés à des environnements agricoles. L'ADN génomique et les données issues des questionnaires étaient disponibles pour 862 enfants asthmatiques et 865 enfants non-asthmatiques (témoins). Un individu a été considéré comme **asthmatique** à partir d'au moins un rapport parental de diagnostic d'asthme effectué par un médecin, ou au moins deux diagnostics de bronchite asthmatique au cours de la vie. Dans le cas contraire, l'individu était considéré comme un témoin. **L'âge de début de l'asthme** a été défini à partir de la question posée aux parents de sujets asthmatiques : « Quel âge avait votre enfant quand les premiers symptômes de sifflements dans la poitrine ont débuté ? à l'âge de … ans, si pendant la première année : à l'âge de … mois ». Chez les non-asthmatiques, l'âge à l'examen a été considéré. Pour prendre en compte le mode d'échantillonnage des sujets, des poids de probabilité ont été introduits dans les analyses statistiques.

L'exposition au tabagisme passif dans l'enfance a été définie à partir de l'exposition au tabagisme passif à n'importe quel moment durant la grossesse. Nous n'avons pas utilisé l'information de l'exposition des enfants au moment de l'enquête (c'est-à-dire le tabagisme actuel par au moins un des deux parents) car celle-ci avait lieu à l'âge de neuf ans, soit après l'âge moyen de début d'asthme des sujets de l'étude (qui était de 2,9 ans). Cependant, 80% des enfants qui étaient exposés à la fumée de tabac *in utero* étaient également exposés au tabagisme passif au moment de l'enquête.

SLSJ

L'étude SLSJ a été décrite dans le cadre du projet 1. Pour l'analyse du deuxième volet de ma thèse, **l'asthme et son âge de début** ont été obtenus à partir des réponses aux questions suivantes « Avez-vous déjà eu des crises d'asthme ? Quel âge aviez-vous quand vous avez eu cette crise ? ». Quand l'âge de début était inférieur à deux ans (41 cas), nous avons considéré l'âge de début à deux ans

pour éviter les incertitudes. Pour les individus n'ayant pas développé d'asthme, nous avons considéré l'âge à l'examen.

L'exposition au tabagisme passif dans la petite enfance a été définie par une exposition passive à la fumée de tabac avant l'âge de deux ans chez les enfants (≤ 16 ans) ou avant cinq ans chez les adultes (> 16 ans).

3. Données génotypiques

Tous les échantillons, excepté MRC, ont été **génotypés** dans le cadre du consortium GABRIEL à l'aide d'une puce Illumina Human610-Quad (EGEA, SLSJ, ECRHS et GABRIELA) ou d'une puce Illumina Human550-Quad (ALSPAC). Les individus de l'échantillon MRC ont été génotypés à l'aide des puces Illumina Sentrix Human-1 et Sentrix HumanHap300 dans le cadre de la première analyse d'association pangénomique de l'asthme [180].

La fiabilité des données de génotypage ont été assurées dans chaque échantillon par des procédures de **contrôle qualité (QC)**, suivant un protocole similaire (voir détail **Tables 2.1 et 2.2, parties génotypage et QC**). Le QC pour MRC avait été fait précédemment et a été décrit dans Moffat *et al.* [180]. D'abord, un contrôle qualité des sujets a été réalisé afin d'exclure les sujets ayant posé des problèmes de génotypage du fait par exemple d'une mauvaise qualité de leur ADN, ou parce qu'ils n'étaient pas issus de la même population ethnique que la majorité des sujets de l'étude. Les individus ont été exclus si :

- ils n'étaient pas d'origine européenne (déterminé à partir d'une analyse en composante principale (ACP) sur chaque échantillon et à l'aide de populations de références issues du consortium international HapMap [222]),
- ils présentaient un pourcentage de données génotypiques manquantes trop important (call rate < 97%),
- ils s'écartaient du taux d'hétérozygotie moyen calculé pour les autosomes et le chromosome X,
- ils présentaient une incohérence entre le sexe rapporté dans les données et le sexe déduit des données génotypiques,

- dans les études familiales, ils présentaient une incohérence entre le lien de parenté rapporté dans les données au cours du recensement et le lien de parenté déduit des données génotypiques.

La détection des sujets d'origine non européenne a été effectuée à l'aide d'une analyse en composante principale (ACP). L'identification des axes de variation a été réalisé à partir de l'ensemble des données génotypiques (SNPs) en faible déséquilibre de liaison provenant de chaque étude du projet et de différents panels de sujets issus de populations européennes, africaines et asiatiques du consortium international HapMap [222]. Les sujets s'écartant du cluster formé par les sujets européens de HapMap ont été exclus. Après exclusion des sujets non européens, des ACP ont été réalisées sur les sujets sans les données HapMap afin de s'assurer de l'absence de stratification résiduelle dans les échantillons. Un QC des marqueurs génétiques a ensuite été effectué, excluant les SNPs renseignés pour un faible taux de sujets (call rate < 97%), une fréquence de l'allèle mineur très faible (MAF < 1% pour ALSPAC, MAF < 5 % pour les autres échantillons) ou montrant un écart à l'équilibre d'Hardy-Weinberg ($P_{HWE}<10^{-4}$).

Imputations

Dans le cadre de mon premier projet de thèse, des imputations ont été réalisées dans l'échantillon MRC afin d'uniformiser les données génotypiques avec celles des échantillons EGEA et SLSJ. Ces imputations ont été réalisées à l'aide de la population de référence HapMap Phase 2 (release 21) avec le logiciel MACH 1.0. Un total de 501 167 SNPs a été conservé pour les analyses.

Pour mon second projet de thèse, l'ensemble des échantillons a été imputé à l'aide de la population de référence HapMap2, afin d'enrichir les données génotypiques (voir **Table 2.1**). Seul les SNPs avec un score de qualité d'imputation élevé (rsq>0.5) et une fréquence d'allèle mineur supérieure ou égale à 1% ont été conservés. Un total de 2,11 millions de SNPs a été conservé pour l'analyse.

Table 2.1. Données d'analyses du premier projet : Génotypage et QC

	EGEA (N=1 660)	**SLSJ (N=1 138)**	**MRC-UK (N=446)**
Génotypage			
Puce de génotypage	Illumina Human610-Quad	Illumina Human610-Quad	Illumina Sentrix Human-1 et Sentrix HumanHap300
Centre de génotypage	Centre National de Génotypage, Evry, France	Centre National de Génotypage, Evry, France	Centre National de Génotypage, Evry, France
QC des sujets			
Call-rate	97%	97%	97%
Hétérozygotie	Exclusion des sujets si <0.30 ou >0.33	Exclusion des sujets si <0.30 ou >0.33	Exclusion des sujets si <0.30 ou >0.33
Exclusion selon l'origine ethnique	Basé sur l'ACP	Basé sur l'ACP	Basé sur l'ACP
QC des SNPs avant imputations			
Fréquence de l'allèle mineur	5%	5%	5%
Test d'équilibre d'Hardy Weinberg (*P*-valeur)	10^{-4}	10^{-4}	10^{-4}
Call-rate	97%	97%	97%
Imputations			
Software	-	-	MACH 1.0
HapMap	-	-	Hapmap2 r21
Filtre de QC des SNPs	-	-	rsq $\geq$ 0.5 & MAF $\geq$ 5%

Table 2.2. Données d'analyses du second projet : Génotypage, QC et imputations

	EGEA (N=1 498)	SLSJ (N=377)	ECRHS (N=1 685)	GABRIELA (N=1 482)	ALSPAC (N=3 231)
Génotypage					
Puce de génotypage	Illumina Human610-Quad	Illumina Human610-Quad	Illumina Human610-Quad	Illumina Human610-Quad	Illumina HumanHap550Quad
Centre de génotypage	Centre National de Génotypage, Evry, France	Centre National de Génotypage, Evry, France	Centre National de Génotypage, Evry, France	Centre National de Génotypage, Evry, France	23andMe sous-traitant le Wellcome Trust Sanger Institute (Cambridge, UK), et le LabCorp, (Burlington, North Carolina, US)
QC des sujets					
Call-rate	97%	97%	97%	97%	97%
Hétérozygotie	Exclusion des sujets si <0.30 ou >0.33	Exclusion des sujets si <0.30 ou >0.33	Exclusion des sujets si <0.30 ou >0.33	Exclusion des sujets si <0.30 ou >0.33	Exclusion des sujets si <0.320 ou >0.345 (Sanger) et <0.310 ou >0.330 (LabCorp)
Exclusion selon l'origine ethnique	Basé sur l'ACP	Basé sur l'ACP	Basé sur l'ACP	Basé sur l'ACP	Basé sur l'ACP
QC des SNPs avant imputations					
Fréquence de l'allèle mineur	5%	5%	5%	5%	1%
Test d'équilibre d'Hardy Weinberg (*P*-valeur)	10^{-4}	10^{-4}	10^{-4}	10^{-4}	$5x10^{-7}$
Call-rate	97%	97%	97%	97%	95%
Imputations					
Software	MACH 1.0	MACH 1.0	MACH 1.0	MACH 1.0	MACH 1.0
HapMap	Hapmap2 r21	Hapmap2 r21	Hapmap2 r21	Hapmap2 r21	Hapmap2 r22
Filtre de QC des SNPs	rsq ≥ 0.5 & MAF ≥ 1%	rsq ≥ 0.5 & MAF ≥ 1%	rsq ≥ 0.5 & MAF ≥ 1%	rsq ≥ 0.5 & MAF ≥ 1%	rsq ≥ 0.5 & MAF ≥ 1%

CHAPITRE III – Analyses d'interactions gène-gène dans l'atopie

1. Résumé

Les maladies allergiques, dont l'asthme, sont des maladies multifactorielles, résultant de nombreux facteurs génétiques et environnementaux. Ces maladies présentent une importante hétérogénéité et des incertitudes de diagnostic qui peuvent poser des problèmes de puissance dans la détection des facteurs de risque. L'étude de l'atopie, définie par une réaction d'hyperréactivité aux allergènes et associée à l'asthme, peut permettre de palier en partie contourner ce problème. La composante génétique de l'atopie est importante, cependant les loci identifiés par les études d'association pangénomiques simple marqueur (GWAS) n'expliquent qu'une petite partie de la composante génétique du risque de la maladie. Une des limites des GWAS est que ces analyses peuvent manquer de puissance pour détecter des variants ayant un effet individuel faible ou interagissant avec d'autres variants.

L'analyse d'interactions gène-gène (GxG) peut permettre la détection de nouveaux gènes impliqués dans le risque de la maladie. Cependant l'analyse d'interactions GxG au niveau génome entier nécessite la réalisation de très nombreux tests statistiques impliquant une correction pour les tests multiples bien plus importante que les GWAS. L'utilisation de filtrages statistiques et/ou basés sur les connaissances *a priori*, qui permettent de réduire à un sous-groupe pertinent les interactions entre gènes à tester, peut apporter un gain de puissance statistique aux analyses d'interactions GxG.

Dans cette étude, nous avons développé une stratégie d'analyse d'interaction GxG pour identifier de nouveaux gènes impliqués dans l'atopie, définie comme une réponse cutanée positive à au moins un aéroallergène. Dans une première étape, nous avons effectué une GWAS de l'atopie et appliqué un filtre statistique afin de sélectionner les SNPs montrant des associations suggestives lors du test simple marqueur. Les SNPs sélectionnés ont été annotés aux gènes. Au cours d'une deuxième étape de filtre basé sur les connaissances contenues dans la littérature scientifique (résumés de PubMed), nous avons sélectionné les paires de gènes montrant des similarités textuelles par fouille de texte

en utilisant la méthode GRAIL. Enfin, lors d'une troisième étape, nous avons testé les interactions entre les SNPs des paires de gènes sélectionnées par les deux types de filtres. Ces analyses, nous avons utilisé trois études familiales recensées à travers des patients asthmatiques : l'échantillon de découverte EGEA (étude française incluant 925 sujets atopiques et 735 non atopiques) et deux échantillons de réplication, SLSJ (étude de canadiens français incluant 641 atopiques et 497 non atopiques) et MRC (étude anglaise avec 106 atopiques et 340 non atopiques). Ces échantillons, leur génotypage et les différentes étapes de contrôle qualité des données ont été décrites au chapitre II.

Dans la première étape de notre stratégie, nous avons effectué une GWAS en deux phases. En phase 1, nous avons effectué une analyse d'association pangénomique dans l'échantillon de découverte (EGEA). En phase 2, les SNPs montrant des associations suggestives ($P_{EGEA} \leq 10^{-4}$) ont été analysés dans les échantillons de réplications (SLSJ et MRC-UK). Nous avons ensuite appliqué une méta-analyse à effets fixes à ces trois échantillons. L'analyse pangénomique a permis d'identifier un SNP du gène *ADGRV1* (adhesion G protein-coupled receptor V1), au locus 5q14 (P_{Meta} = 6.8 x 10^{-9}), significativement associé à l'atopie au seuil génome-entier de $P = 1.5\text{x}10^{-7}$. Nous avons calculé le niveau de significativité génome entier suivant la méthode de Bonferonni à partir du nombre de tests effectifs indépendants parmi l'ensemble des SNPs analysés (méthode du Meff [17]). Les SNPs ont ensuite été annotés aux gènes selon leur position sur le génome. Nous avons décidé de sélectionner deux groupes de gènes pour former des paires : 1) les gènes à moins de 50kb d'au moins un SNP significatif au seuil génome entier : un seul gène, *ADGRV1* ; 2) les gènes à moins de 50kb des SNPs sélectionnés à la première phase du GWAS ($P_{EGEA} \leq 10^{-4}$), correspondant à 30 gènes.

Dans une deuxième étape de filtre, nous avons pris en compte les connaissances contenues dans la littérature scientifique en utilisant la fouille de textes afin de ne conserver que les paires de gènes montrant des similarités textuelles pour l'analyse d'interaction GxG. Nous avons utilisé la méthode GRAIL [108] qui évalue la similarité entre gènes selon la co-occurrences des mots dans les résumés de PubMed dans lesquels les gènes sont cités. Cette approche a été appliquée à *ADGRV1* et 30 gènes ayant au moins un SNP associé à l'atopie ($P_{EGEA} \leq 10^{-4}$). Trois de ces 30 gènes ont montré une relation avec *ADGRV1* au seuil de $P_{\text{GRAIL}} \leq 0.10$.

Dans une troisième étape, nous avons testé les interactions entre tous les SNPs deux à deux, pour chaque paire de gènes sélectionnée. Cette analyse a été faite en deux phases, comme pour la

GWAS. Les interactions ont d'abord été testées dans l'échantillon de découverte, et les paires de SNPs passant le seuil de $P_{EGEA} \leq 5.10^{-3}$ ont été testées dans les échantillons de réplications, puis une méta-analyse des trois études a été réalisée. Le seuil de significativité a été évalué par la méthode de Bonferonni appliquée au nombre effectif de tests indépendants d'interactions SNPxSNP calculé grâce à une extension de la méthode de Li and Ji [17], ce seuil a été estimé à $P = 7.3x10^{-5}$. L'utilisation des deux filtres a permis de réduire le nombre de tests d'interaction par un facteur 9 par rapport à l'utilisation du filtre statistique seulement. L'analyse des interactions entre les SNPs du gène *ADGRV1* avec les SNPs des 3 gènes (*CHD7*, *DNAH5*, *ATP8B1*) a mis en évidence deux interactions significatives entre SNPs des gènes *ADGRV1* et *DNAH5* (dynein axonemal heavy chain 5) ($P_{GxG} \leq 6.1x10^{-5}$).

Les gènes *ADGRV1* et *DNAH5* sont pertinents sur le plan biologique car ils sont tous les deux impliqués dans des ciliopathies et la mobilité ciliaire. En effet, des mutations du gène *ADGRV1* sont à l'origine du syndrome d'Usher de type IIC, une ciliopathie caractérisée par une perte auditive et une déficience visuelle, la protéine codée par le gène est un composant du réseau de protéines Usher qui fonctionnent dans les stéréocils des cellules ciliées de l'oreille interne et des cils photorécepteurs. Des mutations du gène *DNAH5* provoquent une dyskinésie ciliaire primitive de type 3, une ciliopathie combinant manifestations respiratoires des voies aériennes supérieures et inférieures, une infertilité masculine et un *situs inversus*, la protéine codée par ce gène fait partie d'un complexe protéique associé aux microtubules, responsable de la mobilité des cils, en particulier dans les cellules épithéliales respiratoires, où la mobilité ciliaire est essentielle au transport du mucus et à la clairance des voies respiratoires. Bien que les fonctions respectives des protéines ADGRV1 et DNAH5 aient été initialement décrites dans différents organes, des relations entre des fonctions de ces protéines ont été également rapportées (comme indiqué plus en détail dans notre papier), corroborant nos résultats statistiques. Le mode d'action des deux gènes dans l'atopie est encore inconnu, mais on peut émettre l'hypothèse que les protéines ADGRV1 et DNAH5 sont impliquées dans un dysfonctionnement des cils qui déplacent hors des voies respiratoires le mucus sécrété contenant des particules étrangères emprisonnées, ce qui favorise la sensibilisation allergique. Ceci est corroboré par les observations récentes d'une expression différentielle de l'ARN des deux gènes *ADGRV1* et *DNAH5* dans les expectorations de sujets sensibilisés aux acariens par rapport à des témoins non atopiques. Ces résultats sont en faveur d'un rôle des gènes liés aux fonctions ciliaires dans l'atopie.

De plus, les deux gènes ont déjà été retrouvés associés à des maladies respiratoires et des phénotypes associés. Le gène *DNAH5* a été associé à la capacité pulmonaire totale chez des patients ayant une maladie pulmonaire obstructive chronique, et à la sensibilisation aux pollens. De plus, ce gène appartient à la même famille que le gène *DNAH9* qui a montré une interaction avec l'exposition à la fumée de tabac dans la petite enfance dans le risque d'hyperréactivité bronchique dans EGEA et SLSJ. Par ailleurs, une association suggestive du gène *ADGRV1* avec l'asthme a été rapportée par une méta-analyse de GWAS. Cependant, une analyse stratifiée selon le statut asthmatique a montré que l'interaction entre les SNPs des gènes *ADGRV1* et *DNAH5*, associés à l'atopie, apparait indépendante de l'asthme dans notre étude.

Cette étude montre que la stratégie proposée qui combine une analyse d'association pangénomique et une analyse d'interaction GxG, en utilisant deux filtres statistique et basé sur la connaissance, permet d'identifier avec succès des gènes fortement candidats dans un phénotype complexe comme l'atopie.

2. Article publié dans *Journal of Allergy and Clinical Immunology*

A novel role for ciliary function in atopy: *ADGRV1* and *DNAH5* interactions

Pierre-Emmanuel Sugier, MSc,[a,b] Myriam Brossard, PhD,[a]* Chloé Sarnowski, PhD,[a]* Amaury Vaysse, PhD,[a] Andréanne Morin, MSc,[c,d] Lucile Pain, MSc,[d] Patricia Margaritte-Jeannin, PhD,[a] Marie-Hélène Dizier, PhD,[a] William O. C. M. Cookson, MD, DPhil,[e] Mark Lathrop, PhD,[c] Miriam F. Moffatt, DPhil,[e] Catherine Laprise, PhD,[d]‡ Florence Demenais, MD,[a]‡ and Emmanuelle Bouzigon, MD, PhD[a]‡ *Paris, France; Montreal and Chicoutimi, Quebec, Canada; and London, United Kingdom*

Background: Atopy, an endotype underlying allergic diseases, has a substantial genetic component.
Objective: Our goal was to identify novel genes associated with atopy in asthma-ascertained families.
Methods: We implemented a 3-step analysis strategy in 3 data sets: the Epidemiological Study on the Genetics and Environment of Asthma (EGEA) data set (1660 subjects), the Saguenay-Lac-Saint-Jean study data set (1138 subjects), and the Medical Research Council (MRC) data set (446 subjects). This strategy included a single nucleotide polymorphism (SNP) genome-wide association study (GWAS), the selection of related gene pairs based on statistical filtering of GWAS results, and text-mining filtering using Gene Relationships Across Implicated Loci and SNP-SNP interaction analysis of selected gene pairs.
Results: We identified the 5q14 locus, harboring the adhesion G protein–coupled receptor V1 (*ADGRV1*) gene, which showed genome-wide significant association with atopy (rs4916831, meta-analysis *P* value = 6.8×10^{-9}). Statistical filtering of GWAS results followed by text-mining filtering revealed relationships between *ADGRV1* and 3 genes showing suggestive association with atopy ($P \leq 10^{-4}$). SNP-SNP interaction analysis between *ADGRV1* and these 3 genes showed significant interaction between *ADGRV1* rs17554723 and 2 correlated SNPs (rs2134256 and rs1354187) within the dynein axonemal heavy chain 5 *(DNAH5)* gene ($P_{\text{meta-int}} = 3.6 \times 10^{-5}$ and 6.1×10^{-5}, which met the multiple-testing corrected threshold of 7.3×10^{-5}). Further conditional analysis indicated that rs2134256 alone accounted for the interaction signal with rs17554723.
Conclusion: Because both DNAH5 and ADGRV1 contribute to ciliary function, this study suggests that ciliary dysfunction might represent a novel mechanism underlying atopy. Combining GWAS and epistasis analysis driven by statistical and knowledge-based evidence represents a promising approach for identifying new genes involved in complex traits. (J Allergy Clin Immunol 2018;141:1659-67.)

Key words: *Atopy, asthma, genetics, genome-wide association study, gene-gene interaction, text mining, ADGRV1, DNAH5, ciliary function*

Allergies and asthma are among the most common diseases in industrialized countries. Although environmental factors play an important role in allergic diseases, estimates of the heritability of allergy, which range between 25% and 80%, suggest a significant genetic contribution.[1] Genome-wide association studies (GWASs) have identified a number of loci associated with allergic diseases (ie, asthma, atopic dermatitis, and rhinitis),[2,3] but these loci only explain a small part of the genetic risk. Part of the difficulty encountered in identifying the genetic factors involved in these allergic diseases is due to the heterogeneity of these diseases and the uncertainty of diagnosis. However, this problem can be alleviated by the study of an endotype underlying allergic diseases, such as allergic sensitization or atopy.

Atopy is characterized by production of allergen-specific IgE against environmental allergens. Estimates of heritability of atopy range from 40% to 85%.[4,5] Many candidate genetic studies of atopy have been conducted but have often led to inconsistent results.[6] Although the first GWAS of allergic sensitization reported only a few loci,[7-10] 2 recent large-scale meta-analyses of allergic sensitization[11] and self-reported allergy[12] increased the number of associated loci to 10 and 16, respectively. However, other loci can influence atopy because it is well known that a GWAS alone cannot reveal the whole genetic landscape underlying complex phenotypes.

Heterogeneity across studies, which can be caused by variability in the genetic background of the populations, environmental exposures, or study design, might be a limitation of meta-analyses of GWASs for identifying new loci associated

From [a]the Genetic Variation and Human Diseases Unit, INSERM, Université Paris Diderot, Université Sorbonne Paris Cité, Paris; [b]Université Pierre et Marie Curie, Paris; [c]McGill University and Génome Québec Innovation Centre, Montreal; [d]Département des Sciences Fondamentales, Université du Québec à Chicoutimi, Chicoutimi; and [e]the Section of Genomic Medicine, National Heart and Lung Institute, London.

*These authors contributed equally to this work as second authors.

‡These authors contributed equally to this work as senior authors.

Supported by the French National Agency for Research (ANR-11-BSV1-027-GWIS-AM; ANR-USPC-2013-EDAGWAS), Université Pierre et Marie Curie doctoral fellowship, and the Canada Research Chair (held by C.L.), and support from the Canadian Institutes of Health Research (CIHR) enabled the maintenance and continuation of the SLSJ asthma study. C.L. is the director of the Asthma Strategic Group of the Respiratory Health Network of the Fonds de la recherche en santé du Québec (FRSQ) and a member of Allergen Network. Genotyping was supported by a grant from the European Commission (no. LSHB-CT-2006-018996-GABRIEL).

Disclosure of potential conflict of interest: The authors declare that they have no relevant conflicts of interest.

Received for publication September 9, 2016; revised May 30, 2017; accepted for publication June 21, 2017.

Available online September 18, 2017.

Corresponding author: Florence Demenais, MD, UMR-946, INSERM, Université Paris-Diderot, 27 rue Juliette Dodu, 75010 Paris, France. E-mail: florence.demenais@inserm.fr.

The CrossMark symbol notifies online readers when updates have been made to the article such as errata or minor corrections

0091-6749/$36.00

http://dx.doi.org/10.1016/j.jaci.2017.06.050

Abbreviations used
ADGRV1: Adhesion G protein–coupled receptor V1
DNAH5: Dynein axonemal heavy chain 5
EGEA: Epidemiological Study on the Genetics and Environment of Asthma
GRAIL: Gene Relationships Across Implicated Loci
GWAS: Genome-wide association study
LD: Linkage disequilibrium
MRC: Medical Research Council
MRCA: Medical Research Council–funded collection of nuclear families with Asthma
MRCE: Medical Research Council–funded collection of nuclear families with Eczema
OR: Odds ratio
QC: Quality control
SLSJ: Saguenay-Lac-Saint-Jean study
SNP: Single nucleotide polymorphism
SPT: Skin prick test

with a trait. Notably, the importance of data sampling was highlighted recently by a positional cloning study of eczema in which association with *ANO3/MUC15* genetic variants was only found in family samples ascertained through asthmatic patients but not in families ascertained through patients with eczema or in a case-control study of eczema.[13]

Another limitation of GWASs is that they typically focus on analysis of individual single nucleotide polymorphisms (SNPs) and are underpowered to detect genetic factors, which have a small marginal effect but rather interact with each other. Gene-gene interaction analysis (or epistasis analysis) has the ability to reveal novel genes involved in complex traits but raises an enormous multiple-testing problem when performed at the genome-wide level. Statistical and biological filtering pipelines can be used to limit the search for SNP-SNP interactions.[14] Following the "guilt-by-association" assumption, which states that connected genes participate usually in the same or related cellular functions,[15] search for interactions can be restricted to genes pointed out by a preliminary GWAS (eg, interactions of genes harboring significant association signals with genes harboring suggestive associations) and showing relationships based on prior knowledge. One knowledge-based approach that can be particularly useful to prioritize genes for epistasis analysis is text mining of the literature because it can highlight relationships between genes[16] according to their co-occurrence with the same words in scientific articles.

The objective of this study was to identify novel genetic factors influencing atopy by combining a GWAS and epistasis analysis driven by statistical and knowledge-based evidence in 3 family samples ascertained through asthmatic patients: the French Epidemiological Study on the Genetics and Environment of Asthma (EGEA; 1660 subjects), the French-Canadian Saguenay-Lac-Saint-Jean (SLSJ) study (1138 subjects), and the Medical Research Council (MRC) UK study (446 subjects). Our overall analysis strategy included 3 main steps: (1) a genome-wide single-SNP association analysis, (2) the selection of related gene pairs based on statistical filtering from GWAS results and text-mining filtering using the Gene Relationships Across Implicated Loci (GRAIL) approach,[17] and (3) an SNP-SNP interaction analysis for the selected gene pairs.

METHODS

Study data sets and definition of atopy

The EGEA study combines a case-control and family-based study of asthma. The whole study population includes 388 families ascertained through at least 1 asthmatic proband recruited in chest clinics (1705 probands and first-degree relatives) plus 415 population-based control subjects (total of 2120 subjects). All subjects were born in France and were of European ancestry. The protocol of this study has been described elsewhere.[18-20] Atopy was assessed by using skin prick tests (SPT) performed in 1978 subjects. A positive SPT response was defined as a wheal diameter of 3 mm or greater to at least 1 of 11 aeroallergens belonging to 3 groups (indoor allergens, outdoor allergens, and molds). After quality control (QC) of genotypic data, 925 atopic and 735 nonatopic subjects were included in the analysis.

The Saguenay-Lac-Saint-Jean and Quebec City Familial Asthma Collection (SLSJ) consists of a French-Canadian founder population panel of 253 multigenerational families from the Saguenay-Lac-Saint-Jean region ascertained through 2 asthmatic probands.[21] This study has been described elsewhere.[21] Skin tests were done in 1195 SLSJ subjects, and atopy was defined similarly as in EGEA. After QC of genotypic data, the analysis data set included 641 atopic and 497 nonatopic subjects.

The Medical Research Council–funded collection of nuclear families with Asthma (MRCA) UK study includes 207 nuclear families recruited through at least 1 proband with childhood-onset asthma. The study protocol has been described elsewhere.[22] Atopy was defined similarly as in EGEA. To increase the number of unaffected subjects (control subjects), we included subjects from the Medical Research Council–funded collection of nuclear families with Eczema (MRCE) UK data set that were recruited through probands with eczema. Only subjects without asthma, without eczema, and with low IgE levels were used as control subjects in this study. We checked that age and sex distributions were similar in MRCA and MRCE samples. After QC of genotypic data, the analysis sample included 106 atopic and 340 nonatopic subjects. The whole UK sample will be subsequently designated as the MRC sample.

Protocols of EGEA, the SLSJ study, and MRC studies have been approved by local ethical committees. All adult participants and children's legal guardians provided written informed consent.

Genotyping

Both the EGEA and SLSJ data sets were genotyped with the Illumina 610-Quad Array (Illumina, San Diego, Calif) as part as of the GABRIEL asthma consortium GWAS.[23] Stringent quality criteria were applied to select both patients and SNPs, and these have been detailed previously.[23,24] After QC, there was a final set of 501,167 autosomal SNPs for analysis. Offspring in MRCA families and MRCE control subjects were genotyped by using the Illumina Sentrix HumanHap300 BeadChip (307,981 autosomal SNPs) as part of the first asthma GWAS.[22,23] QC for MRC samples has been detailed elsewhere.[22,25] SNP imputation was performed by using MACH v1.00 software[26] and HapMap2 release 21 CEU haplotypes as a reference panel to make the number of SNPs in MRC sample as large as the number in EGEA and SLSJ samples. Imputed SNPs were kept for analysis if their imputation quality score (rsq)[27] was 0.5 or greater and minor allele frequency was 5% or greater.

Descriptive statistics and analysis strategy

Descriptive statistics of atopy together with sex, age, and asthma status were assessed in each data set by using Stata software (version 14.1; distributed by Stata, College Station, Tex). The workflow of our 3-step analysis strategy is summarized in Fig 1 and presented in the following paragraphs.

Genome-wide single-SNP analysis

We performed a 2-stage GWAS. In the first stage association analysis between individual SNPs and atopy was carried out in the EGEA data set. This analysis was based on a logistic regression model assuming an additive model

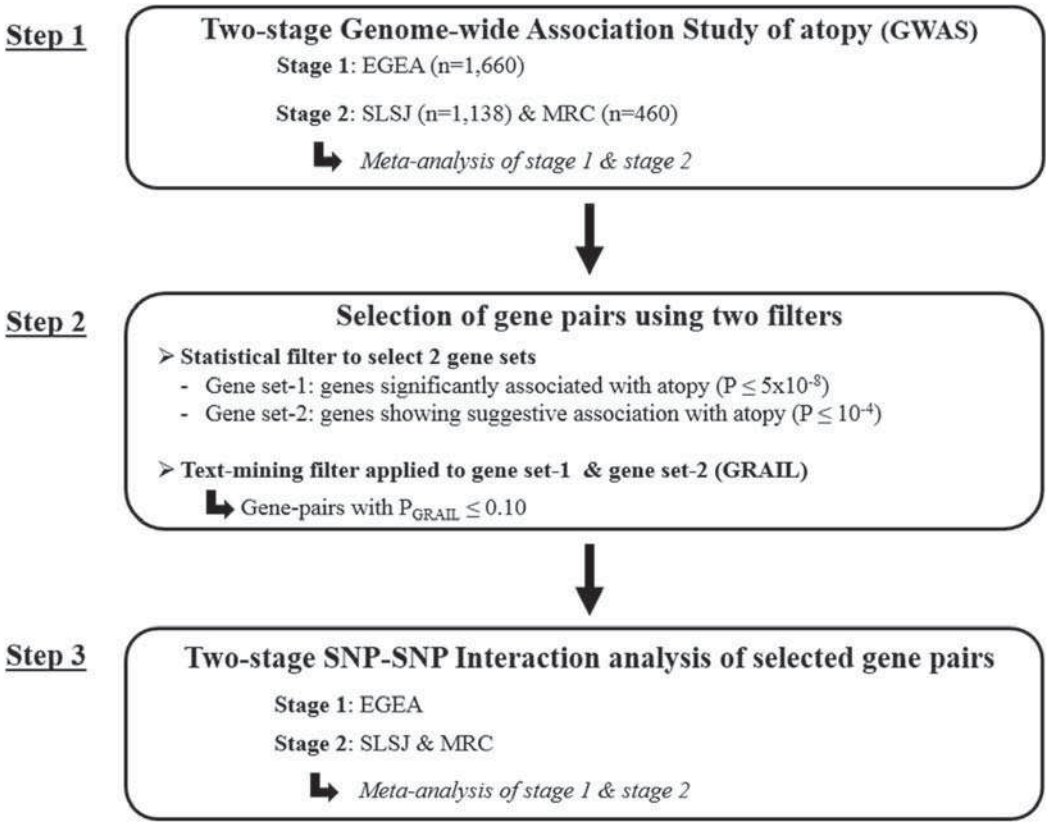


FIG 1. Three-step analysis strategy.

for SNP effect by using Stata software (version 14.1). This model was adjusted for significant effects of age and sex and 2 principal components to account for population structure. We took into account familial dependencies using the cluster and robust options of the logit function in Stata software. Test of SNP effect was based on a Wald test. In a second stage the SNPs reaching a *P* value of 10^{-4} or less in EGEA were followed up in the SLSJ and MRC studies. The association analysis in the SLSJ and MRC studies used the same model as in EGEA. Results of stage 2 data sets and then of the 3 data sets were combined by using a fixed-effects meta-analysis. SNPs were declared to be significantly associated with atopy if the meta-analysis *P* value (P_{meta}) of 3 data sets reached the genome-wide significance level of 1.5×10^{-7}. This threshold was obtained by dividing the type I error of 5% by the effective number of independent SNPs in the Illumina 610-Quad array.[28]

Selection of gene pairs using both statistical and text-mining filters

Statistical filtering consisted of selecting 2 sets of genes using the GWAS results: genes showing significant association with atopy (set 1) and genes showing suggestive association with atopy (set 2). Set 1 included all genes at a distance of 50 kb or less from SNPs reaching the genome-wide significance level in the GWAS meta-analysis. Set 2 included all genes that were at most 50 kb apart from SNPs with a *P* value of 10^{-4} or less in the stage 1 EGEA data set and were not part of set 1. To assign SNPs to genes, we used the National Center for Biotechnology Information's dbSNP Build 137 and human Genome Build 37.3.

We further filtered gene pairs (formed by crossing set 1 genes with set 2 genes) through GRAIL[17] text mining of PubMed abstracts (available in October 2014). For each gene, GRAIL builds a vector of words in which the elements of this vector are weights that take values between 0 and 1 depending on how often a word is found with a gene in an abstract. Then GRAIL computes pairwise similarity between genes from gene/word vectors and ranks the similarities between each gene from set 1 and all genes of the genome. The P_{GRAIL} value of a gene from set 2 with a gene from set 1 is equal to the proportion of all genes that have similarity with the set 1 gene greater than the similarity between set 2 and set 1 genes (ie, rank divided by total number of genes across the genome). We used the threshold of a P_{GRAIL} value of .10 or less, as recommended,[17] to select related gene pairs for further epistasis analysis.

SNP-SNP interaction analysis for selected gene pairs

As for the single-SNP association analysis, we performed a 2-stage SNP-SNP interaction analysis. At stage 1, we analyzed all SNP-SNP interactions for the GRAIL-selected gene pairs in the EGEA data set. For each gene, we considered all SNPs lying within gene boundaries. Pairwise SNP-SNP interactions were evaluated by using logistic regression, assuming an additive model for SNP main effects and interaction and adjusting for the same covariates (age, sex, and principal components) as in the GWAS by using Stata software (version 14.1). We used the same coding scheme as usually proposed for SNP-SNP interaction modeling.[29] We modeled the additive effect of an SNP by coding the genotypes of homozygotes for the minor allele, heterozygotes, and homozygotes for the major allele as 1, 0, and −1; the interaction term between 2 SNPs was obtained by means of multiplication of these genotypic values for the 2 SNPs. The test of interaction was based on a likelihood ratio test that follows a χ^2 distribution with 1 *df*. We discarded all SNP pairs for which 1 or more of the 9 genotype combinations appeared in fewer than 5 subjects (cases or control subjects). In a second stage all SNP pairs showing suggestive evidence for interaction in EGEA ($P_{int} \leq 5 \times 10^{-3}$) were followed up in the SLSJ and MRC studies. The results of the stage 2 data sets and then of the 3 data sets were meta-analyzed by using a fixed-effects model.

To correct for multiple testing, for each gene pair investigated, we computed the effective number of independent interaction tests from the eigenvalues of the correlation matrix of products of SNP variables by using an

TABLE I. *ADGRV1* locus on 5q14 showing significant association with atopy

				Stage 1	
				EGEA (n = 1,660)	
SNP	Position (kb)*	Alleles†	MAF‡	β (SE)§	*P* value‖
rs4244205	90,188	A/G	0.41	−0.35 (0.08)	1.1×10^{-5}
rs4916831	90,212	A/G	0.44	−0.40 (0.08)	1.0×10^{-6}
rs10060641	90,213	T/C	0.38	−0.39 (0.08)	7.2×10^{-7}
rs12054681	90,217	C/A	0.37	−0.39 (0.08)	9.7×10^{-7}
rs949787	90,251	G/T	0.28	−0.33 (0.08)	5.5×10^{-5}

*Position in kilobases according to National Center for Biotechnology Information's dbSNP Build 137; †major allele/minor allele; ‡minor allele frequency; and §β is the regression coefficient for a 1-unit increase of the effect allele in logistic regression, assuming an additive model. SE represents the SE associated with the regression coefficient. ‖*P* value associated with the Wald test of SNP effect; ¶*P* value associated with the Wald test of meta-analyzed SNP effect in the stage 2 data sets (SLSJ and MRC studies); #*P* value associated with the Wald test of meta-analyzed SNP effect in the 3 data sets (EGEA and the SLSJ and MRC studies) (the *P* value is shown in boldface when it reached the multiple-testing corrected threshold of 1.5×10^{-7}); and ***P* value associated with the Cochran Q test of homogeneity across the 3 data sets.

extension of the method of Li and Ji.[30] The corrected threshold to declare an interaction statistically significant was equal to the 5% type I error divided by the sum of the effective number of independent interaction tests over all gene pairs tested.

Stratified analyses according to asthma status

Because family samples were ascertained through asthmatic probands, we investigated whether SNP associations and SNP-SNP interactions detected with atopy might be related to the presence of asthma. Single-SNP and SNP-SNP interaction analyses were repeated in the 2 groups of asthmatic and nonasthmatic subjects separately. These analyses were performed for SNPs that showed significant results in meta-analyses of the 3 data sets. Homogeneity of the odds ratios (ORs) between the 2 groups was tested by using the Cochran Q statistic.[31]

RESULTS

Descriptive statistics

A total of 1660 EGEA, 1138 SLSJ study, and 446 MRC study subjects were included in this study. The proportion of atopic subjects was similar in EGEA and SLSJ studies (55.7% and 56.3%, respectively) but was lower in the MRC study (23.8%, $P \leq 10^{-3}$). In each study there was a higher proportion of male subjects among atopic than in nonatopic subjects, and atopic subjects were younger than nonatopic subjects (see Table E1 in this article's Online Repository at www.jacionline.org). As expected, the proportion of asthmatic patients was higher in atopic than in nonatopic subjects in all data sets (see Table E1). In EGEA 78.0% (75.0% and 78.3% in the SLSJ and MRC studies, respectively) of atopic subjects had positive SPT responses to indoor allergens, 55.5% (77.5% and 52.8% in the SLSJ and MRC studies, respectively) to outdoor allergens, and 34.8% (14.8% and 12.3% in the SLSJ and MRC studies, respectively) to molds.

Genome-wide single-SNP analysis

In the stage 1 EGEA data set, no SNP reached the genome-wide significance level of 1.5×10^{-7} (see quantile-quantile and Manhattan plots in Figs E1 and E2 in this article's Online Repository at www.jacionline.org). However, 73 SNPs in 47 loci showed associations with atopy exceeding the screening threshold of a *P* value of 10^{-4} or less. These SNPs were followed up in the stage 2 SLSJ and MRC data sets and meta-analyzed (see Table E2 in this article's Online Repository at www.jacionline.org). The SNP rs4916831 within the adhesion G protein–coupled receptor V1 *(ADGRV1)* gene at 5q14 reached the genome-wide significance level ($P_{meta} = 6.8 \times 10^{-9}$) in the overall meta-analysis of the 3 data sets (Table I). Four other SNPs at that locus in moderate linkage disequilibrium (LD) with rs4916831 (r^2 range, 0.51-0.79) showed a suggestive association ($4.3 \times 10^{-7} \leq P_{meta} \leq 3.8 \times 10^{-6}$, Table I).

Selection of gene pairs by using both statistical and text-mining filters

Gene set 1 included *ADGRV1*, the only gene significantly associated with atopy. There were 30 genes that were fewer than 50 kb apart from the 65 SNPs at 46 loci with a *P* value of 10^{-4} or less in EGEA (after exclusion of *ADGRV1* SNPs), and these formed gene set 2 (see Table E3 in this article's Online Repository at www.jacionline.org). When GRAIL was applied to 30 gene pairs (date accessed: April 24, 2015) formed by each of these 30 genes with *ADGRV1*, 3 genes were related to *ADGRV1* at a P_{GRAIL} value of less than .10: dynein axonemal heavy chain 5 *(DNAH5)* on 5p15 ($P_{GRAIL} = .084$), *CHD7* on 8q12 ($P_{GRAIL} = 3.2 \times 10^{-3}$), and *ATP8B1* on 18q21 ($P_{GRAIL} = .016$).

SNP-SNP interaction analysis for selected gene pairs

In the stage 1 EGEA data set, the 3 GRAIL-selected gene pairs (*ADGRV1/DNAH5*, *ADGRV1/CHD7*, and *ADGRV1/ATP8B1)* were each examined for SNP-SNP interactions, making a total of 5324 SNP pairs. There were 37 SNP pairs that reached a P_{int} value of 5×10^{-3} or less in EGEA, and these were followed up in the SLSJ and MRC studies at stage 2. Two of these SNP pairs, which are related to the *ADGRV1-DNAH5* gene pair, met the multiple-testing corrected threshold, which is estimated to be 7.3×10^{-5} (see Table E4 in this article's Online Repository at www.jacionline.org), in the meta-analysis of the 3 data sets (Table II). The 2 significant interactions involved the same SNP, rs17554723, within *ADGRV1* and 2 SNPs within *DNAH5*, rs2134256 ($P_{meta\text{-}int} = 3.6 \times 10^{-5}$) and rs1354187 ($P_{meta\text{-}int} = 6.1 \times 10^{-5}$), that are in moderate LD ($r^2 = 0.50$, $D' = 0.95$). However, further conditional regression analysis in each of the strata defined by genotypes at *ADGRV1* rs17554723 showed that *DNAH5* rs1354187 was no longer significantly

TABLE I. *(Continued)*

Stage 2								
SLSJ study (n = 1,138)		MRC studies (n = 446)		Meta-analysis		Overall meta-analysis		
β (SE)§	*P* value‖	β (SE)§	*P* value‖	β (SE)§	P_{stage2}¶	β (SE)§	P_{meta}#	$P_{Cochran}$**
−0.19 (0.10)	6.0×10^{-2}	−0.14 (0.18)	.45	−0.18 (0.09)	4.4×10^{-2}	−0.27 (0.06)	3.8×10^{-6}	.35
−0.32 (0.11)	2.3×10^{-3}	−0.21 (0.17)	.23	−0.29 (0.09)	1.2×10^{-3}	−0.35 (0.06)	$\mathbf{6.8 \times 10^{-9}}$	.59
−0.17 (0.11)	.11	−0.24 (0.19)	.22	−0.18 (0.09)	4.8×10^{-2}	−0.30 (0.06)	4.3×10^{-7}	.23
−0.14 (0.11)	.18	−0.29 (0.21)	.16	−0.17 (0.09)	6.5×10^{-2}	−0.30 (0.06)	7.8×10^{-7}	.18
−0.23 (0.10)	2.0×10^{-2}	−0.19 (0.20)	.36	−0.22 (0.09)	1.2×10^{-2}	−0.28 (0.06)	3.2×10^{-6}	.65

associated with atopy ($P \geq .15$) when conditioning on *DNAH5* rs2134256. The most significant SNP pair shows a pattern of interaction in which the ORs for atopy associated with the TT (or CC) genotype at *DNAH5* rs2134256 are in opposite directions according to genotype AA (or GG) at *ADGRV1* rs17554723 (Fig 2). This pattern was consistent in all 3 data sets (Fig 2).

Stratified analyses according to asthma status

Association analyses of atopy with the genome-wide significant *ADGRV1* rs4916831 SNP in asthmatic and nonasthmatic subjects did not show any relationship with the presence of asthma in the stage 1 and 2 data sets and meta-analysis of the 3 data sets ($P_{Cochran}$ for test of homogeneity between the 2 groups ≥ .82; see Table E5, *top*, in this article's Online Repository at www.jacionline.org). In the meta-analysis the evidence for association was even stronger, although not significantly so in nonasthmatic subjects ($P = 7.8 \times 10^{-6}$) than in asthmatic patients ($P = 1.4 \times 10^{-4}$). Similarly, interaction analyses for *ADGRV1* and *DNAH5* SNPs did not show any relationship with asthma ($P_{Cochran} \geq .30$; see Table E5, *bottom*). The evidence for interaction was only significant in nonasthmatic subjects (see Table E6, in this article's Online Repository at www.jacionline.org); this can be explained, at least in part, by the larger sample size of nonasthmatic (n = 1849) than asthmatic (n = 1354) subjects.

Functional annotations of SNPs showing significant results

All SNPs that show significant association (or interaction) with atopy are intronic. The 2 *ADGRV1* SNPs, rs4916831 and rs17554723, on 5q14, which were detected through GWAS and interaction analysis, are 120 kb apart in introns 83 (rs4916831) and 70 (rs17554723) and are in low LD ($r^2 = 0.20$, $D' = 0.75$). The 2 *DNAH5* SNPs (rs2134256 and rs1354187) at 5p15.2 are located in introns 58 and 60 (8 kb apart), but only rs2134256 accounts for the interaction signal (see above). By interrogating the Genotype-Tissue Expression database,[32] rs4916831 was found to be associated with *ADGRV1* expression in esophageal mucosa ($P = 7.5 \times 10^{-7}$).[32] We also investigated whether the *ADGRV1* and *DNAH5* SNPs (as well as their proxies, $r^2 \geq 0.80$) map to functionally important regulatory regions using HaploRegV4.[33] As shown in Table E6, these SNPs and/or proxies map to binding sites of various transcription factors. In addition, 4 proxies of *ADGRV1* rs4916831 map to enhancer histone marks in the lung and skin, whereas a proxy of *DNAH5* rs2134256 maps to promoter and enhancer marks in hematopoietic stem cells.

DISCUSSION

By combining genome-wide single-SNP analysis and epistasis analysis driven by statistical and knowledge-based evidence in 3 asthma-ascertained family data sets, we identified significant association of atopy at a novel 5q14 locus harboring the *ADGRV1* gene and significant interaction between *ADGRV1* and *DNAH5* genetic variants.

Interaction between *ADGRV1* and *DNAH5* variants has biological relevance because these 2 genes are involved in ciliopathies and ciliary function. Ciliopathies comprise a group of disorders associated with genetic mutations encoding defective proteins, which result in either abnormal formation or function of cilia.[34] Mutations in the *ADGRV1* gene cause Usher syndrome type IIC, a ciliopathy characterized by hearing loss and visual impairment,[35,36] whereas mutations of the *DNAH5* gene cause primary ciliary dyskinesia type 3, a ciliopathy that combines upper and lower tract respiratory manifestations, male infertility, and situs inversus.[37] The ADGRV1 protein (also called GPR98) is a component of the Usher protein network that functions in stereocilia of inner-ear hair cells and photoreceptor cilia. The heavy chain of axonemal dynein, as encoded by *DNAH5*, is part of a microtubule-associated motor protein complex responsible for cilia mobility, especially in respiratory epithelial cells, where ciliary motility is essential for mucus transport and airway clearance.[38] Although the respective function of the ADGRV1 and DNAH5 proteins was initially described in different organs, these proteins can also have related functions. Indeed, the cilium in photoreceptors is ultrastructurally very similar to the nasal ciliated epithelium, and the nasal ciliated epithelium of patients with Usher syndrome II was found to have a lower ciliary beat frequency than healthy control subjects.[39] Moreover, Usher syndrome has been reported to be associated with bronchiectasis, sinusitis, and reduced nasal mucociliary clearance.[40]

In addition to the involvement of both ADGRV1 and DNAH5 proteins in ciliary functions, which supports the statistical interaction found between these 2 genes, both *DNAH5* and *ADGRV1* have been associated with respiratory diseases and related phenotypes. A recent GWAS reported significant

TABLE II. SNP pairs showing significant interaction for atopy

					Stage 1		
					EGEA (n = 1,660)		
					Main effect	Interaction	
SNPs	Chromosome*	Genes†	Alleles‡	MAF§	β (SE)‖	β (SE)‖	P_{int}¶
rs17554723	5	*ADGRV1*	A/G	0.33	−0.06 (0.10)	−0.38 (0.12)	3.0×10^{-3}
rs2134256‡‡	5	*DNAH5*	T/C	0.25	0.13 (0.09)		
rs17554723	5	*ADGRV1*	A/G	0.33	0.03 (0.09)	−0.34 (0.11)	3.3×10^{-3}
rs1354187‡‡	5	*DNAH5*	T/C	0.36	0.08 (0.08)		

*Chromosome represents the chromosome number at which the SNP is located; †symbol of the gene where the SNP lies; ‡major allele/minor allele; §minor allele frequency; and ‖β for the main effect is the regression coefficient for a 1-unit increase of the effect allele in logistic regression, assuming an additive model. β for interaction is the regression coefficient for homozygotes for the minor allele at the 2 loci or homozygotes for the major allele at the 2 loci with respect to heterozygotes at either 1 or the 2 loci by using the coding scheme under an additive genetic model described in the methods section. SE represents the SE associated with the regression coefficient.
¶*P* value of the likelihood ratio test for interaction between SNPs (which follows a χ^2 distribution with 1 *df* assuming an additive model); #*P* value associated with the Wald test of meta-analyzed interaction effect in the stage 2 data sets (SLSJ and MRC studies); **P* value associated with the Wald test of meta-analyzed interaction effect in the 3 data sets (EGEA and SLSJ and MRC studies) ($P_{meta\text{-}int}$ is shown in boldface when it reached the multiple-testing corrected threshold of 7.3×10^{-5}); ††*P* value associated with the Cochran Q test of homogeneity across the 3 data sets; and ‡‡The 2 *DNAH5* SNPs, rs1354187 and rs2134256, showing significant interaction with *ADGRV1* SNP are in moderate LD ($r^2 = 0.50$, $D' = 0.95$).

association of *DNAH5* variants with total lung capacity in patients with chronic obstructive pulmonary disease[41] and suggestive association with IgE grass sensitization.[9] However, the SNP reported by that latter study was not in LD with the *DNAH5* SNPs interacting with *ADGRV1* SNP in this study ($r^2 < 0.13$). Based on an approach similar to ours, which combined genome-wide expression data in nasal epithelial cells, allele frequency variation between populations, and literature search to select candidate genes, nominal association of asthma with *DNAH5* was reported, and stronger association was found with *KIF3*, a gene involved in transport of protein complexes within cilia and potentially in allergen clearance as *DNAH5*.[42] In addition, *DNAH5* belongs to the same gene family as *DNAH9*, which showed interaction with environmental tobacco smoke exposure for bronchial hyperresponsiveness in EGEA and SLSJ families.[43] Moreover, suggestive association of *ADGRV1* with asthma has been reported recently by a meta-analysis of GWASs.[44] Although most previously reported associations concern asthma or respiratory phenotypes, the interaction between *ADGRV1* and *DNAH5* SNPs associated with atopy in the present study appears independent of asthma, as shown by the stratified analysis on asthma. Although the mechanism by which these 2 genes influence atopy is still unknown, we can hypothesize that they are involved in dysfunction of cilia that move secreted mucus-containing trapped foreign particles up and out of the airways, which favors allergic sensitization. This is supported by recent observations of differential mRNA expression of both *ADGRV1* and *DNAH5* genes in sputum from house dust mite–sensitized wheezing subjects compared with nonatopic control subjects.[45] Furthermore, both *DNAH5* and other genes of the same family, including *DNAH9*, were among the highest-ranking coexpression hubs in one of the house dust mite–induced wheezing–associated gene modules, which was strongly enriched with genes involved in the function of ciliated epithelial cells.[45] All these observations suggest cilia-related genes might constitute an important emerging pathway for atopy.

The strategy used in this study that enabled identifying novel relevant candidates for atopy combined genome-wide single-SNP analysis and gene-gene interaction analysis based on both statistical filtering of GWAS results and text-mining filtering. It is of note that our 3-step strategy was designed *a priori* and that SNP-SNP interaction tests were only performed for gene pairs selected through our 2 filtering processes. The genome-wide single-SNP analysis pointed to 1 gene *(ADGRV1)* that harbored the lead SNP rs4916831 reaching genome-wide significance and 4 other SNPs showing suggestive association. By increasing the density of SNPs through HapMap2-based imputation at that locus, an additional SNP ($r^2 = 0.80$ with rs4916831) reached genome-wide significance, and 6 other SNPs had *P* values within 1 order of magnitude of the genome-wide threshold (results not shown), which strengthens our finding. Further conditional analysis in that region showed that association with atopy was accounted for only by the lead genotyped SNP. The subsequent statistical and text-mining filters used before epistasis analysis made it possible to detect gene-gene interaction by decreasing the multiple testing burden. Indeed, use of both filters reduced the number of interaction tests by 9-fold compared with use of the statistical filter only. The text-mining filter was based on GRAIL, which was shown to be successful in pointing out true disease regions that were validated.[17] Although many sources of biological information can be used to connect genes, such as coexpression gene networks or protein-protein interaction networks, the advantage of GRAIL is to provide a broader framework for revealing gene-gene relationships of any origin through literature search. However, GWASs and candidate gene studies, which are driven by researchers' expectations, can create a bias toward genes that are frequently reported in the literature. An appropriate approach would be to use the existing knowledge and to correct for potential bias, but to our knowledge, such a method does not yet exist.

In conclusion, this study shows that the proposed strategy that combines GWAS and epistasis analysis driven by statistical and knowledge-based evidence can successfully identify strong candidate genes for complex phenotypes as atopy. The interaction between *DNAH5* and *ADGRV1*, 2 genes involved in ciliary function, is of biological relevance and provides a novel mechanism underlying atopy. Further studies, including

TABLE II. *(Continued)*

Stage 2									Overall meta-analysis			
SLSJ study (n = 1,138)			MRC studies (n = 446)			Meta-analysis						
Main effect	Interaction		Main effect	Interaction		Main effect	Interaction		Main effect	Interaction		
β (SE)‖	β (SE)‖	P_{int}¶	β (SE)‖	β (SE)‖	P_{int}¶	β (SE)‖	β (SE)‖	$P_{stage2\text{-}int}$#	β (SE)‖	β (SE)‖	$P_{meta\text{-}int}$**	$P_{Cochran}$††
0.02 (0.13)	−0.42 (0.16)	.01	0.13 (0.23)	−0.28 (0.33)	.40	0.04 (0.11)	−0.39 (0.14)	.006	−0.02 (0.08)	−0.38 (0.09)	**3.6×10^{-5}**	.84
−0.06 (0.13)			0.03 (0.26)			−0.04 (0.12)			0.06 (0.07)			
0.13 (0.11)	−0.35 (0.14)	.01	0.19 (0.21)	−0.16 (0.29)	.59	0.14 (0.10)	−0.32 (0.12)	.009	0.08 (0.06)	−0.33 (0.08)	**6.1×10^{-5}**	.92
−0.04 (0.11)			0.05 (0.22)			−0.02 (0.10)			0.04 (0.06)			

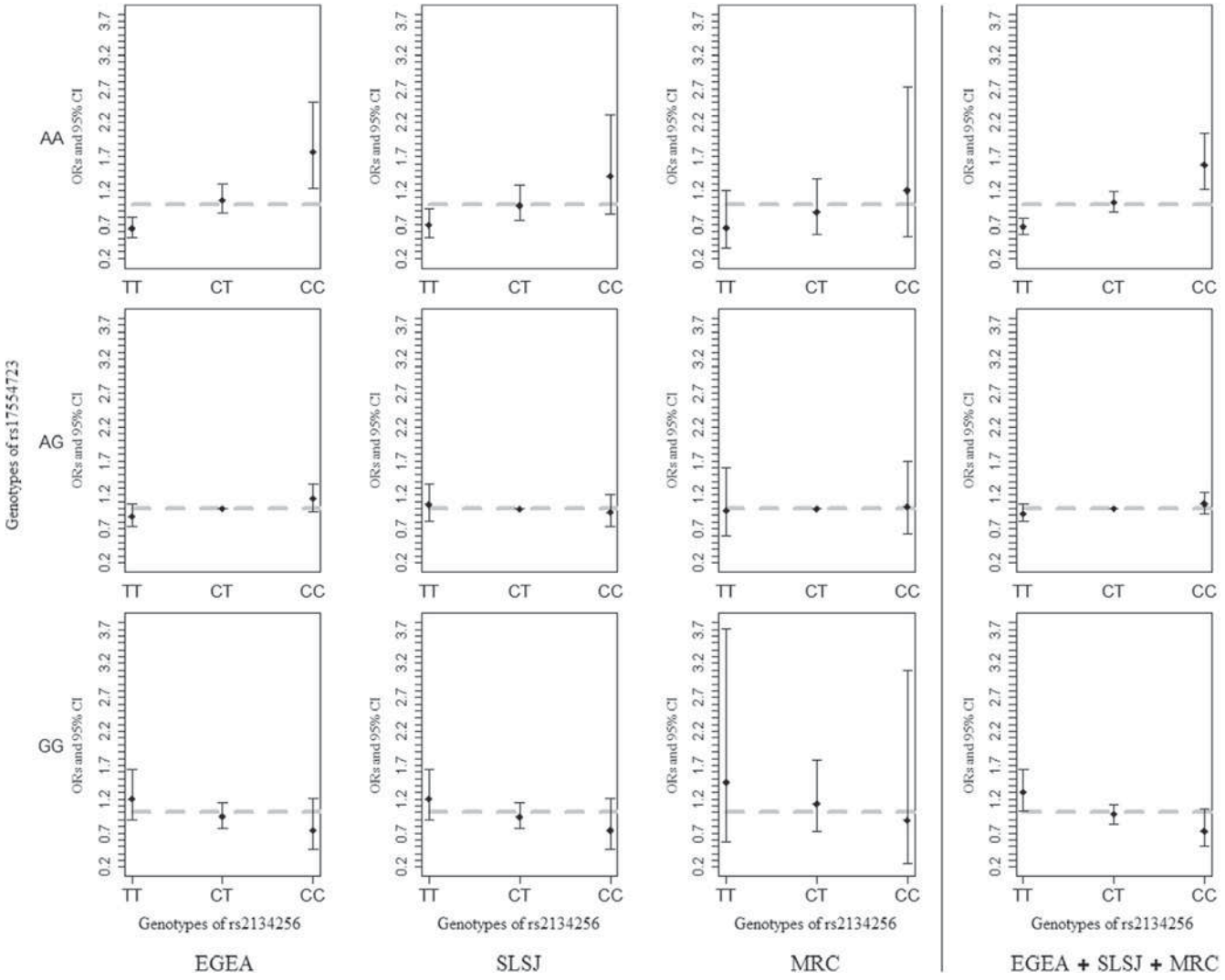


FIG 2. ORs and 95% CIs for atopy associated with each genotype at *DNAH5* rs2134256 (TT, CT, and CC) in each of the strata defined by genotypes at *ADGRV1* rs17554723 (AA, AG, or GG). These ORs were calculated by using the genotype-coding scheme defined in the text and are shown for each of the 3 data sets (EGEA and the SLSJ and MRC studies) and for the combined data set.

functional and experimental studies, are needed to confirm the current findings and identify the functional variants.

EGEA: We thank all those who participated in the setting of the study and on the various aspects of the examinations involved: interviewers; technicians for lung function testing and skin prick tests; blood sampling; IgE determinations; coders; those involved in QC, data, and sample management; and all those who supervised the study in all EGEA centers. We thank the 3 CIC-INSERM of Necker, Grenoble, and Marseille who supported the EGEA study and in which subjects were examined. We thank the biobanks in Lille (CIC-INSERM) and at Annemasse (Etablissement français du sang), where EGEA biological samples are stored. Finally, we thank the EGEA Cooperative Group members as follows.

Coordination: V. Siroux (epidemiology, principal investigator since 2013), F. Demenais (genetics), I. Pin (clinical aspects), R. Nadif (biology), F. Kauffmann (principal investigator 1992-2012); *Respiratory epidemiology—INSERM U 700, Paris*: M. Korobaeff (EGEA1), F. Neukirch (EGEA1); *INSERM U 707, Paris*: I. Annesi-Maesano (EGEA1-2); *INSERM CESP/U 1018, Villejuif*: F. Kauffmann, N. Le Moual, R. Nadif, M. P. Oryszczyn (EGEA1-2), R. Varraso; *INSERM U 823, Grenoble*: V. Siroux. Genetics: *INSERM U 393, Paris*: J. Feingold; *INSERM U 946, Paris*: E. Bouzigon, F. Demenais, M. H. Dizier; *CNG, Evry*: I. Gut (now CNAG, Barcelona, Spain), M. Lathrop (now University of McGill, Montreal, Canada). *Clinical centers—Grenoble*: I. Pin, C. Pison; *Lyon*: D. Ecochard (EGEA1), F. Gormand, Y. Pacheco; *Marseille*: D. Charpin (EGEA1), D. Vervloet (EGEA1-2); *Montpellier*: J. Bousquet; *Paris Cochin*: A. Lockhart (EGEA1), R. Matran (now in Lille); *Paris Necker*: E. Paty (EGEA1-2), P. Scheinmann (EGEA1-2); *Paris-Trousseau*: A. Grimfeld (EGEA1-2), J. Just. *Data and quality management—INSERM ex-U155* (EGEA1): J. Hochez; *INSERM CESP/U 1018, Villejuif*: N. Le Moual; *INSERM ex-U780*: C. Ravault (EGEA1-2); *INSERM ex-U794*: N. Chateigner; *Grenoble*: J. Ferran (EGEA1-2).

SLSJ: We thank all participants included in the SLSJ asthma familial collection. Catherine Laprise built, coordinates, and manages the SLSJ study. Drs Paul Bégin and Charles Morin confirmed the respiratory diagnosis for the adults and children, respectively. We also thank the laboratory technicians (Nadia Mior and Denise Morin), research professionals (Anne-Marie Madore), and nurses (from the ECOGENE-21 Biocluster). Catherine Laprise is the Canada Research Chair in Environment and Genetics of Respiratory Disorders and Allergy, Director of the Asthma Strategic Group of the Respiratory Health Network (RHN) of Fonds de la recherche en santé du Québec (FRSQ), and researcher of the AllerGen NCE.

MRC: We thank all the families who participated and were recruited as part of the MRCA and MRCE collections (coordination: M. F. Moffatt, W. O. C. M Cookson, and J. I. Harper). We thank the research nurses involved, as well as personnel who conducted blood extractions, IgE determinations, and sample management (Anna Dixon). Genotyping of the cohorts was done by the CNG, Evry: I. Gut (now CNAG, Barcelona, Spain), M. Lathrop (now University of McGill, Montreal, Canada).

Key Messages

- ***ADGRV1* genetic variants are associated with atopy in families with asthma.**
- **Interaction between *ADGRV1* and *DNAH5* variants is associated with atopy; these 2 genes are involved in ciliary function.**
- **Use of a strategy that combines GWAS and epistasis analysis driven by statistical and knowledge-based evidence can be used to successfully identify new genes underlying complex traits.**

REFERENCES

1. Ober C, Yao TC. The genetics of asthma and allergic disease: a 21st century perspective. Immunol Rev 2011;242:10-30.
2. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 2014;42:D1001-6.
3. Bonnelykke K, Sparks R, Waage J, Milner JD. Genetics of allergy and allergic sensitization: common variants, rare mutations. Curr Opin Immunol 2015;36:115-26.
4. Los H, Postmus PE, Boomsma DI. Asthma genetics and intermediate phenotypes: a review from twin studies. Twin Res 2001;4:81-93.
5. Thomsen SF, Ulrik CS, Kyvik KO, Ferreira MA, Backer V. Multivariate genetic analysis of atopy phenotypes in a selected sample of twins. Clin Exp Allergy 2006;36:1382-90.
6. Vercelli D. Discovering susceptibility genes for asthma and allergy. Nat Rev Immunol 2008;8:169-82.
7. Andiappan AK, Wang de Y, Anantharaman R, Parate PN, Suri BK, Low HQ, et al. Genome-wide association study for atopy and allergic rhinitis in a Singapore Chinese population. PLoS One 2011;6:e19719.
8. Castro-Giner F, Bustamante M, Ramon Gonzalez J, Kogevinas M, Jarvis D, Heinrich J, et al. A pooling-based genome-wide analysis identifies new potential candidate genes for atopy in the European Community Respiratory Health Survey (ECRHS). BMC Med Genet 2009;10:128.
9. Ramasamy A, Curjuric I, Coin LJ, Kumar A, McArdle WL, Imboden M, et al. A genome-wide meta-analysis of genetic variants associated with allergic rhinitis and grass sensitization and their interaction with birth order. J Allergy Clin Immunol 2011;128:996-1005.
10. Wan YI, Strachan DP, Evans DM, Henderson J, McKeever T, Holloway JW, et al. A genome-wide association study to identify genetic determinants of atopy in subjects from the United Kingdom. J Allergy Clin Immunol 2011;127:223-31, e1-3.
11. Bonnelykke K, Matheson MC, Pers TH, Granell R, Strachan DP, Alves AC, et al. Meta-analysis of genome-wide association studies identifies ten loci influencing allergic sensitization. Nat Genet 2013;45:902-6.
12. Hinds DA, McMahon G, Kiefer AK, Do CB, Eriksson N, Evans DM, et al. A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. Nat Genet 2013;45:907-11.
13. Dizier MH, Margaritte-Jeannin P, Madore AM, Esparza-Gordillo J, Moffatt M, Corda E, et al. The ANO3/MUC15 locus is associated with eczema in families ascertained through asthma. J Allergy Clin Immunol 2012;129:1547-53.e3.
14. Sun X, Lu Q, Mukherjee S, Crane PK, Elston R, Ritchie MD. Analysis pipeline for the epistasis search—statistical versus biological filtering. Front Genet 2014;5:106.
15. Li ZC, Huang MH, Zhong WQ, Liu ZQ, Xie Y, Dai Z, et al. Identification of drug-target interaction from interactome network with "guilt-by-association" principle and topology features. Bioinformatics 2016;32:1057-64.
16. Luo Y, Riedlinger G, Szolovits P. Text mining in cancer gene and pathway prioritization. Cancer Inform 2014;13:69-79.
17. Raychaudhuri S, Plenge RM, Rossin EJ, Ng AC, Purcell SM, Sklar P, et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet 2009;5:e1000534.
18. Kauffmann F, Dizier MH, Annesi-Maesano I, Bousquet J, Charpin D, Demenais F, et al. EGEA (Epidemiological study on the Genetics and Environment of Asthma, bronchial hyperresponsiveness and atopy)—descriptive characteristics. Clin Exp Allergy 1999;29(suppl 4):17-21.
19. Kaufmann F, Dizier MH, Pin I, Paty E, Gormand F, Vervloet D, et al. Epidemiological study of the genetics and environment of asthma, bronchial hyperresponsiveness, and atopy. Am J Respir Crit Care Med 1997;156:123-9.
20. Bouzigon E, Nadif R, Le Moual N, Dizier MH, Aschard H, Boudier A, et al. [Genetic and environmental factors of asthma and allergy: Results of the EGEA study]. Rev Mal Respir 2015;32:822-40.
21. Laprise C. The Saguenay-Lac-Saint-Jean asthma familial collection: the genetics of asthma in a young founder population. Genes Immun 2014;15:247-55.
22. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 2007;448:470-3.
23. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, et al. A large-scale, consortium-based genomewide association study of asthma. N Engl J Med 2010;363:1211-21.
24. Sarnowski C, Sugier PE, Granell R, Jarvis D, Dizier MH, Ege M, et al. Identification of a new locus at 16q12 associated with time to asthma onset. J Allergy Clin Immunol 2016;138:1071-80.

25. Liang L, Morar N, Dixon AL, Lathrop GM, Abecasis GR, Moffatt MF, et al. A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. Genome Res 2013;23:716-26.
26. Li Y, Abecasis GR. Mach 1.0: rapid haplotype reconstruction and missing genotype inference. Am J Hum Genet 2006;S79:2290.
27. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet 2009;10:387-406.
28. Li MX, Yeung JM, Cherny SS, Sham PC. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference data sets. Hum Genet 2012;131:747-56.
29. Herold C, Steffens M, Brockschmidt FF, Baur MP, Becker T. INTERSNP: genome-wide interaction analysis guided by a priori information. Bioinformatics 2009;25:3275-81.
30. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity 2005;95:221-7.
31. Cochran WG. The comparison of percentages in matched samples. Biometrika 1950;37:256-66.
32. The GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015;348:648-60.
33. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 2012;40:D930-4.
34. Fliegauf M, Benzing T, Omran H. When cilia go bad: cilia defects and ciliopathies. Nat Rev Mol Cell Biol 2007;8:880-93.
35. Weston MD, Luijendijk MW, Humphrey KD, Moller C, Kimberling WJ. Mutations in the VLGR1 gene implicate G-protein signaling in the pathogenesis of Usher syndrome type II. Am J Hum Genet 2004;74:357-66.
36. Besnard T, Vache C, Baux D, Larrieu L, Abadie C, Blanchet C, et al. Non-USH2A mutations in USH2 patients. Hum Mutat 2012;33:504-10.
37. Leigh MW, Zariwala MA, Knowles MR. Primary ciliary dyskinesia: improving the diagnostic approach. Curr Opin Pediatr 2009;21:320-5.
38. Olbrich H, Horvath J, Fekete A, Loges NT, Storm van's Gravesande K, Blum A, et al. Axonemal localization of the dynein component DNAH5 is not altered in secondary ciliary dyskinesia. Pediatr Res 2006;59:418-22.
39. Armengot M, Salom D, Diaz-Llopis M, Millan JM, Milara J, Mata M, et al. Nasal ciliary beat frequency and beat pattern in retinal ciliopathies. Invest Ophthalmol Vis Sci 2012;53:2076-9.
40. Bonneau D, Raymond F, Kremer C, Klossek JM, Kaplan J, Patte F. Usher syndrome type I associated with bronchiectasis and immotile nasal cilia in two brothers. J Med Genet 1993;30:253-4.
41. Lee JH, McDonald ML, Cho MH, Wan ES, Castaldi PJ, Hunninghake GM, et al. DNAH5 is associated with total lung capacity in chronic obstructive pulmonary disease. Respir Res 2014;15:97.
42. Kovacic MB, Myers JM, Wang N, Martin LJ, Lindsey M, Ericksen MB, et al. Identification of KIF3A as a novel candidate gene for childhood asthma using RNA expression and population allelic frequencies differences. PLoS One 2011; 6:e23714.
43. Dizier MH, Nadif R, Margaritte-Jeannin P, Barton SJ, Sarnowski C, Gagne-Ouellet V, et al. Interaction between the DNAH9 gene and early smoke exposure in bronchial hyperresponsiveness. Eur Respir J 2016;47: 1072-81.
44. Almoguera B, Vazquez L, Mentch F, Connolly J, Pacheco JA, Sundaresan AS, et al. Identification of four novel loci in asthma in European and African American populations. Am J Respir Crit Care Med 2017;195:456-63.
45. Jones AC, Troy NM, White E, Hollams EM, Gout AM, Ling KM, et al. Persistent activation of interlinked Th2-airway epithelial gene networks in sputum-derived cells from aeroallergen-sensitized symptomatic atopic asthmatics. bioRxiv 2018; 8:1511.

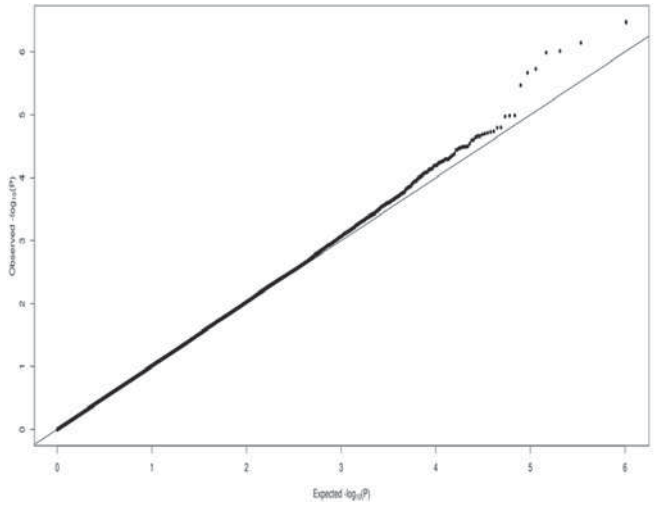


FIG E1. Quantile-quantile plot of GWAS results of atopy in the stage 1 EGEA data set. Quantile-quantile plot for the association of atopy with 501,167 genotyped SNPs that passed QC in the stage 1 EGEA data set. *Dots* represent the distribution of observed $-\log_{10}(P)$ values against the expected $-\log_{10}(P)$ values from a theoretical χ^2 distribution. The *straight line* represents the theoretical distribution of expected $-\log_{10}(P)$ values under the null hypothesis of no association. There was no evidence of any systematic bias: the genomic inflation factor (λ) was equal to 1.017.

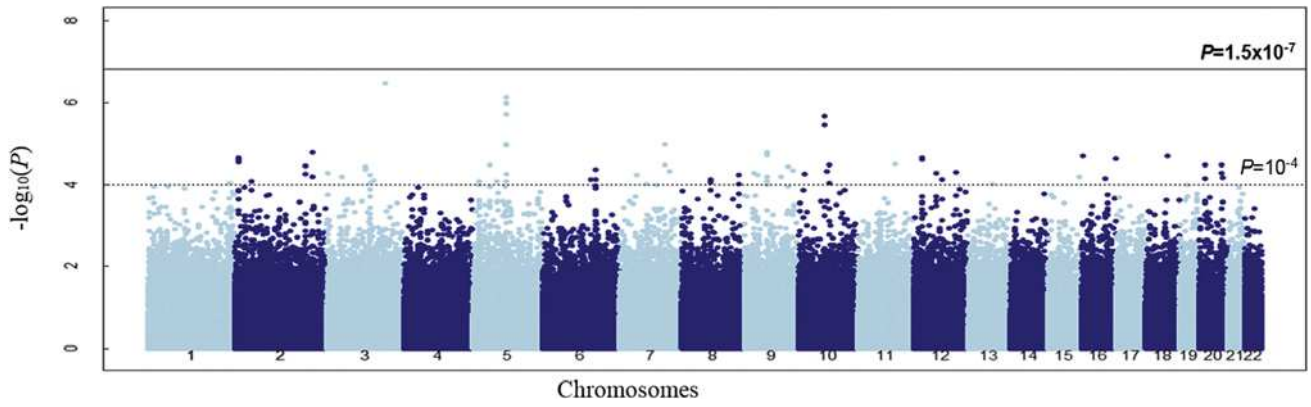


FIG E2. Manhattan plot of GWAS results for atopy in the stage 1 EGEA data set. The *y-axis* shows the $-\log_{10}$ *P* values of the association test of individual SNPs with atopy in the stage 1 EGEA data set, and the *x-axis* shows the SNP chromosomal positions. The *solid horizontal line* indicates the genome-wide significance level ($P = 1.5 \times 10^{-7}$), and the *dotted horizontal line* indicates the suggestive association level ($P = 10^{-4}$) to select SNPs for follow-up in stage 2 SLSJ and MRC data sets.

TABLE E1. Descriptive statistics of EGEA, SLSJ, and MRC data sets

	EGEA			SLSJ study			MRC studies		
	Nonatopic subjects (n = 735)	Atopic subjects (n = 925)	*P* value*	Nonatopic subjects (n = 497)	Atopic subjects (n = 641)	*P* value*	Nonatopic subjects (n = 340)	Atopic subjects (n = 106)	*P* value*
Sex (male), no. (%)	329 (44.8)	539 (58.3)	$<10^{-4}$	205 (41.3)	317 (49.5)	5.9×10^{-3}	117 (34.4)	68 (64.2)	$<10^{-4}$
Age (y), mean (SD)	35.9 (16.7)	27.5 (15.7)	$<10^{-4}$	45.1 (23.4)	32.5 (18.4)	$<10^{-4}$	19.6 (14.7)	12.7 (4.3)	$<10^{-4}$
Asthma, no. (%)	114 (16.0)	514 (56.3)	$<10^{-4}$	144 (29.0)	405 (63.2)	$<10^{-4}$	84 (24.7)	93 (87.7)	$<10^{-4}$
Indoor allergens, no. (%)†	—	720 (78.0)		—	481 (75.0)		—	83 (78.3)	
Outdoor allergens, no. (%)‡	—	512 (55.5)		—	497 (77.5)		—	56 (52.8)	
Molds, no. (%)§	—	320 (34.8)		—	95 (14.8)		—	13 (12.3)	

**P* values for tests of distribution of each feature (sex, age, and asthma) between atopic and nonatopic subjects in each data set (χ^2 test for sex and asthma; Wilcoxon-Mann-Whitney test for age).
†Indoor allergens included house dust mite, cockroach, and cat in EGEA; house dust mite, dust, and cat in the SLSJ study; and dust mite and cat in the MRC studies.
‡Outdoor allergens included timothy grass, olive, birch, ragweed, and pellitory in EGEA; herbs, trees, and animals in the SLSJ study; and timothy grass in the MRC studies.
§Molds included *Alternaria*, *Aspergillus*, and *Cladosporium* species in EGEA; *Alternaria*, *Aspergillus*, *Cladosporium*, *Hormodendrum*, and *Penicillium* species in the SLSJ study; and *Aspergillus* and *Cladosporium* species in the MRC studies.

TABLE E2. Loci showing suggestive association with atopy at a *P* value of 10^{-4} or less in the stage 1 EGEA data set and followed up in the stage 2 SLSJ and MRC data sets

						Stage 1		Stage 2						Overall meta-analysis	
						EGEA (n = 1660)		SLSJ study (n = 1138)		MRC studies (n = 446)		Meta-analysis			
SNP	Chromosome*	Position (kb)	Nearest genes (kb distance)†	Alleles‡	EAF§	OR (95% CI)‖	*P* value#	OR (95% CI)‖	*P* value#	OR‖	*P* value#	OR (95% CI)‖	P_{stage2}¶	OR (95% CI)‖	P_{meta}**
rs1252146	1	237,154	*MTR* (87); *RYR2* (51)	C/T	0.31	1.41 (1.19-1.67)	8.9×10^{-5}	0.75 (0.60-0.93)	.009	1.03 (0.68-1.58)	.88	0.80 (0.66-0.97)	.02	1.10 (0.97-1.25)	.15
rs9428818	1	239,172	*ZP4* (1,118); *CHRM3* (620)	C/T	0.08	1.74 (1.32-2.29)	8.9×10^{-5}	1.60 (1.07-2.41)	.02	1.07 (0.52-2.21)	.86	1.46 (1.02-2.08)	.04	1.63 (1.31-2.02)	1.3×10^{-5}
rs13409750	2	7,610	*RNF144A* (425); *LINC00299* (538)	C/T	0.80	0.64 (0.53-0.79)	2.5×10^{-5}	1.05 (0.84-1.32)	.66	0.95 (0.63-1.45)	.83	1.03 (0.84-1.26)	.78	0.82 (0.71-0.94)	5.9×10^{-3}
rs7567384	2	7,611	*RNF144A* (426); *LINC00299* (537)	C/T	0.80	0.65 (0.53-0.79)	2.8×10^{-5}	1.05 (0.84-1.32)	.66	0.96 (0.63-1.45)	.84	1.03 (0.85-1.25)	.77	0.82 (0.71-0.95)	6.8×10^{-3}
rs1561574	2	7,647	*RNF144A* (463); *LINC00299* (501)	C/T	0.15	1.60 (1.29-2.00)	2.2×10^{-5}	1.07 (0.83-1.37)	.6	1.13 (0.73-1.75)	.57	1.08 (0.87-1.35)	.46	1.32 (1.13-1.54)	4.3×10^{-4}
rs3819892	2	37,510	*PRKD*	G/T	0.86	0.63 (0.50-0.79)	8.1×10^{-5}	1.09 (0.81-1.48)	.56	0.89 (0.46-1.7)	.72	1.05 (0.80-1.39)	.71	0.78 (0.65-0.93)	5.7×10^{-3}
rs10497664	2	186,260	*ZNF804A* (456); *FSIP2* (343)	G/T	0.41	1.38 (1.18-1.61)	5.5×10^{-5}	0.89 (0.74-1.08)	.24	0.79 (0.54-1.14)	.21	0.87 (0.74-1.03)	.11	1.11 (0.99-1.25)	.06
rs13006333	2	186,268	*ZNF804A* (464); *FSIP2* (335)	G/T	0.43	1.38 (1.18-1.61)	3.4×10^{-5}	0.89 (0.74-1.08)	.23	0.76 (0.52-1.11)	.16	0.86 (0.73-1.02)	.09	1.12 (1.00-1.25)	.05
rs12328639	2	211,676	*CPS*1 (132); *ERBB4* (564)	A/G	0.19	1.50 (1.23-1.84)	6.4×10^{-5}	1.11 (0.89-1.39)	.35	1.04 (0.64-1.69)	.88	1.10 (0.90-1.34)	.36	1.29 (1.12-1.48)	4.8×10^{-4}
rs1016403	2	211,704	*CPS1* (161); *ERBB4* (536)	A/G	0.16	1.60 (1.29-1.99)	1.6×10^{-5}	1.17 (0.92-1.48)	.19	1.01 (0.63-1.63)	.95	1.14 (0.92-1.40)	.23	1.35 (1.16-1.56)	9.9×10^{-5}
rs2689860	3	801	*CHL1* (349); *CNTN6* (334)	A/C	0.63	1.37 (1.18-1.59)	5.1×10^{-5}	0.83 (0.68-1.02)	.08	1.10 (0.77-1.57)	.60	0.89 (0.75-1.06)	.20	1.14 (1.02-1.28)	.02
rs445518	3	28,004	*EOMES* (240); *CMC1* (279)	C/T	0.32	0.71 (0.61-0.84)	6.5×10^{-5}	1.21 (0.99-1.47)	.07	1.08 (0.75-1.55)	.69	1.17 (0.99-1.40)	.07	0.90 (0.80-1.02)	.09
rs1007368	3	93,961	*NSUN3* (115); *EPHA6* (2,573)	A/G	0.87	1.59 (1.28-1.99)	4.2×10^{-5}	0.99 (0.71-1.40)	.98	1.37 (0.75-2.48)	.30	1.07 (0.80-1.44)	.63	1.38 (1.16-1.65)	3.8×10^{-4}
rs1584930	3	93,962	*NSUN3* (116); *EPHA6* (2,572)	A/G	0.13	0.62 (0.50-0.78)	3.6×10^{-5}	1.00 (0.71-1.41)	1	0.73 (0.4-1.33)	.31	0.92 (0.69-1.25)	.61	0.72 (0.60-0.86)	3.1×10^{-4}
rs3900940	3	108,148	*MYH15*	C/T	0.68	1.39 (1.18-1.64)	9.3×10^{-5}	0.93 (0.74-1.16)	.51	1.20 (0.84-1.73)	.32	1.00 (0.82-1.21)	.98	1.21 (1.07-1.37)	3.1×10^{-3}
rs11917965	3	108,189	*MYH15*	A/C	0.41	0.72 (0.62-0.85)	5.9×10^{-5}	1.05 (0.85-1.29)	.66	0.81 (0.57-1.15)	.24	0.98 (0.82-1.17)	.83	0.83 (0.74-0.93)	1.6×10^{-3}
rs7633227	3	116,831	*LSAMP-AS4* (180); *IGSF11* (1,789)	C/T	0.62	0.73 (0.62-0.85)	7.9×10^{-5}	0.97 (0.79-1.18)	.73	1.13 (0.77-1.65)	.54	1.00 (0.84-1.19)	.98	0.84 (0.75-0.94)	3.2×10^{-3}
rs4681369	3	147,419	*FLJ30375* (279); *AGTR1* (997)	G/T	0.20	0.62 (0.51-0.74)	3.4×10^{-7}	1.01 (0.80-1.29)	.91	1.03 (0.69-1.54)	.89	1.02 (0.83-1.25)	.87	0.77 (0.67-0.89)	2.3×10^{-4}
rs795540	5	13,818	*DNAH5*	C/T	0.56	0.73 (0.62-0.85)	8.3×10^{-5}	1.02 (0.84-1.23)	.86	0.95 (0.67-1.35)	.76	1.00 (0.85-1.18)	.99	0.85 (0.75-0.95)	4.4×10^{-3}
rs17194068	5	39,821	*DAB2* (396); *PTGER4* (859)	A/G	0.28	1.44 (1.21-1.70)	3.3×10^{-5}	0.81 (0.66-1.00)	.05	0.87 (0.6-1.26)	.45	0.82 (0.69-0.99)	.04	1.11 (0.98-1.26)	.11
rs7711329	5	65,126	*NLN* (1); *ERBB2IP* (96)	C/T	0.19	0.68 (0.56-0.83)	8.3×10^{-5}	1.07 (0.83-1.37)	.61	1.07 (0.65-1.77)	.80	1.07 (0.85-1.33)	.57	0.82 (0.71-0.95)	8.7×10^{-3}

(Continued)

TABLE E2. *(Continued)*

						Stage 1		Stage 2							
						EGEA (n = 1660)		SLSJ study (n = 1138)		MRC studies (n = 446)		Meta-analysis		Overall meta-analysis	
SNP	Chromo-some*	Position (kb)	Nearest genes (kb distance)†	Alleles‡	EAF§	OR (95% CI)‖	*P* value#	OR (95% CI)‖	*P* value#	OR‖	*P* value#	OR (95% CI)‖	P_{stage2}¶	OR (95% CI)‖	P_{meta}**
rs10942608	5	90,083	*ADGRV1*	C/T	0.23	0.71 (0.60-0.84)	8.2×10^{-5}	0.99 (0.81-1.21)	.92	0.75 (0.49-1.16)	.19	0.94 (0.78-1.13)	.52	0.81 (0.71-0.92)	8.4×10^{-4}
rs4244205	5	90,189	*ADGRV1*	A/G	0.41	0.71 (0.61-0.83)	1.1×10^{-5}	0.83 (0.68-1.01)	.06	0.87 (0.62-1.24)	.45	0.84 (0.70-1.00)	.04	0.76 (0.68-0.86)	3.8×10^{-6}
rs4916829	5	90,195	*ADGRV1*	T/G	0.35	0.70 (0.61-0.83)	1.0×10^{-5}	0.94 (0.78-1.16)	.59	0.88 (0.59-1.31)	.52	0.93 (0.75-1.18)	.44	0.79 (0.70-0.89)	1.2×10^{-4}
rs4916831	5	90,212	*ADGRV1*	A/G	0.44	0.67 (0.58-0.79)	1.0×10^{-6}	0.72 (0.59-0.89)	.002	0.81 (0.58-1.14)	.23	0.75 (0.63-0.89)	.001	0.71 (0.63-0.79)	6.8×10^{-9}
rs10060641	5	90,213	*ADGRV1*	T/C	0.38	0.67 (0.58-0.79)	7.2×10^{-7}	0.85 (0.68-1.04)	.11	0.79 (0.54-1.16)	.22	0.83 (0.69-1.00)	.05	0.74 (0.66-0.83)	4.3×10^{-7}
rs12054681	5	90,218	*ADGRV1*	C/A	0.37	0.68 (0.58-0.79)	9.7×10^{-7}	0.87 (0.70-1.06)	.18	0.75 (0.5-1.12)	.16	0.84 (0.70-1.01)	.06	0.74 (0.66-0.84)	7.8×10^{-7}
rs949787	5	90,251	*ADGRV1*	G/T	0.28	0.72 (0.61-0.84)	5.5×10^{-5}	0.79 (0.66-0.96)	.02	0.83 (0.56-1.23)	.36	0.80 (0.67-0.95)	.01	0.76 (0.67-0.85)	3.2×10^{-6}
rs12522571	5	90,259	*ADGRV1*	A/G	0.22	0.66 (0.56-0.78)	1.9×10^{-6}	0.90 (0.73-1.12)	.36	0.75 (0.46-1.23)	.25	0.88 (0.72-1.07)	.19	0.74 (0.65-0.85)	7.4×10^{-6}
rs7769042	6	106,156	*PREP* (305); *PRDM1* (378)	A/G	0.57	1.38 (1.18-1.62)	7.4×10^{-5}	1.14 (0.93-1.39)	.21	0.61 (0.43-0.87)	.01	0.98 (0.82-1.16)	.79	1.18 (1.05-1.33)	5.9×10^{-3}
rs12660166	6	120,894	*MIR3144* (557); *C6orf170* (507)	C/T	0.25	0.70 (0.59-0.84)	7.4×10^{-5}	1.11 (0.87-1.42)	.41	0.95 (0.62-1.45)	.80	1.07 (0.86-1.32)	.55	0.83 (0.73-0.95)	7.2×10^{-3}
rs12194792	6	121,085	*MIR3144* (348); *C6orf170* (316)	C/T	0.58	1.37 (1.18-1.60)	4.3×10^{-5}	0.90 (0.74-1.09)	.27	1.04 (0.69-1.58)	.83	0.92 (0.78-1.10)	.36	1.16 (1.03-1.30)	.01
rs1880617	7	53,138	*POM121L12* (33); *FLJ45974* (585)	G/T	0.39	1.36 (1.17-1.59)	9.5×10^{-5}	1.02 (0.81-1.27)	.89	0.85 (0.57-1.25)	.40	0.97 (0.80-1.18)	.77	1.19 (1.06-1.35)	4.3×10^{-3}
rs12530936	7	91 187	*FZD1* (289); *MTERF* (315)	C/T	0.13	0.64 (0.51-0.80)	9.9×10^{-5}	0.86 (0.63-1.17)	.34	0.83 (0.43-1.6)	.57	0.85 (0.65-1.13)	.27	0.72 (0.60-0.85)	1.8×10^{-4}
rs10279056	7	115,847	*TFEC* (176); *TES* (4)	A/G	0.57	1.40 (1.20-1.62)	1.0×10^{-5}	0.90 (0.74-1.10)	.30	1.05 (0.75-1.48)	.77	0.94 (0.79-1.11)	.45	1.17 (1.05-1.31)	4.9×10^{-3}
rs17138756	7	116,135	*TES* (236); *CAV2* (4)	A/G	0.10	0.58 (0.45-0.75)	3.4×10^{-5}	0.99 (0.77-1.28)	.95	0.93 (0.51-1.7)	.83	0.98 (0.78-1.24)	.89	0.77 (0.65-0.92)	3.9×10^{-3}
rs10253511	7	130,797	*MKLN1*	C/T	0.91	0.59 (0.46-0.76)	4.8×10^{-5}	0.88 (0.61-1.25)	.46	1.97 (0.82-4.74)	.13	0.98 (0.71-1.37)	.92	0.71 (0.58-0.87)	1.6×10^{-3}
rs4237038	8	61,731	*CHD7*	A/G	0.25	1.44 (1.20-1.73)	7.3×10^{-5}	0.97 (0.80-1.18)	.77	1.08 (0.7-1.67)	.74	0.99 (0.83-1.18)	.89	1.19 (1.05-1.35)	7.6×10^{-3}
rs10957162	8	61,758	*CHD7*	A/G	0.27	1.44 (1.20-1.72)	8.5×10^{-5}	0.99 (0.82-1.21)	.95	0.98 (0.63-1.51)	.92	0.99 (0.83-1.18)	.92	1.19 (1.05-1.35)	7.3×10^{-3}
rs7357565	8	134,817	*ST3GAL1* (233); *ZFAT* (673)	A/G	0.51	0.73 (0.62-0.86)	9.7×10^{-5}	1.00 (0.82-1.22)	1	0.63 (0.41-0.96)	.03	0.92 (0.77-1.10)	.36	0.81 (0.72-0.91)	3.9×10^{-4}
rs4399592	8	134 824	*ST3GAL1* (240); *ZFAT* (666)	C/T	0.65	1.41 (1.19-1.66)	5.8×10^{-5}	1.00 (0.81-1.22)	.99	1.67 (1.11-2.52)	.01	1.11 (0.92-1.33)	.27	1.26 (1.12-1.43)	2.1×10^{-4}
rs1408793	9	12,628	*PTPRD* (2,000); *TYRP1* (65)	A/C	0.21	0.68 (0.57-0.82)	5.1×10^{-5}	0.94 (0.73-1.21)	.62	1.67 (1.02-2.72)	.04	1.06 (0.85-1.32)	.62	0.81 (0.71-0.94)	5.00×10^{-3}

(Continued)

TABLE E2. *(Continued)*

					Stage 1		Stage 2							
					EGEA (n = 1660)		SLSJ study (n = 1138)		MRC studies (n = 446)		Meta-analysis		Overall meta-analysis	
SNP	Chromosome*	Position (kb)	Nearest genes (kb distance)†	Alleles‡	EAF§	OR (95% CI)‖	*P* value#	OR (95% CI)‖	*P* value#	OR‖	*P* value#	OR (95% CI)‖	P_{stage2}¶	OR (95% CI)‖	P_{meta}**
rs1433831	9	18,869	*ADAMTSL1*	A/G	0.34	0.72 (0.61-0.84)	5.7×10^{-5}	1.13 (0.91-1.41)	.27	0.79 (0.54-1.14)	.20	1.03 (0.85-1.24)	.78	0.83 (0.74-0.94)	4.0×10^{-3}
rs2309394	9	71,671	*FXN*	A/G	0.56	0.72 (0.61-0.85)	8.4×10^{-5}	1.02 (0.84-1.24)	.83	1.49 (1.01-2.18)	.04	1.1 (0.93-1.30)	.27	0.88 (0.78-0.99)	.04
rs9314854	9	71,686	*FXN*	C/T	0.52	1.37 (1.18-1.61)	6.3×10^{-5}	1.18 (0.97-1.44)	.10	0.85 (0.56-1.28)	.43	1.11 (0.93-1.33)	.24	1.25 (1.11-1.41)	1.6×10^{-4}
rs7870295	9	71,686	*FXN*	A/G	0.55	1.42 (1.21-1.66)	1.6×10^{-5}	1.15 (0.94-1.40)	.17	0.79 (0.52-1.19)	.25	1.07 (0.89-1.28)	.46	1.25 (1.11-1.41)	2.0×10^{-4}
rs4745580	9	71,690	*FXN*	C/T	0.55	1.41 (1.21-1.65)	1.9×10^{-5}	1.16 (0.95-1.41)	.14	1.19 (0.78-1.79)	.42	0.89 (0.75-1.07)	.21	0.8 (0.71-0.90)	1.5×10^{-4}
rs7859021	9	71,694	*FXN*	A/G	0.48	0.73 (0.63-0.86)	8.6×10^{-5}	0.84 (0.69-1.02)	.07	1.14 (0.65-1.99)	.64	1.02 (0.73-1.44)	.89	1.41 (1.15-1.72)	1.1×10^{-3}
rs2771064	9	103,921	*PLPPR1*	A/G	0.89	1.68 (1.30-2.16)	6.4×10^{-5}	0.96 (0.63-1.48)	.86	1.20 (0.82-1.75)	.35	1.19 (1.00-1.42)	.05	0.90 (0.80-1.02)	.09
rs13286744	9	117,444	*TMEM268* (35); *TNFSF15* (103)	C/T	0.42	0.71 (0.60-0.84)	3.6×10^{-5}	1.19 (0.98-1.45)	.09	0.88 (0.49-1.57)	.66	1.17 (0.87-1.57)	.31	1.46 (1.19-1.79)	2.4×10^{-4}
rs12000625	9	126,686	*DENND1A*	A/C	0.93	1.78 (1.35-2.34)	4.4×10^{-5}	1.29 (0.91-1.82)	.15	0.95 (0.58-1.57)	.85	0.86 (0.67-1.10)	.22	1.20 (1.02-1.41)	.03
rs17515236	10	9,540	*GATA3* (1,423); *CELF2* (1,507)	A/G	0.87	1.57 (1.26-1.95)	5.4×10^{-5}	0.83 (0.63-1.10)	.2	0.86 (0.51-1.45)	.57	0.73 (0.54-0.97)	.03	1.24 (1.05-1.47)	.01
rs11005971	10	59,649	*MIR3924* (584); *IPMK* (302)	C/T	0.15	1.64 (1.33-2.01)	3.4×10^{-6}	0.68 (0.48-0.96)	.03	1.13 (0.67-1.9)	.65	1.38 (1.03-1.84)	.03	0.80 (0.68-0.95)	.01
rs7101032	10	59,672	*MIR3924* (607); *IPMK* (279)	C/T	0.85	0.61 (0.49-0.74)	2.1×10^{-6}	1.50 (1.07-2.12)	.02	0.69 (0.38-1.26)	.23	0.89 (0.67-1.17)	.41	1.24 (1.01-1.54)	.04
rs12771265	10	71,411	*C10orf35* (18); *COL13A1* (150)	A/G	0.52	1.38 (1.19-1.61)	3.2×10^{-5}	0.87 (0.72-1.05)	.14	1.04 (0.71-1.52)	.85	0.90 (0.76-1.07)	.22	1.14 (1.02-1.28)	.02
rs4746906	10	71,516	*C10orf35* (122); *COL13A1* (45)	C/T	0.16	1.53 (1.24-1.90)	9.1×10^{-5}	0.93 (0.74-1.17)	.52	0.73 (0.44-1.21)	.22	0.89 (0.72-1.10)	.28	1.16 (1.00-1.35)	.05
rs618929	11	96,229	*JRKL* (102); *CNTN5* (2,662)	A/G	0.54	1.39 (1.19-1.63)	3.2×10^{-5}	0.81 (0.66-1.00)	.05	1.23 (0.87-1.76)	.24	0.90 (0.75-1.08)	.27	1.16 (1.03-1.30)	.02
rs4475974	12	18,262	*RERGL* (19); *PIK3C2G* (152)	G/T	0.60	0.71 (0.61-0.83)	2.5×10^{-5}	1.05 (0.86-1.29)	.62	0.73 (0.5-1.06)	.10	0.97 (0.81-1.16)	.73	0.81 (0.72-0.92)	7.1×10^{-4}
rs4457807	12	18,306	*RERGL* (63); *PIK3C2G* (109)	A/G	0.39	0.70 (0.60-0.83)	2.2×10^{-5}	1.09 (0.92-1.29)	.32	1.16 (0.78-1.72)	.45	1.10 (0.94-1.29)	.23	0.89 (0.79-0.99)	.04
rs10783599	12	54,186	*CALCOCO1* (65); *HOXC13* (147)	A/G	0.32	1.39 (1.18-1.62)	5.1×10^{-5}	0.80 (0.66-0.97)	.02	1.24 (0.81-1.89)	.33	0.86 (0.72-1.02)	.09	1.11 (0.99-1.25)	.07
rs7315435	12	69,284	*CPM*	A/G	0.80	1.45 (1.21-1.75)	7.4×10^{-5}	1.03 (0.80-1.34)	.81	1.04 (0.64-1.69)	.87	1.03 (0.82-1.30)	.77	1.27 (1.10-1.47)	1.1×10^{-3}
rs2434080	12	106,454	*C12orf75* (689); *NUAK1* (3)	A/G	0.59	0.72 (0.62-0.85)	5.0×10^{-5}	1.04 (0.84-1.28)	.73	0.90 (0.58-1.41)	.66	1.01 (0.84-1.22)	.90	0.83 (0.74-0.94)	2.4×10^{-3}
rs1327751	13	75,282	*LINC00347* (150); *TBC1D4* (577)	C/T	0.11	1.64 (1.28-2.11)	9.7×10^{-5}	1.05 (0.74-1.47)	.79	0.89 (0.52-1.52)	.67	1.00 (0.75-1.33)	.99	1.33 (1.10-1.60)	3.2×10^{-4}

(Continued)

TABLE E2. *(Continued)*

						Stage 1		Stage 2							
						EGEA (n = 1660)		SLSJ study (n = 1138)		MRC studies (n = 446)		Meta-analysis		Overall meta-analysis	
SNP	Chromosome*	Position (kb)	Nearest genes (kb distance)†	Alleles‡	EAF§	OR (95% CI)‖	*P* value#	OR (95% CI)‖	*P* value#	OR‖	*P* value#	OR (95% CI)‖	P_{stage2}¶	OR (95% CI)‖	P_{meta}**
rs8030108	15	98,056	*SPATA8* (726); *ARRDC4* (448)	C/T	0.44	0.74 (0.64-0.86)	6.4×10^{-5}	1.02 (0.84-1.23)	.84	1.37 (0.94-1.99)	.10	1.08 (0.91-1.28)	.36	0.87 (0.78-0.97)	.02
rs13338087	16	4,105	*ADCY9*	G/T	0.39	1.41 (1.20-1.65)	2.0×10^{-5}	0.84 (0.69-1.02)	.08	0.84 (0.56-1.25)	.39	0.84 (0.70-1.00)	.05	1.12 (1.00-1.26)	.05
rs12929999	16	76,915	*MIR4719* (13); *MON1B* (309)	C/T	0.19	1.49 (1.22-1.81)	7.2×10^{-5}	1.09 (0.85-1.40)	.5	0.94 (0.62-1.43)	.78	1.05 (0.85-1.30)	.66	1.27 (1.10-1.47)	1.2×10^{-3}
rs11647758	16	89,094	*CBFA2T3* (51); *ACSF3* (66)	C/T	0.82	1.53 (1.26-1.86)	2.3×10^{-5}	1.10 (0.88-1.38)	.38	1.04 (0.66-1.64)	.86	1.09 (0.89-1.33)	.39	1.3 (1.13-1.49)	2.7×10^{-4}
rs2571225	18	55,462	*ATP8B1*	C/T	0.72	0.68 (0.57-0.81)	2.0×10^{-5}	0.96 (0.78-1.18)	.67	0.69 (0.47-1.03)	.07	0.89 (0.74-1.07)	.22	0.77 (0.68-0.88)	8.3×10^{-5}
rs6056732	20	9,577	*PAK5*	A/C	0.91	0.56 (0.42-0.73)	3.2×10^{-5}	0.73 (0.52-1.03)	.07	0.55 (0.28-1.06)	.07	0.68 (0.51-0.93)	.01	0.61 (0.50-0.75)	2.5×10^{-6}
rs6056733	20	9,577	*PAK5*	A/G	0.12	1.62 (1.28-2.06)	7.2×10^{-5}	1.28 (0.98-1.67)	.07	1.80 (1.06-3.07)	.03	1.37 (1.08-1.74)	.01	1.49 (1.26-1.77)	3.7×10^{-6}
rs8117366	20	52,365	*ZNF217* (65); *BCAS1* (439)	G/T	0.81	1.51 (1.24-1.84)	3.2×10^{-5}	0.94 (0.73-1.21)	.64	1.47 (0.88-2.46)	.14	1.03 (0.82-1.29)	.81	1.28 (1.11-1.49)	9.5×10^{-4}
rs6013784	20	52,365	*ZNF217* (65); *BCAS1* (439)	C/T	0.21	0.68 (0.57-0.82)	5.2×10^{-5}	1.08 (0.84-1.40)	.54	0.68 (0.41-1.13)	.14	0.99 (0.79-1.24)	.91	0.79 (0.68-0.91)	1.3×10^{-3}
rs4389378	20	54,592	*CBLN4* (12); *MC3R* (232)	A/G	0.24	0.72 (0.62-0.85)	6.8×10^{-5}	0.93 (0.75-1.15)	.51	1.12 (0.76-1.63)	.57	0.97 (0.80-1.17)	.77	0.82 (0.72-0.92)	1.1×10^{-3}

*Chromosome number at which the SNP is located.
†Gene where the SNP is located is indicated; otherwise, the genes on either side of the SNP are indicated together with the distance between the SNP and gene boundary (using Build 37.3).
‡Baseline/effect allele.
§Effect allele frequency.
‖OR associated with the effect allele under an additive model along with the 95% CI associated with the OR.
¶*P* value associated with the Wald test of meta-analyzed SNP effect in the stage 2 data sets (SLSJ and MRC studies).
#*P* value associated with the Wald test statistic of SNP effect.
***P* value associated with the Wald test of the meta-analyzed SNP effect in the 3 data sets (EGEA and SLSJ and MRC studies).

TABLE E3. List of genes (gene set 2) tested for their relationship with *ADGRV1* (gene set 1) by using GRAIL

Gene symbol*	Nearest SNP (kb distance)†	Cytogenetic band	Position (kb)‡	GWAS *P* values in EGEA	P_{GRAIL}
CHD7	rs4237038	8q12	61,731	7.3×10^{-5}	3.2×10^{-3}
ATP8B1	rs2571225	18q21	55,462	2.0×10^{-5}	.016
DNAH5	rs795540	5p15	13,818	8.3×10^{-5}	.084
NDUFAF7	rs3819892 (34)	2p22	37,510	8.1×10^{-5}	1
PRKD3	rs3819892	2p22	37,510	8.1×10^{-5}	1
MYH15	rs3900940	3q13	108,148	9.3×10^{-5}	1
NLN	rs7711329 (1)	5q12	65,126	8.3×10^{-5}	1
PDE1C	rs11763324	7p14	32,171	2.3×10^{-5}	1
POM121L12	rs1880617 (33)	7p12	53,138	9.5×10^{-5}	1
TES	rs10279056 (4)	7q31	115,847	1.0×10^{-5}	1
CAV2	rs17138756 (4)	7q31	116,135	3.4×10^{-5}	1
LINC-PINT	rs10253511 (3)	7q32	130,797	4.8×10^{-5}	1
MKLN1	rs10253511	7q32	130,797	4.8×10^{-5}	1
ADAMTSL1	rs1433831	9p22	18,869	$5.7 \times 10\times^{5}$	1
FXN	rs2309394	9q21	71,671	8.4×10^{-5}	1
PRKACG	rs2309394 (42)	9q21	71,671	8.4×10^{-5}	1
TJP2	rs7859021 (42)	9q21	71,694	8.6×10^{-5}	1
PLPPR1	rs2771064	9q31	103,921	6.4×10^{-5}	1
TMEM268	rs13286744 (36)	9q32	117,444	3.6×10^{-5}	1
DENND1A	rs12000625	9q33	126,686	4.4×10^{-5}	1
JMJD1C	rs10995495	10q21	65,062	4.7×10^{-5}	1
C10orf35	rs12771265 (18)	10q22	71,411	3.2×10^{-5}	1
COL13A1	rs4746906 (46)	10q22	71,516	9.1×10^{-5}	1
RERGL	rs4475974 (19)	12p12	18,262	2.5×10^{-5}	1
CPM	rs7315435	12q15	69,284	7.4×10^{-5}	1
MDM2	rs7315435 (45)	12q15	69,284	7.4×10^{-5}	1
NUAK1	rs2434080 (3)	12q23	106,454	5.0×10^{-5}	1
ADCY9	rs13338087	16p13	4,105	2.0×10^{-5}	1
PAK5	rs6056732	20p12	9,577	3.2×10^{-5}	1
CBLN4	rs4389378 (12)	20q13	54,592	6.7×10^{-5}	1

*Genes are located at most 50 kb apart from SNPs with *P* values of 10^{-4} or less in the stage 1 EGEA data set.
†Distance (in kilobases; build 37.3) of the nearest SNP to each gene is indicated in parentheses when the SNP lies outside of a gene.
‡SNP position in kilobases (Build 37.3).

TABLE E4. Number of SNP pairs tested for the 3 selected gene pairs and multiple-testing corrected threshold

Gene set 1	Gene set 2	No. of SNP pairs tested	Effective no. of independent SNP pairs tested	Total effective no. of independent SNP pairs tested*
ADGRV1	*CHD7*	808	110	
ADGRV1	*ATP8B1*	1450	223	688
ADGRV1	*DNAH5*	3066	355	

*Total effective number of independent SNP pairs tested is estimated by the sum of the effective number of independent SNP pairs tested by gene pair over all gene pairs examined. The multiple-testing corrected threshold is thus equal to the 5% type I error divided by the total effective number of independent SNP pairs tested ($T = 7.3 \times 10^{-5}$).

TABLE E5. Stratified analysis on asthma for the *ADGRV1* SNP significantly associated with atopy at the genome-wide level and on asthma for the interacting *ADGRV1*/*DNAH5* SNP pair significantly associated with atopy

					Stage 1		Stage 2		Overall meta-analysis	
SNPs	**Gene***	**Alleles†**	**MAF‡**	**Group**	**OR (95% CI)§**	***P* value‖**	**OR (95% CI)§**	**P_{stage2} value¶**	**OR (95% CI)§**	**P_{meta} value#**
rs4916831	*ADGRV1*	A/G	0.44	Asthmatic patients	0.67 (0.48-0.92)	.01	0.69 (0.53-0.89)	3.9×10^{-3}	0.68 (0.56-0.83)	1.4×10^{-4}
				Nonasthmatic subjects	0.69 (0.57-0.84)	1.8×10^{-4}	0.72 (0.55-0.94)	1.4×10^{-2}	0.70 (0.60-0.82)	7.8×10^{-6}
				$P_{Cochran}$ value**		.86		.82		.82

					Stage 1		Stage 2		Overall meta-analysis	
					Interaction effect		**Interaction effect**		**Interaction effect**	
SNPs	**Genes***	**Alleles†**	**MAF‡**	**Group**	**OR (95% CI)§**	**P_{int}‖**	**OR (95% CI)§**	**$P_{stage2\text{-}int}$ value¶**	**OR (95% CI)§**	**$P_{meta\text{-}int}$ value#**
rs17554723	*ADGRV1*	A/G	0.33	Asthmatic patient	0.77 (0.47-1.24)	.28	0.73 (0.48-1.12)	.15	0.74 (0.54-1.02)	.07
rs2134256	*DNAH5*	T/C	0.25	Nonasthmatic subjects	0.64 (0.46-0.89)	7.3×10^{-3}	0.53 (0.34-0.82)	4.2×10^{-3}	0.60 (0.46-0.78)	1.1×10^{-4}
				$P_{Cochran}$ value**		.55		.29		.30

The 2 groups contained 1354 asthmatic patients (628, 549, and 177 for EGEA and the SLSJ and MRC studies, respectively) and 1849 nonasthmatic subjects (997, 583, and 269 for EGEA and the SLSJ and MRC studies, respectively).

*Gene symbol.

†Major allele/minor allele.

‡Minor allele frequency.

§OR for single-SNP effect (top) or interaction between SNPs (bottom) associated with atopy assuming an additive model.

‖*P* and P_{int} are *P* values associated with the Wald test of SNP effect (top) and interaction effect (bottom).

¶P_{stage2} and $P_{stage2\text{-}int}$ are *P* values associated with the Wald test of meta-analyzed single-SNP effect (top) and interaction between SNPs (bottom) in the stage 2 data sets (SLSJ and MRC studies).

#P_{meta} and $P_{meta\text{-}int}$ are *P* values associated with the Wald test of meta-analyzed single-SNP effect (top) and interaction between SNPs (bottom) in the 3data sets (EGEA and SLSJ and MRC studies).

***P* value associated with the Cochran Q test for homogeneity between the 2 groups of asthmatic and nonasthmatic subjects.

TABLE E6. Functional annotations of *ADGRV1* and *DNAH5* SNPs

Gene	SNP	Position on chromosome 5	Variant location	r^2 value with the significant SNP(s) in our analysis	Regulatory elements			
					Promoter histone marks	Enhancer histone marks	DNaseI hypersensitivity site	Transcription factor binding sites
ADGRV1	***rs4916831***	90,916,459	Intronic	—	No	No	No	SOX, TCF12
	rs4637585	90,889,700	Intronic	0.8	No	Yes (lung)	No	DBX, FOXO, GATA, MEF2, NCX, PAX, PDX, SOX, ZFP
	rs10045202	90,890,830	Intronic	0.8	No	No	No	NERF1A
	rs4244205	90,893,069	Intronic	0.8	No	No	No	DOBOX4
	rs4580808	90,894,483	Intronic	0.82	No	Yes (lung)	No	PAX4, SREBP, SIN3AK20
	rs4496735	90,894,600	Intronic	0.82	No	No	No	EVI1, OSF2, SIX5
	rs10035307	90,906,182	Intronic	0.83	No	No	No	FOX, SP2
	rs6858917	90,909,376	Intronic	0.96	No	Yes (skin, lung)	No	AHR, CTCF, RAD21, SMAD
	rs6889986	90,911,582	Intronic	0.99	No	Yes (lung)	No	ARID3A, CDX, DBX1, FOXL1, HOXB13, HOXB9, HOXD10, MEF2, NCX, PBX1, POU2F2, POU3F2, SOX, TATA, TEF, ZFP105
	rs10078568	90,915,831	Intronic	0.8	No	No	No	CDX
	rs17554723	90,793,497	Intronic	—	No	No	No	—
	rs11745546	90,794,895	Intronic	1	No	No	No	CCNT2, E2F, MZF1
	rs13186025	90,827,511	Intronic	0.8	No	No	No	DMRT2, HDAC2
DNAH5	***rs2134256***	13,768,544	Intronic	—	No	No	No	SMC3, TCF12
	rs6862904	13,750,128	Intronic	0.83	No	No		MRG, NRSF, PBX3, TGIF1, ZNF143
	rs11745096	13,750,128	Intronic	0.84	Yes (blood)	Yes (blood)	No	—
	rs34789506	13,757,027	Intronic	0.93	No	No	No	GR
	rs11742383	13,762,604	Intronic	0.95	No	No	No	GR
	rs13156606	13,765,320	Intronic	0.93	No	No	No	BATF, HDAC2, IRF
	rs60499013	13,772,775	Intronic	0.94	No	No	No	CDX, DMRT2, DMRT3, DBX1, FOXA, HDAC2, HOXA10, HOXB13, HOXD10, NF, PAX, STAT

The *ADGRV1* rs4916831 SNP is significantly associated with atopy, and the *ADGRV1* rs17554723 and *DNAH5* rs2134256 SNPs show significant interaction in atopy (in boldface). These 3 SNPs and their proxies ($r^2 \geq 0.8$ by using the 1,000 Genomes Project database) were mapped to regulatory elements by using the HaploReg v4.1 tool (http://archive.broadinstitute.org/mammals/haploreg/haploreg.php).

CHAPITRE V – Analyses pangénomiques d'interactions gène-exposition au tabagisme passif dans l'enfance dans le délai de survenue de l'asthme

1. Résumé

Les études d'association à l'échelle du génome (GWAS) de l'asthme ont permis d'identifier de nouveaux loci associés à l'asthme, mais comme pour d'autres maladies complexes, les variants génétiques de ces loci ne représentent qu'une partie de la composante familiale de l'asthme. Cela pourrait être en partie dû à l'hétérogénéité phénotypique et génotypique de la maladie [223]. L'asthme présente un large spectre de manifestations cliniques, dans lequel l'âge de début joue un rôle important. Les méthodes d'analyse de survie peuvent permettre de prendre en compte la variabilité de l'âge d'apparition de l'asthme dans la définition de la maladie. De plus, considérer des mécanismes complexes comme des interactions gène-environnement peut conduire à identifier de nouveaux variants génétiques qui ont des effets marginaux faibles et qui confèrent un risque d'asthme seulement en présence d'une exposition environnementale.

Cette partie de ma thèse avait pour objectif d'identifier des gènes influençant le délai de survenue de l'asthme et en particulier, les gènes pouvant interagir avec l'exposition au tabagisme passif pendant la petite enfance dans le délai de survenue de l'asthme dans l'enfance.

La recherche d'interactions de variants génétiques avec le tabagisme passif dans le délai de survenue de l'asthme fait suite à une méta-analyse d'associations pangénomiques de ce délai de survenue portant sur neuf études internationales indépendantes et incluant 13 886 sujets dont 5 462 asthmatiques avec un large spectre d'âge de début de la maladie,

permettant ainsi d'en étudier la variabilité, et faisant partie du consortium européen sur l'asthme GABRIEL.

Une GWAS du délai d'apparition de l'asthme a été réalisé séparément dans chaque étude à l'aide d'un modèle de survie en considérant, pour les sujets asthmatiques, l'âge déclaré de début d'asthme et pour les sujets non-asthmatiques, l'âge au moment de l'examen, puis les résultats de ces GWAS ont ensuite été combinés par méta-analyse. Cette étude a permis d'identifier cinq loci au seuil génome-entier dont un nouveau locus en 16q12 incluant des variants qui corrèlent avec l'expression de *CYLD* (cylindromatosis turban tumor syndrome gene) et *NOD2* (nucleotide-binding oligomerization domain 2), deux gènes candidats pour l'asthme ; et de confirmer l'implication de quatre autres locus : 2q12, 6p21, 9p24 et 17q12-q21. Pour déterminer si ces cinq régions présentaient des signaux d'association distincts, nous avons conduit des analyses d'association conditionnelles séparément dans chacune de ces régions. Ces analyses ont identifié deux signaux additionnels dans les régions 9p24 et 17q12-q21. De plus, nous avons montré 1) que les SNPs des régions 9p24 et 17q12-q21 étaient associés à une apparition plus précoce de l'asthme (~6 ans) alors que ceux en 16q12 étaient associés à une apparition plus tardive de la maladie (~12 ans) ; et 2) qu'être porteur d'un grand nombre d'allèles à risque à ces sept loci était associé à une apparition plus précoce de l'asthme (4 ans pour les porteurs de six à huit allèles vs 9-12 ans pour les porteurs d'un ou deux allèles). Enfin, pour chaque individu, nous avons calculé un score polygénique de risque génétique correspondant à la somme, sur l'ensemble des sept SNPs distincts détectés au seuil génome entier, de la valeur du génotype du SNP (sous un codage additif) pondéré par son effet estimé sur le délai de survenue de l'asthme en méta-analyse. Nous avons ainsi détecté une forte association entre l'âge de début de l'asthme et le score de risque génétique avec une médiane d'âge de début de 10 ans au premier quintile Q1 de la distribution du score de risque génétique et de 6 ans au cinquième quintile Q5. Cette étude à laquelle j'ai contribué a été publiée dans le *Journal of Allergy and Clinical Immunology* (disponible en annexe).

Parmi les neuf études incluses dans ce travail, cinq disposaient de données sur l'exposition au tabagisme passif pendant la petite enfance (ELTS), incluant 8 273 sujets (2 874 asthmatiques et 5 399 non-asthmatiques). Considérer un type d'asthme spécifique, comme l'asthme dans l'enfance (apparaissant avant l'âge de 16 ans) qui peut résulter de facteurs étiologiques distincts de l'asthme de l'adulte, peut permettre d'accroitre la

puissance de détection de facteurs génétiques. Mon travail a consisté à mettre en évidence des gènes interagissant avec l'ELTS dans le délai de survenue de l'asthme dans l'enfance par une méta-analyse des cinq études d'interactions gène-environnement sur l'ensemble du génome en utilisant un modèle de survie. Dans la mesure où des effets marginaux des SNPs avaient été détectés par l'étude pangénomique mentionnée ci-dessus, nous avons choisi de nous focaliser sur les tests d'interaction GxE à 1ddl, plus puissants pour détecter des interactions avec des SNPs sans effets marginaux détectables. Des études d'associations pangénomiques ont d'abord été conduites séparément chez les sujets exposés et les non exposés dans chaque étude, puis une méta-analyse a été réalisée dans chacune des deux strates définies selon l'exposition au tabac (sujets exposés / non exposés). Les résultats de ces deux méta-analyses de GWAS ont ensuite été combinés selon la méthode proposée par Aschard *et al* [145], afin d'obtenir les statistiques des tests d'interaction GxE et les estimations des effets d'interaction. Un signal a été détecté au seuil génome entier au niveau du locus 13q21. Trois autres loci ont également été détectés à un seuil suggestif de $P \leq 5 \times 10^{-6}$: 2p22, 14q22, 20p12. Ces résultats sont corroborés dans les quatre loci par la consistence des résultats des cinq études dans chaque strate de sujets exposés/non exposés (tests d'hétérogénéité selon les études non significatifs ; forest plots montrés dans le papier).

Des annotations fonctionnelles détaillées ont été conduites. Celles-ci reposaient sur l'interrogation de bases de données d'expression des gènes, de bases de données épigénomiques générées par ROADMAP et ENCODE et résumées dans l'outil HaploReg v4, et de la base de données PhenoScanner qui répertorie les résultats d'études génétiques d'association avec des maladies/traits et des phénotypes moléculaires comme la méthylation de l'ADN. Nous avons interrogé différentes bases d'expression des gènes, afin d'identifier des cis-eQTLs (SNPs associés à l'expression des gènes situés à moins d'1Mb des SNPs du GWAS) dans le sang et les poumons. Le SNP au loci 2p22 s'est notamment révélé être un fort cis-eQTL du gène *CYP1B1* (cytochrome P450 family 1 subfamily B member 1) dans le sang. Aucun autre cis-eQTL n'a été trouvé aux trois autres loci. Cependant, le SNP montrant l'effet d'interaction le plus significatif (top SNP) au locus 13q21 est localisé dans un intron du gène *KLHL1* (kelch like family member 1). Ce gène est le seul gène codant pour une protéine qui se trouve dans une région d'1Mb autour de ce SNP. De même, le top SNP au locus 20p12 est situé dans un intron du gène

MACROD2 (MACRO domain containing 2) qui est le seul gène codant pour une protéine dans une région de 500kb autour de ce SNP. Le top SNP dans la région 14q22 est en fort DL avec des SNPs dans le promoteur du gène *NIN* (ninein). L'interrogation des bases de données épigénomiques montrent que les SNPs identifiés par l'analyse GxE ou des SNPs en DL avec ces SNPs colocalisent avec des marques d'histones (promoteurs ou enhancers), des régions sensibles au clivage par la DNase I, et des sites de fixation de facteurs de transcription particulièrement pertinents : CTCF aux loci 2p22, 14q22 et 20p12 et AhR/Arnt au loci 14q22. CTCF fonctionne comme une protéine activatrice de la transcription, ou comme répresseur. Il a été récemment démontré que le CTCF est un facteur qui contrôle la co-expression de gènes dans les voies respiratoires de patients asthmatiques. Ahr et Arnt jouent un rôle majeur dans la régulation de la réponse aux composants de la fumée du tabac. L'interrogation de la base de données PhenoScanner a montré que les principaux SNPs des quatre loci étaient fortement associés aux niveaux de méthylation de l'ADN dans le sang total. Des associations de méthylation de l'ADN avec les SNPs aux régions 2p22 et 14q22 ont également été observées dans les neutrophiles et les cellules immunitaires.

Une recherche extensive dans la littérature a montré que les quatre gènes - *KLHL1, CYP1B1, NIN, MACROD2* - ont une fonction biologique pertinente en lien avec l'exposition au tabac. Notamment, des modifications de la méthylation de l'ADN associées à l'exposition aux composants de la fumée de tabac ont été rapportées dans les loci où sont situés ces gènes dans différentes études épigénétiques. De plus, ces sites CpG sont situés à proximité des sites CpG associés aux SNPs interagissant avec l'ELTS identifiés par notre analyse (comme l'a indiqué la base PhenoScanner).

Une analyse de pathways biologiques (classes d'ontologie des gènes) par la méthode GSEA [59] a également mis en évidence plusieurs voies biologiques significativement enrichies en gènes interagissant avec ELTS. Onze pathways étaient enrichis en gènes interagissant avec l'ELTS (au seuil FDR $\leq$ 5%). Trois d'entre eux avaient un FWER $\leq$ 5% : la réponse de la défense immunitaire aux bactéries (GO:0042742), la phosphorylation oxydative (GO:0006119), le processus métabolique des stérols (GO:0016125). Le premier pathway inclus des gènes qui jouent un rôle important dans l'immunité innée. Le second est composé de gènes codant pour des protéines impliquées dans la fonction de la chaîne respiratoire mitochondriale. Le troisième contient des gènes

impliqués dans le métabolisme des lipides. Ces trois pathways contiennent des gènes qui ont une fonction biologique susceptible de jouer un rôle dans l'asthme et qui peut être altérée par les composants de la fumée de cigarette.

En conclusion, cette étude a identifié de nouveaux locus en interaction avec l'exposition au tabagisme passif durant la petite enfance dans le délai de survenue de l'asthme dans l'enfance. Les gènes candidats au niveau de ces loci ont des fonctions biologiquement pertinentes liées à l'exposition à la fumée de tabac. La colocalisation des variants génétiques détectés avec des éléments de régulation, leur association avec la méthylation de l'ADN dans le sang, et la présence à proximité de modifications de méthylation de l'ADN associées à l'exposition à la fumée du tabac, incitent fortement à entreprendre de nouvelles études épigénétiques et fonctionnelles afin d'élucider les mécanismes sous-jacents.

2. Article soumis

Genome-wide interaction study of early-life smoking exposure on time-to-asthma onset in childhood

Running title: gene-smoking exposure interaction in childhood asthma

Pierre-Emmanuel Sugier[1,2,3], Chloé Sarnowski[1,2], Raquel Granell[4], Catherine Laprise[5], Markus J Ege[6,7], Patricia Margaritte-Jeannin[1,2], Marie-Hélène Dizier[1,2], Cosetta Minelli[8], Miriam F. Moffatt[9], Mark Lathrop[10], William O.C.M. Cookson[9], A. John Henderson[4], Erika von Mutius[6,7,11], Manolis Kogevinas[12,13,14], Florence Demenais[1,2]*, Emmanuelle Bouzigon[1,2]*

1. Inserm, UMR-946, Genetic Variation and Human Diseases Unit, Paris, France

2. Université Paris Diderot, Sorbonne Paris Cité, Institut Universitaire d'Hématologie, Paris, France

3. Université Pierre et Marie Curie, Paris, France

4. MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, UK

5. Département des sciences fondamentales, Université du Québec à Chicoutimi, Saguenay, QC, Canada

6. Dr von Hauner Children's Hospital, Ludwig Maximilian University, Munich, Germany

7. Comprehensive Pneumology Center Munich (CPC-M), German Center for Lung Research, Munich, Germany

8. Population Health & Occupational Disease, National Heart and Lung Institute, Imperial College, London, UK

9. Section of Genomic Medicine, National Heart Lung Institute, Imperial College London, London, UK

10. McGill University and Génome Québec Innovation Centre, Montréal, Canada

11. Helmholtz Zentrum München - German Research Center for Environmental Health, Institute for Asthma and Allergy Prevention, Neuherberg, Germany

12. ISGlobal, Barcelona, Spain; Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain

13. CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain

14. Municipal Institute of Medical Research (IMIM-Hospital del Mar), Barcelona, Spain

* These authors contributed equally to this work.

Corresponding author:

Emmanuelle Bouzigon, MD, PhD

UMR-946, INSERM / Université Paris-Diderot, Institut de Génétique Moléculaire, 27 rue Juliette Dodu, 75010 Paris, France

Tel: 33 (0) 1 72 63 93 05 / Fax: 33 (0) 1 72 63 93 49

Email: emmanuelle.bouzigon@inserm.fr

ABSTRACT

Background: Asthma, a heterogeneous disease with variable age of onset, likely results from gene-by-environment interactions. Early-life tobacco smoke (ELTS) exposure is a major asthma risk factor. Only a few loci were reported to interact with ELTS exposure in asthma.

Objective : Our aim was to identify new loci interacting with ELTS exposure on time-to-asthma onset (TAO) in childhood.

Methods: We conducted a gene-environment-wide interaction analysis of ELTS exposure on time-to-asthma onset in childhood in five European-ancestry studies (totaling 8,273 subjects) using survival analysis methods. The results of all five genome-wide analyses were meta-analyzed.

Results: The 13q21 locus showed genome-wide significant interaction with ELTS exposure (P=4.3x10^{-8} for rs7334050 within *KLHL1* with consistent results across the five studies). Suggestive interactions (P<5x10^{-6}) were found at three other loci: 20p12 (rs13037508 within *MACROD2*; P=4.9x10^{-7}), 14q22 (rs7493885 near *NIN*; P=2.9x10^{-6}) and 2p22 (rs232542 near *CYP1B1*; P=4.1x10^{-6}). Functional annotations and the literature showed that the lead SNPs at these four loci are associated with DNA methylation levels in the blood and these CpG sites colocalize with DNA methylation changes triggered by tobacco smoke components, which strongly support our findings. Pathway analysis pointed out three pathways enriched in genes interacting with ELTS (defense response to bacteria, oxidative phosphorylation, sterol metabolic process). These pathways have a biological function relevant to asthma that may be altered by smoking exposure.

Conclusion: We identified novel candidate genes interacting with ELTS exposure on time-to-asthma onset in childhood. These genes have plausible biological relevance

related to tobacco smoke exposure. Further epigenetic and functional studies are needed to confirm these findings and to shed light on the underlying mechanisms.

Keywords:

gene-environment interaction, environmental tobacco smoke exposure, childhood asthma, time-to-asthma onset

INTRODUCTION

Asthma, one of the most common chronic diseases, results from the interplay between genetic and environmental factors. Asthma has variable age of onset and variable expression over the life span. It is recognized that childhood-onset asthma may be distinct from later-onset asthma and may represent a more homogeneous subgroup often associated with allergy. The genetic component of asthma is substantial but the asthma loci, identified so far, explain only a part of the genetic risk.[1] One potential reason for this missing heritability is gene-environment (GxE) interaction because some genetic variants may confer risk only in the presence of environmental exposures.

Tobacco smoke exposure in early-life is a major risk factor for asthma. Interactions between genetic variants and early-life tobacco smoke (ELTS) exposure on asthma have been first identified by genome-wide linkage scans and candidate gene studies.[2] ELTS exposure was then found to increase the risk of early-onset asthma associated with the 17q12-21 variants identified by the first asthma GWAS.[3] Recently, a gene-environment-wide interaction study (GEWIS) of childhood-onset asthma reported interactions between *in utero* exposure to maternal smoking and the 18p11 locus and between exposure to parental smoking in childhood and the 6q26 locus but none of these interactions reached genome-wide significance.[4] A subsequent GEWIS conducted for adult-onset asthma and active smoking revealed suggestive interactions on chromosomes 9p23 and 12p12.[5] Therefore, similarly to the previously reported heterogeneity of SNP effect on asthma risk according to age of onset of asthma[3,6], the SNP-smoke exposure interactions may differ between childhood-onset and later-onset asthma and may vary according to time of exposure. Only two GWAS have considered either age of asthma onset in asthmatic children[7] or time-to-asthma onset (TAO) in asthmatic and non-asthmatic subjects[8] and

have led to the discovery of new asthma loci, but no GEWIS has yet taken into account the age of asthma onset.

In order to identify new asthma risk loci, we conducted a genome-wide interaction analysis of ELTS exposure on time-to-asthma onset in childhood in five European ancestry studies.

METHODS

Study populations

The total sample consisted of 8273 subjects of European-ancestry (2874 subjects with childhood-onset asthma and 5399 non-asthmatic subjects) from five studies, which were part of the Gabriel asthma consortium.[6] The five datasets were from three population-based studies (the ALSPAC[9] birth corhort from UK, the pan-European ECRHS cohort study[10] including sixteen centres from eight countries and the cross-sectional GABRIELA[11] survey conducted in rural areas of Austria, Germany and Switzerland) and two family-based studies (the French EGEA cohort study[12] and the French-Canadian SLSJ study[13]). A detailed description of these studies can be found in the Supplementary Information. These were the only Gabriel consortium studies of sufficient sample size that had data on age of asthma onset, ELTS exposure and imputed SNP data (Table S1-A). Written informed consent was signed by all participants or by kin or guardians for minors/children. Ethical approval was obtained for each study from the appropriate institutional ethics committees (ethical approval numbers are provided in the Supplementary Information).

Definition of time to asthma onset and ELTS exposure

Definition of asthma was based on report of doctor's diagnosis and/or on standardized questionnaires, as used in previous GWAS.[6,8] Subjects with childhood asthma were those with asthma onset ≤ 16 years of age while non-asthmatics were those that never had asthma up to their last follow-up. To model TAO, we used age of asthma onset in asthmatics and, in non-asthmatics, age at last examination. ELTS exposure was defined as being exposed to maternal smoking during pregnancy and/or parental smoking in early childhood (generally, before 5 years of age).

Genotyping and imputation

SNP genotyping and imputation and quality control (QC) criteria are summarized in Table S1-B. Genotyping was carried out using the Illumina 610-Quad for all studies except for ALSPAC where the Illumina 550-Quad array was used. We used HapMap2 imputed data as available in Gabriel consortium studies. Imputation was performed as previously described.[6] We kept for analysis SNPs with imputation quality score (rsq) ≥ 0.5 and minor allele frequency ≥ 1%, making a total of 2.11 million SNPs for analysis. At the loci showing genome-wide significant interaction with ELTS exposure, we further imputed the summary statistics from the observed gene-environment interaction statistics (Z scores) using the 1000 Genomes phase 3 EUR panel as reference panel and a novel summary statistics imputation method.[14]

Gene-environment-wide interaction analysis

To maximize statistical power, we conducted a meta-analysis of the five datasets, as previously done for the time-to-asthma onset GWAS.[8] Each dataset was split in ELTS-

exposed ($ELTS^+$) and ELTS-unexposed ($ELTS^-$) subjects. A genome-wide association (GWA) analysis of TAO was conducted in each stratified dataset. The effect of individual SNPs on TAO was estimated using a Cox proportional-hazard model while adjusting for sex and principal components to correct for population stratification and assuming an additive genetic model for SNP effect. A robust sandwich estimator of the variance with cluster on family was used to take into account familial dependencies in the family studies. The complex sampling design of the GABRIELA study was taken into account by using survey regression techniques to estimate robust standard errors ('svy' command in Stata). In each $ELTS^+/ELTS^-$ stratum, the SNP effect sizes estimated from each of the five studies were combined using a fixed-effect meta-analysis with inverse variance weighting. The SNPxELTS interaction effect was estimated as the difference (D_{HR}) between the $ELTS^+$ and $ELTS^-$ combined SNP effect sizes; the variance of D_{HR} was estimated, as previously explained.[15] The test statistic for SNPxELTS interaction (D_{HR} divided by the square root of its variance) was compared to a standard normal distribution. We applied a Bonferroni correction to correct for multiple testing. A SNPxELTS interaction was declared as genome-wide significant if it reached the threshold of 5×10^{-8}; an interaction reaching the threshold of $5x10^{-6}$ was considered as suggestive. Furthermore, the consistency of SNP effect size estimates across all five studies in each $ELTS^+/ELTS^-$ stratum was assessed using the Cochran's Q homogeneity test, and the extent of heterogeneity was estimated using the I^2 statistic, which describes the percentage of variation across studies that is due to heterogeneity rather than to chance.[16] All analyses were conducted using Stata© V14.1.

In each region showing genome-wide significant interaction with ELTS exposure, we conducted approximate conditional analysis to potentially identify multiple distinct

signals. We used the genome-wide complex trait analysis (GCTA) software[17] which is based on the GWAS summary meta-analysis statistics in each ELTS$^+$/ELTS$^-$ stratum and takes into account the correlations among SNPs estimated from a reference population (here, the Hapmap2 CEU reference panel). After conducting conditional analysis in each ELTS+/ELTS- stratum by adjusting for the effect of the most significant SNP, the interaction effect and test statistic were computed as explained above. If the lead SNP-adjusted interaction effect remained significant after correction for the number of SNPs investigated in the region, a second round of conditional analysis was carried out.

Functional annotations of the loci interacting with ELTS exposure

To provide biological insight into our findings, we conducted a bioinformatic assessement of the loci detected by our genome-wide interaction analysis. At each locus, we defined a list of SNPs to be interrogated that included the most significant SNP interacting with ELTS (designated as lead SNP) and all SNPs in LD with the lead SNP (r^2 comprised between 0.5 and 1). To pinpoint the most likely candidate genes at the identified loci, we searched for cis-expression quantitative trait loci (eQTLs) within at most 1 Mb of each investigated SNP by interrogating four eQTL studies in the blood (peripheral blood[18], lymphoblastoid cell lines[19,20], monocytes[21]) and the GTEx database that contains eQTL data from many tissues.[22] To complement the eQTL analysis, we searched for missense variants potentially tagged by the interaction signals using the HaploReg v4.1 tool.[23] To get greater insight into how the genetic variants interacting with ELTS may functionally influence TAO, we investigated whether the SNPs from the aforementioned SNP set were located in the vicinity of cis-regulatory DNA elements and transcription factor (TF) binding sites, using ROADMAP/ENCODE functional genomics data generated in a wide

range of human cell types[24] and summarized in HaploReg v4.1.[23] We also conducted a search in the Phenoscanner database[25] to assess whether the SNPs were previously reported in genetic association studies with diseases and traits as well as molecular phenotypes including DNA methylation.

Pathway analysis

To identify biological pathways enriched in genes interacting with ELTS on TAO, we applied the gene-set enrichment analysis (GSEA) to the GEWIS test statistics using GenGen.[26] The SNPs were assigned to genes if they were located within 50 kb of the gene boundaries (to include regulatory regions neary the gene). GSEA derives an enrichment score to detect the gene-sets significantly enriched in genes interacting with ELTS compared to the whole genome. The gene-sets were the Gene Ontology (GO) categories provided by GenGen.[26] Statistical significance of the gene-set enrichment scores was determined by 10,000 SNPxELTS interaction statistics permutations. We computed empirical P-values and, to adjust for multiple testing, the false discovery rate (FDR) and the family-wise error rate (FWER). We used a FDR≤5% to declare statistical significance; a FWER ≤5% was further used to select the most significantly enriched pathways.

RESULTS

Figure 1 presents the flow chart of the study together with the sample size of each of the five datasets by $ELTS^{+}$/$ELTS^{-}$ stratum (total n=8273, of which 3187 were $ELTS^{+}$ and 5086 were $ELTS^{-}$).

Gene-environment-wide interaction analysis

The Manhattan plot of interaction *P*-values for the genome-wide interaction analysis of ELTS exposure on TAO is shown in Figure 2. There was little inflation in the interaction test statistics (QQ plot in Figure S1). A genome-wide significant interaction with ELTS exposure was found at the 13q21 locus (P=4.3x10^{-8} for rs7334050). Besides the lead SNP at that locus, there were five additional variants that showed suggestive interactions (1.3x10^{-7} < P <3.9x10^{-6}; Table S2 and regional plot in Figure 3). Suggestive interactions (P≤5x10^{-6}) were also observed at three other loci on chromosomes 20p12, 14q22 and 2p22. The results at all four loci are shown in Table 1 for the lead SNPs and in Table S2 for the additional SNPs. The minor allele (MAF=0.14) of the significant SNP, rs7334050, at 13q21 conferred an increased risk in ELTS exposed subjects (HR_{ELTS+} = 1.34, 95% Confidence Interval (CI), 1.19-1.52) and a decreased risk in unexposed subjects (HR_{ELTS-} = 0.85, 95% CI, 0.76-0.95); the increase in risk in the exposed group was stronger (P=2.6x10^{-6}) than the decrease in risk in the unexposed group (P=3.2x10^{-3}). The SNPxELTS interaction hazard ratio was always in the same direction for all five studies (ranging from 1.16 to 3.39). The SNP rs7334050 also showed consistent effect sizes across the five studies in each ELTS^{+}/ELTS^{-} stratum with no evidence for heterogeneity (P for Cochran's Q test >0.31; the I^2 estimates were equal to 0.0 in both ELTS^{+} and ELTS^{-} strata; Table S2 and Figure 4 for forest plot). Moreover, conditional analysis adjusting for the 13q21 genome-wide significant SNP did not show significant evidence for any additional interaction signal. All lead SNPs at the other three loci showed an opposite direction of effect according to exposure (Table 1) with consistent effect sizes across studies in both ELTS^{+}/ELTS^{-} strata (P for Cochran's Q test being greater than 0.15; Table S2, Figure S2 for forest plots).

Functional annotations of the loci interacting with ELTS exposure

The search for cis-eQTLs at the four loci detected by this GEWIS showed that the lead SNP rs232542 at 2p22 and three proxies (r^2 >0.99) were strong cis-eQTLs for the *CYP1B1* gene ($7.0x10^{-34} \leq p \leq 3.6x10^{-33}$) in the blood.[18] No cis-eQTL was found at the other three loci, which may be due to the fact that gene expression in eQTL studies is measured in basic conditions without exposure to any environmental factor. However, the 13q21 lead SNP (rs7334050) is located in an intron of *KLHL*1, which is the only protein coding gene within 1 Mb on each side of that SNP. Similarly, the best SNP (rs13037508) at 20p12 is an intronic variant in *MACROD2*, the sole protein coding gene within 500 kb of that SNP. The lead SNP (rs7493885) at 14q22 is closest to *NIN* and is in LD (r^2=0.65) with SNPs within *NIN* promoter. Moreover, the interrogated SNPs at all four loci did not tag any missense variant.

The colocalization of lead SNPs and proxies at the four loci with regulatory elements are shown in Table 2. The lead *KLHL1* SNP maps to binding sites of transcription factors (TFs) and a nearby SNP in strong LD (r^2=0.87) maps to histone marks in fetal lung and DNase I hypersensitive sites (DHSs) in hematopoietic stem cells. The lead SNPs (and proxies) at the other three loci colocalize with histone marks and/or DHSs in blood cells and the lungs and TF binding sites. Notably, these TFs include CTCF at 2p22, 14q22 and 20p12 loci and Ahr (Arhyl hydrocarbon receptor) and its partner Arnt (Arhyl hydrocarbon receptor nuclear translocator) at 14q22. CTCF functions as a transcriptional activator, repressor or insulator protein. It was recently shown that CTCF is a major driver of gene co-expression in the airways of asthmatic patients.[27] The Ahr and Arnt TFs are known to play a major role in the regulation of biological responses to tobacco smoke components. Finally, interrogation of the Phenoscanner database showed that the lead SNPs at all four

loci were strongly associated with DNA methylation levels in the whole blood (Table 2). Associatons of DNA methylation with 2p22 and 14q22 SNPs were also observed in neutrophils and immune cells.

Pathway analysis

Pathway analysis identified 11 GOs significantly enriched in genes interacting with ELTS exposure on TAO (FDR≤5%, Table 3). The three GOs with smallest FDR and with FWER<5% were: defense response to bacterium (GO:0042742), oxidative phosphorylation (GO:0006119) and sterol metabolic process (GO:0016125). The genes driving the enrichment score of each of these three GOs are shown in Table S3. The first GO includes genes encoding defensins and toll-like receptors that play a key role in innate immunity which was found to be affected by cigarette smoke.[28] The second GO is made of genes encoding proteins involved in the mitochondrial respiratory chain function, known to play a role in asthma and altered by cigarette smoking[29] while the third GO contains genes involved in lipid metabolism that is also disrupted by tobacco smoke exposure.[30]

DISCUSSION

To our knowledge, this is the first study to examine gene-by-ELTS exposure interactions on time-to-asthma onset using a genome-wide approach. We identified a significant interaction with ELTS exposure at the 13q21 locus and suggestive interactions at three other loci on chromosomes 2p22, 14q22 and 20p12. The evidence for these interactions rests on the results obtained in five large European-ancestry studies and the consistency of results across studies.

The SNP (rs7334050) showing significant interaction with ELTS exposure is located within the Kelch-like 1 (*KLHL1*) gene. *KLHL1* encodes a neuronal acting-binding protein that modulates voltage-gated calcium channels.[31] Calcium channels regulate the contraction of airway smooth muscle, which plays a key role in bronchial-hyperresponsiveness (BHR), and have an effect in cytokine production and development of airway inflammation.[32] Cigarette smoke was shown to enhance the expression of Ca2+ regulatory proteins leading to increased cell proliferation of airway smooth muscle and cytokine generation.[33] Interestingly, *KLHL1* was found to be one of the genes that had mutation frequencies in air-pollution related lung tumors significantly associated with lifetime exposure to benzo(a)pyrene.[34] It is also one of the genes with nearby DNA methylation modifications associated with prenatal exposure to drinking water arsenic in the cord blood of newborns.[35] Benzo(a)pyrene and arsenic are both known components of cigarette smoke. Interestingly, the *KLHL1* CpG site, reported in the arsenic exposure study, is in close vicinity (613 bp apart) of the CpG associated with the significant ELTS-interacting SNP (rs7334050). Finally, the *KLHL1* locus was found associated with sneeze reflex in response to bright light exposure by a GWAS [36] (although the reported SNP was not in LD with our lead SNP; r^2=0). Sneezing and coughing are reflex responses which protect the airways from various chemical challenges including smoking exposure.

Although the other three loci did not reach genome-wide significance, they harbor relevant candidates with biological function related to tobacco smoke exposure. The 20p12 lead SNP is within *MACROD2*, a gene encoding a deacetylase involved in removing ADP-ribose from mono-ADP-ribosylated proteins. This type of post-translational modification can modulate signal transduction pathways, stress pathways and DNA repair as may occur after DNA damage caused by carcinogens.[37] Moreover,

MACROD2 belonged to the epigenetic signature of cigarette smoking and was one of the few genes colocalizing with CpG sites showing methylation levels in former smokers that never returned to never-smoker levels after 30 years of smoking cessation, as reported by a large epigenome-wide association study (EWAS).[38] The 14q22 lead SNP is located nearby *NIN*, which encodes a protein with a key role in ciliogenesis.[39] Genes involved in cilia function have been previously reported to interact with tobacco smoke exposure, either in early life for *DNAH9* on BHR[2] or in childhood for *PACRG* on childhood asthma.[4] Functional annotations of the 14q22 locus indicated that a proxy of the lead SNP colocalized with the binding sites of AhR and Arnt TFs, which play a crucial role in the biological response to polyaromatic hydrocarbons, as produced by cigarette smoking. The latter TFs were shown to downregulate *NIN* expression in the RPTEC/TERT1 cell model.[40] In addition, differentially DNA methylated sites near *NIN* were found associated with maternal smoking during pregnancy by a large-scale EWAS in newborns[41]; the CpG site associated with sustained smoking during pregnancy lies nearby (23 kb apart) the CpG associated with the ELTS-interacting SNP rs8020067. Finally, at 2p22, the *CYP1B1* gene, whose expression is strongly associated with SNPs interacting with ELTS in this study, encodes a member of the cytochrome P450 superfamily of enzymes which metabolizes procarcinogens that are components of tobacco smoke. The induction of *CYP1B1* expression in response to smoke exposure was recently confirmed by an EWAS of cigarette smoking in lung cells.[42] Moreover, *CYP1B1* was one the genes, as *NIN*, with DNA methylation changes associated with maternal smoking during pregnancy.[41] It is noteworthy that one of the CpGs significantly associated with sustained smoking during pregnancy is the CpG associated with the ELTS-interacting SNP rs232542. Moreover, DNA methylation alterations annotated to *CYP1B1* were also reported to be associated

with prenatal exposure to drinking water arsenic[35] and to active smoking[38], the respective CpG sites being less than 38 kb apart from the rs232542-associated CpG . Therefore, candidate genes at all four loci identified by this study show alterations (somatic mutations and/or DNA methylation changes) related to tobacco smoke exposure, which strongly support our findings. The colocalization of smoke exposure-related methylation changes with CpG sites associated with the ELTS-interacting SNPs, detected by this study, provides further support. Further asthma studies integrating genetic and epigenetic data together with exposure to tobacco smoke in early life will enable to confirm these findings and uncover the underlying causal mechanisms, in particular the potential mediation of the interactive SNP effect with ELTS exposure on childhood asthma through DNA methylation. In addition, the pathways identified by pathway analysis contain genes with a biological function relevant to asthma and reported to be altered by smoking exposure.

To our knowledge, none of our findings have been previously reported by asthma GWAS (GWAS-Catalog of Published Genome-Wide Association Studies[43] and the Phenoscanner database[25]) or by asthma GEWIS. One of the two suggestive interactions, previously found by the sole GEWIS of smoke exposure on childhood asthma[4], was replicated in our study (P=1.5x10^{-3} at 18p11 reported for *in utero* exposure) while the other one was not (P=0.86 at 6q26 reported for childhood exposure). None of the loci identified in our study was reported by that published study except for one SNP (rs4670230) on chromosome 2p22 that modestly interacted with *in utero* exposure (P=2.1x10^{-4}) but was not correlated with our lead signal (r^2=0). The published study discovery dataset, which underwent GEWI analysis, and the current study had comparable sample sizes. However, the difference in the results might be partly due to

differences in the definition of tobacco smoke exposure (the overall proportion of exposed subjects was 13% for *in utero* exposure and 51% for childhood exposure in the published study *versus* 36% for early-life exposure in this study), the outcome examined (asthma status *versus* time-to-asthma onset) and the model used for analysis (logistic regression v*ersus* Cox model).

Up to now, few GEWIS have been conducted for asthma-related phenotypes and only one reported a genome-wide significant result for a rare variant (MAF=1.5%) interacting with dust mite exposure on lung function.[44] One of the difficulties in GEWIS is the need of large scale studies to detect significant interaction, which in turn might be affected by heterogeneity in outcome and exposure definition of the participating studies. To overcome these limitations, this study was restricted to childhood asthma and we paid attention to use a definition of ELTS exposure so that exposure was likely to occur before the onset of asthma. We meta-analyzed all five studies to maximize statistical power. Power computation indicated that our GEWIS had 50% and 70% power of detecting a SNPxELTS interaction hazard ratio of 1.7 and 1.8 respectively when the MAF is 0.20 or greater. Our findings were also supported by the consistency of the results across the five studies at all four loci. We did not detect any cross-study heterogeneity for all significant and suggestive interactions. We also verified that all studies contributed to the interaction signals: for example, for the 13q21 genome-wide significant signal, the study-specific contribution (estimated by the ratio of the interaction test statistic of each study to the meta-analyzed interaction test statistic) ranged from 9% to 45%. Similar results were observed at the other three loci detected by this GEWIS. Even though interactions of these three loci with ELTS exposure did not reach genome-wide significance, combining statistical results with biological and functional data greatly strengthened the evidence for

the potential involvement of these loci. Moreover, our study which includes more than 8000 subjects, stands among the largest studies considered to date in GEWIS of asthma phenotypes. We detected flip-flop patterns of interactions (i.e. opposite SNP effect size according to exposure), as previously reported at several loci for asthma and other diseases.[45] We did not confirm the interaction of ELTS exposure with the 17q12-21 variants[3] because, at that locus, ELTS exposure did not show a flip-flop effect but rather a synergistic effect which is best detected by other analysis approaches than the GxE interaction test used in this study.[46]

In conclusion, this study identified new loci interacting with ELTS exposure on childhood asthma. Candidate genes at these loci have biologically relevant functions related to tobacco smoke exposure. The colocalization of the ELTS-interacting variants with regulatory elements, their association with DNA methylation in the blood and the presence of nearby DNA methylation alterations associated with tobacco smoke exposure prompts for further epigenetic and functional studies to provide more insight into the underlying mechanisms.

ACKNOWLEDGEMENTS

We thank all participants who provided data for each study and to our valued colleagues who contributed to data collection, phenotypic characterization of the samples and genotyping.

This work was supported by the French National Agency for Research (ANR-11-BSV1-027-GWIS-AM, ANR-15-EPIG-0004-05 RESET-AID), Université Pierre et Marie Curie and Région Ile-de-France (DIM-SEnT) doctoral fellowships, the Fonds de Dotation Recherche en Santé Respiratoire. The UK Medical Research Council and Wellcome Trust

(grant: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. The Canada Research Chair held by C Laprise and the funding supports from Canadian Institutes of Health Research (CIHR) enabled the maintenance and continuation of the SLSJ asthma study. C. Laprise is the director of the Asthma Strategic Group of the Respiratory Health Network of the Fonds de la recherche en santé du Québec (FRSQ) and member of Allergen network. Genotyping was supported by grants from the European Commission (No. LSHB-CT-2006-018996-GABRIEL) and the Wellcome Trust (WT084703MA). GABRIELA was supported by the European Commission as part of GABRIEL (a multidisciplinary study to identify the genetic and environmental causes of asthma in the European Community), European Commission Research Grant (LSHB-CT-2006-018996) and by the European Research Council (ERS-2009-AdG, project HERA 250268).

CONFLICT OF INTEREST

EvM received personal fees from Pharma Ventures, personal fees from Peptinnovate Ltd., OM Pharma SA, European Commission/European Research Council Executive Agency, Tampereen Yliopisto, University of Turku, HAL Allergie GmbH, Ökosoziales Forum Oberösterreich, Mundipharma Deutschland GmbH & Co. KG, outside the submitted work. All other authors have no relevant conflicts of interest

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

REFERENCES

1. Demenais F, Margaritte-Jeannin P, Barnes KC, et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat Genet.* 2018;50(1):42-53.
2. Dizier MH, Nadif R, Margaritte-Jeannin P, et al. Interaction between the DNAH9 gene and early smoke exposure in bronchial hyperresponsiveness. *Eur Respir J.* 2016;47(4):1072-1081.
3. Bouzigon E, Corda E, Aschard H, et al. Effect of 17q21 variants and smoking exposure in early-onset asthma. *N Engl J Med.* 2008;359(19):1985-1994.
4. Scholtens S, Postma DS, Moffatt MF, et al. Novel childhood asthma genes interact with in utero and early-life tobacco smoke exposure. *J Allergy Clin Immunol.* 2014;133(3):885-888.
5. Vonk JM, Scholtens S, Postma DS, et al. Adult onset asthma and interaction between genes and active tobacco smoking: The GABRIEL consortium. *PLoS One.* 2017;12(3):e0172716.
6. Moffatt MF, Gut IG, Demenais F, et al. A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med.* 2010;363(13):1211-1221.
7. Forno E, Lasky-Su J, Himes B, et al. Genome-wide association study of the age of onset of childhood asthma. *J Allergy Clin Immunol.* 2012;130(1):83-90 e84.
8. Sarnowski C, Sugier PE, Granell R, et al. Identification of a new locus at 16q12 associated with time to asthma onset. *J Allergy Clin Immunol.* 2016;138(4):1071-1080.
9. Boyd A, Golding J, Macleod J, et al. Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol.* 2013;42(1):111-127.
10. Kogevinas M, Zock JP, Jarvis D, et al. Exposure to substances in the workplace and new-onset asthma: an international prospective population-based study (ECRHS-II). *Lancet.* 2007;370(9584):336-341.
11. Ege MJ, Strachan DP, Cookson WO, et al. Gene-environment interaction for childhood asthma and exposure to farming in Central Europe. *J Allergy Clin Immunol.* 2011;127(1):138-144, 144 e131-134.
12. Kauffmann F, Dizier MH, Pin I, et al. Epidemiological study of the genetics and environment of asthma, bronchial hyperresponsiveness, and atopy: phenotype issues. *Am J Respir Crit Care Med.* 1997;156(4 Pt 2):S123-129.
13. Laprise C. The Saguenay-Lac-Saint-Jean asthma familial collection: the genetics of asthma in a young founder population. *Genes Immun.* 2014;15(4):247-255.
14. Rueger S, McDaid A, Kutalik Z. Evaluation and application of summary statistic imputation to discover new height-associated loci. *PLoS Genet.* 2018;14(5):e1007371.
15. Winkler TW, Justice AE, Graff M, et al. The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study. *PLoS Genet.* 2015;11(10):e1005378.
16. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med.* 2002;21(11):1539-1558.
17. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet.* 2011;88(1):76-82.

18. Westra HJ, Peters MJ, Esko T, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet.* 2013;45(10):1238-1243.
19. Liang L, Morar N, Dixon AL, et al. A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Res.* 2013;23(4):716-726.
20. Grundberg E, Small KS, Hedman AK, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet.* 2012;44(10):1084-1089.
21. Zeller T, Wild P, Szymczak S, et al. Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. *PLoS One.* 2010;5(5):e10693.
22. Consortium TG. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science.* 2015;348(6235):648-660.
23. Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* 2016;44(D1):D877-881.
24. Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015;518(7539):317-330.
25. Staley JR, Blackshaw J, Kamat MA, et al. PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics.* 2016;32(20):3207-3209.
26. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet.* 2007;81(6):1278-1283.
27. Pascoe CD, Obeidat M, Arsenault BA, et al. Gene expression analysis in asthma using a targeted multiplex array. *BMC Pulm Med.* 2017;17(1):189.
28. Qiu F, Liang CL, Liu H, et al. Impacts of cigarette smoking on immune responsiveness: Up and down or upside down? *Oncotarget.* 2017;8(1):268-284.
29. Hoffmann RF, Zarrintan S, Brandenburg SM, et al. Prolonged cigarette smoke exposure alters mitochondrial structure and function in airway epithelial cells. *Respir Res.* 2013;14:97.
30. He BM, Zhao SP, Peng ZY. Effects of cigarette smoking on HDL quantity and function: implications for atherosclerosis. *J Cell Biochem.* 2013;114(11):2431-2436.
31. Aromolaran KA, Benzow KA, Koob MD, Piedras-Renteria ES. The Kelch-like protein 1 modulates P/Q-type calcium current density. *Neuroscience.* 2007;145(3):841-850.
32. Valverde MA, Cantero-Recasens G, Garcia-Elias A, Jung C, Carreras-Sureda A, Vicente R. Ion channels in asthma. *J Biol Chem.* 2011;286(38):32877-32882.
33. Wylam ME, Sathish V, VanOosten SK, et al. Mechanisms of Cigarette Smoke Effects on Human Airway Smooth Muscle. *PLoS One.* 2015;10(6):e0128778.
34. Yu XJ, Yang MJ, Zhou B, et al. Characterization of Somatic Mutations in Air Pollution-Related Lung Cancer. *EBioMedicine.* 2015;2(6):583-590.
35. Rojas D, Rager JE, Smeester L, et al. Prenatal arsenic exposure and the epigenome: identifying sites of 5-methylcytosine alterations that predict functional changes in gene expression in newborn cord blood and subsequent birth outcomes. *Toxicol Sci.* 2015;143(1):97-106.
36. Pickrell JK, Berisa T, Liu JZ, Segurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet.* 2016;48(7):709-717.
37. Butepage M, Eckei L, Verheugd P, Luscher B. Intracellular Mono-ADP-Ribosylation in Signaling and Disease. *Cells.* 2015;4(4):569-595.

38. Joehanes R, Just AC, Marioni RE, et al. Epigenetic Signatures of Cigarette Smoking. *Circ Cardiovasc Genet.* 2016;9(5):436-447.
39. Gupta GD, Coyaud E, Goncalves J, et al. A Dynamic Protein Interaction Landscape of the Human Centrosome-Cilium Interface. *Cell.* 2015;163(6):1484-1499.
40. Limonciel A, Moenks K, Stanzel S, et al. Transcriptomics hit the target: Monitoring of ligand-activated and stress response pathways for chemical testing. *Toxicol In Vitro.* 2015;30(1 Pt A):7-18.
41. Joubert BR, Felix JF, Yousefi P, et al. DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *Am J Hum Genet.* 2016;98(4):680-696.
42. Stueve TR, Li WQ, Shi J, et al. Epigenome-wide analysis of DNA methylation in lung tissue shows concordance with blood studies and identifies tobacco smoke-inducible enhancers. *Hum Mol Genet.* 2017;26(15):3014-3027.
43. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 2014;42(Database issue):D1001-1006.
44. Forno E, Sordillo J, Brehm J, et al. Genome-wide interaction study of dust mite allergen on lung function in children with asthma. *J Allergy Clin Immunol.* 2017;140(4):996-1003 e1007.
45. Ober C, Vercelli D. Gene-environment interactions in human disease: nuisance or opportunity? *Trends Genet.* 2011;27(3):107-115.
46. Gauderman WJ, Mukherjee B, Aschard H, et al. Update on the State of the Science for Analytical Methods for Gene-Environment Interactions. *Am J Epidemiol.* 2017;186(7):762-770.
47. Bonder MJ, Luijk R, Zhernakova DV, et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet.* 2017;49(1):131-138.
48. Chen L, Ge B, Casale FP, et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell.* 2016;167(5):1398-1414 e1324.
49. Gaunt TR, Shihab HA, Hemani G, et al. Systematic identification of genetic influences on methylation across the human life course. *Genome Biol.* 2016;17:61.

TABLE 1. Loci showing interaction ($P<5x10^{-6}$) with early-life tobacco smoke (ELTS) exposure on time-to-asthma onset in childhood

CHR	SNP	Position (kb, build 37)	Closest Gene (distance of SNP to gene in kb)	E/R	EAF	SNP x ELTS interaction[a]		SNP effect in subjects exposed to ELTS		SNP effect in subjects unexposed to ELTS	
						HR [95% CI]	2-sided *P*	HR_{ELTS} [95% CI]	2-sided *P*	HR_{ELTS-} [95% CI]	2-sided *P*
2p22	rs232542	38 328	*CYP1B1* (25)	T/C	0.31	0.75[0.67-0.85]	$4.1x10^{-6}$	0.85[0.77-0.93]	$7.0x10^{-4}$	1.12[1.04-1.22]	$1.7x10^{-3}$
13q21	**rs7334050**	**70 645**	***KLHL1* (0)**	**G/T**	**0.14**	**1.58[1.34-1.86]**	**$4.3x10^{-8}$**	1.34[1.19-1.52]	$2.6x10^{-6}$	0.85[0.76-0.95]	$3.2x10^{-3}$
14q22	rs7493885	51 317	*NIN* (20)	G/T	0.27	1.37[1.20-1.56]	$2.9x10^{-6}$	1.16[1.05-1.29]	$2.7x10^{-3}$	0.85[0.78-0.93]	$2.5x10^{-4}$
20p12	rs13037508	14 928	*MACROD2* (0)	T/A	0.37	0.68[0.58-0.79]	$4.9x10^{-7}$	0.83[0.74-0.94]	$3.6x10^{-3}$	1.23[1.12-1.35]	$7.9x10^{-6}$

CHR, chromosome; SNP, single nucleotide polymorphism; E, effect allele / R, Reference allele; EAF, effect allele frequency; HR, hazard ratio; CI: confidence interval. In bold, SNP showing interaction with ELTS at the genome-wide significance threshold ($P<5x10^{-8}$)

[a]The interaction effect size between SNP and ELTS exposure was estimated as the difference between the $ELTS^+$ and $ELTS^-$ combined SNP effect sizes obtained from the fixed-effects meta-analyses of the five studies in each $ELTS^{+/}ELTS^-$ stratum.

TABLE 2. Functional annotations of SNPs at the four loci interacting with ELTS exposure on time-to-asthma onset in childhood

Locus	SNP (r2 with lead SNP)	Position (kb)	Regulatory elements			DNA methylation			
			Histone marks[a]	DNase I hyper sensitive sites[a]	Transcription factor binding sites	CpG site (position, kb)	Tissue	[b] 2-sided *P*	Study
13q21	**rs7334050**	70 645	No	No	FOXL1, PBX-1, POU1F1	cg14273027 (70 682)	Blood	$1.1x10^{-6}$	BIOSQTL[47]
	rs73214641 (0.87)	70 653	Yes (fetal lung)	Yes (blood stem cells)					
20p12	**rs13037508**	14 928	Yes (lung, fetal lung)	Yes (stem cells)	BCL CEBPD,PAX-5 FOXD3, HDAC2, IRF, POU2F2, POU3F2, STAT, P30O, RXRA	cg04470754 (14 904)	Blood	$6.4x10^{-57}$	BIOSQTL[47]
	rs2423868 (0.71)	14 929	Yes (lung)	Yes (lung, fetal lung)	**CTCF**, RAD21, RFX5				
14q22	**rs7493885**	51 317	Yes (blood stem cells)	No		cg25597366 (51 313)	Blood	$1.2x10^{-58}$	BIOSQTL[47]
							Neutrophils	$7.1x10^{-15}$	BLUEPRINT[48]
							Monocytes	$5.8x10^{-13}$	BLUEPRINT[48]
							T cells	$8.3x10^{-9}$	BLUEPRINT[48]
							Cord blood	$4.7x10^{-28}$	ALSPAC[49]
							Blood	$3.6x10^{-46}$	ALSPAC[49]
	rs8020067 (1)	51 318	Yes (T cells, blood stem cells)	No	**AHR::ARNT**, ARNT, HBP1, PAX-4				
	rs4901062 (0.99)	51 315	Yes (lung, fetal lung)	No	**CTCF**, PITX2, RAD21				

Locus	SNP (r2 with lead SNP)	Position (kb)	Regulatory elements			DNA methylation			
			Histone marks[a]	DNase I hyper sensitive sites[a]	Transcription factor binding sites	CpG site (position, kb)	Tissue	[b] 2-sided *P*	Study
2p22	**rs232542**	38 328	Yes (fetal lung, blood cells)	No	YY1	cg02486145 (38 334)	Blood	$4.7x10^{-179}$	BIOSQTL[47]
							Neutrophils	$1.9x10^{-11}$	BLUEPRINT[48]
							Monocytes	$1.0x10^{-6}$	BLUEPRINT[48]
							T cells	$2.9x10^{-11}$	BLUEPRINT[48]
							Cord blood	$5.4x10^{-33}$	ALSPAC[49]
							Blood	$4.4x10^{-34}$	ALSPAC[49]
	rs232540 (0.99)	38 329	Yes (stem cells)	Yes (blood cells)	**CTCF**, ELTS, EVI-1, MYF, PEBP, RAD21, SMC3, TAL1				

r^2:linkage disequilibrium measure between a SNP and lead SNP (in bold) at a locus; kb: kilobase (build 37)

[a]Histone marks represent promoters or enhancers. When regulatory elements colocalized with SNPs in tissues biologically relevant to asthma (blood cells and/or lung tissue), "Yes" is indicated in the corresponding column (data retrieved from Haploreg v4.1[23])

[b] *P* is the *P*-value for association of SNP with DNA methylation levels at a CpG site (data retrieved from the Phenoscanner database[25]).

TABLE 3. Gene ontology categories enriched in genes interacting with ELTS exposure on time-to-asthma onset in childhood

Gene Ontology (GO)	GO Type	Gene set size (number of genes)	Gene Ontology names	Empirical *P*-value	FDR	FWER
GO:0042742	P	67	Defense response to bacterium	$<10^{-4}$	0.001	0.001
GO:0006119	P	52	Oxidative phosphorylation	0.0007	0.01	0.03
GO:0016125	P	76	Sterol metabolic process	$<10^{-4}$	0.01	0.03
GO:0050962	P	28	Detection of light stimulus involved in sensory perception	0.0002	0.02	0.09
GO:0009583	P	30	Detection of light stimulus	0.0005	0.02	0.11
GO:0006873	P	199	Cellular ion homeostasis	0.0004	0.03	0.16
GO:0003729	F	42	mRNA binding	0.0007	0.03	0.15
GO:0051338	P	185	Regulation of transferase activity	0.001	0.03	0.25
GO:0016811	F	42	Hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides	0.001	0.03	0.21
GO:0007005	P	61	Mitochondrion organization	0.001	0.03	0.24
GO:0010563	P	20	Negative regulation of phosphorus metabolic process	0.001	0.03	0.29

P=biological process, F=molecular function; Empirical *P*-value: *P*-value for enrichment score estimated by 10,000 permutations of SNPxELTS test statistics: FDR:false discovery rate; Family-wise error rate.

Figure Legends

FIGURE 1. Analysis strategy of the genome-wide interaction study of time-to-asthma onset in childhood**.** This figure shows the number of subjects (N) by study in each early-life tobacco smoke (ELTS) exposure stratum.

FIGURE 2. Manhattan plot of the genome-wide interaction analysis for time-to-asthma-asthma onset in childhood**.** The x axis represents chromosomal location and the y axis represents $-\log_{10}(P)$ for tests of interaction between individual SNPs and ELTS exposure on time-to-asthma onset. The red horizontal line denotes the genome-wide significant threshold of $p=5x10^{-8}$ and the black dashed horizontal line a suggestive threshold of $p=5x10^{-6}$.

FIGURE 3. Regional plot of the 13q21 region. The x axis represents physical location in megabase (build 37) and the Y axis represents $-\log_{10}(P)$ for tests of interaction between individual SNPs and ELTS exposure on time-to-asthma onset. The rs ID is shown for the lead SNP in the region. For remaining SNPs, the color indicates the r^2 measure of linkage disequilibrium (LD) with the lead SNP.

FIGURE 4. Forest Plot of the 13q21 lead SNP (rs7334050) according to early-life tobacco smoke (ELTS) exposure. Hazard ratios (HR) and 95% Confidence Intervals (CI) are plotted by study and by ELTS exposure stratum. The combined HR estimates over all five studies in each stratum and the combined SNP by ELTS exposure interaction HR are plotted as a diamond.

FIGURE 1. Analysis strategy of the genome-wide interaction study of time-to-asthma onset in childhood

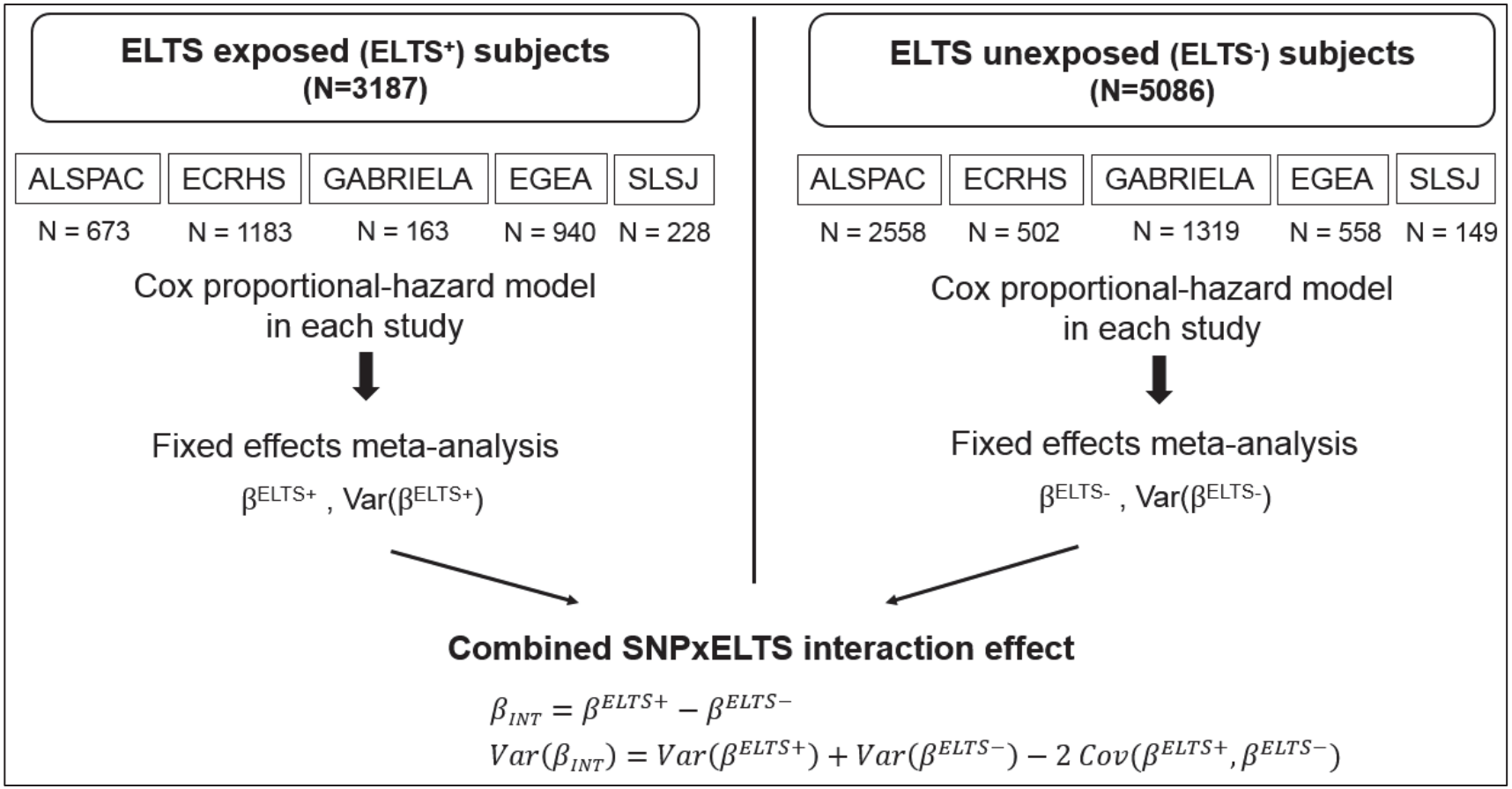


FIGURE 2. Manhattan plot of the genome-wide interaction analysis for time-to-asthma-asthma onset in childhood

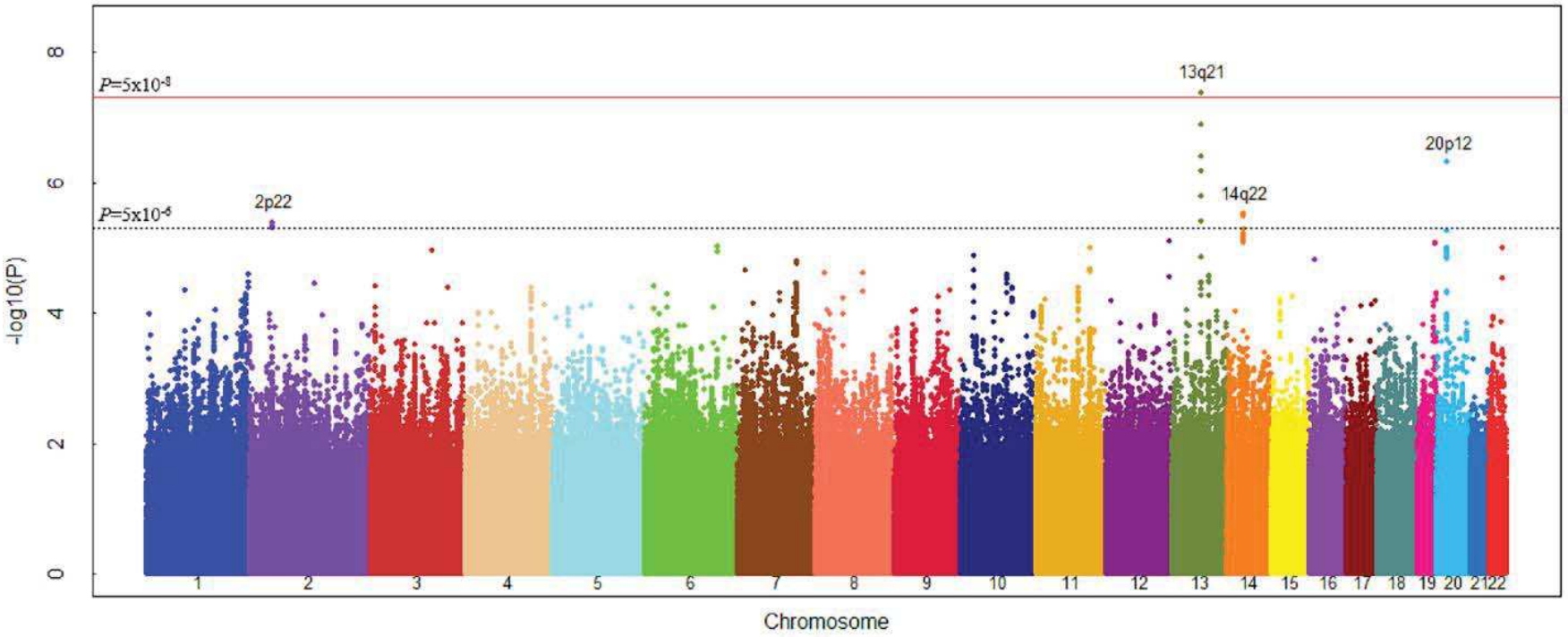


FIGURE 3. Regional plot of the 13q21 region

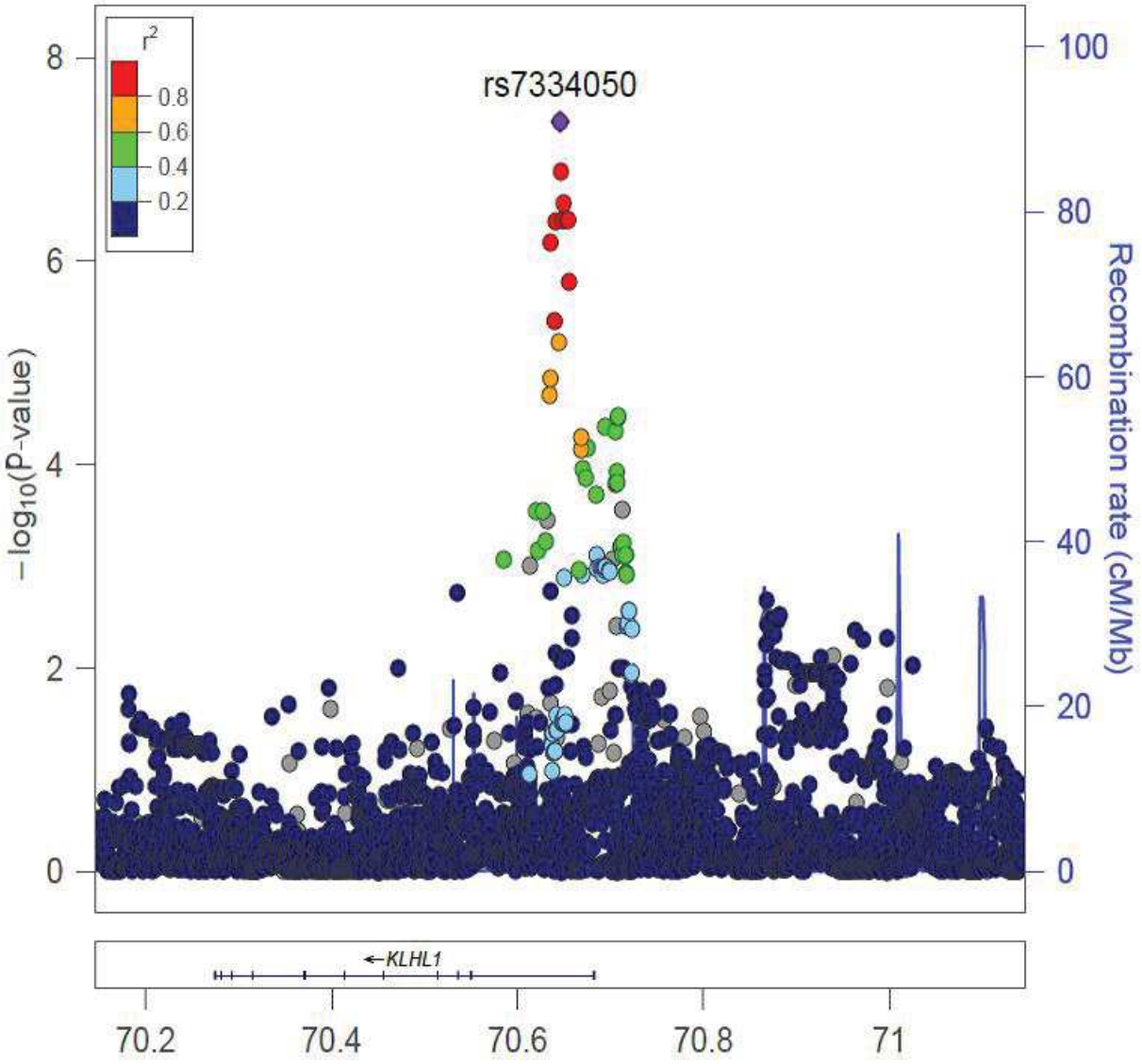
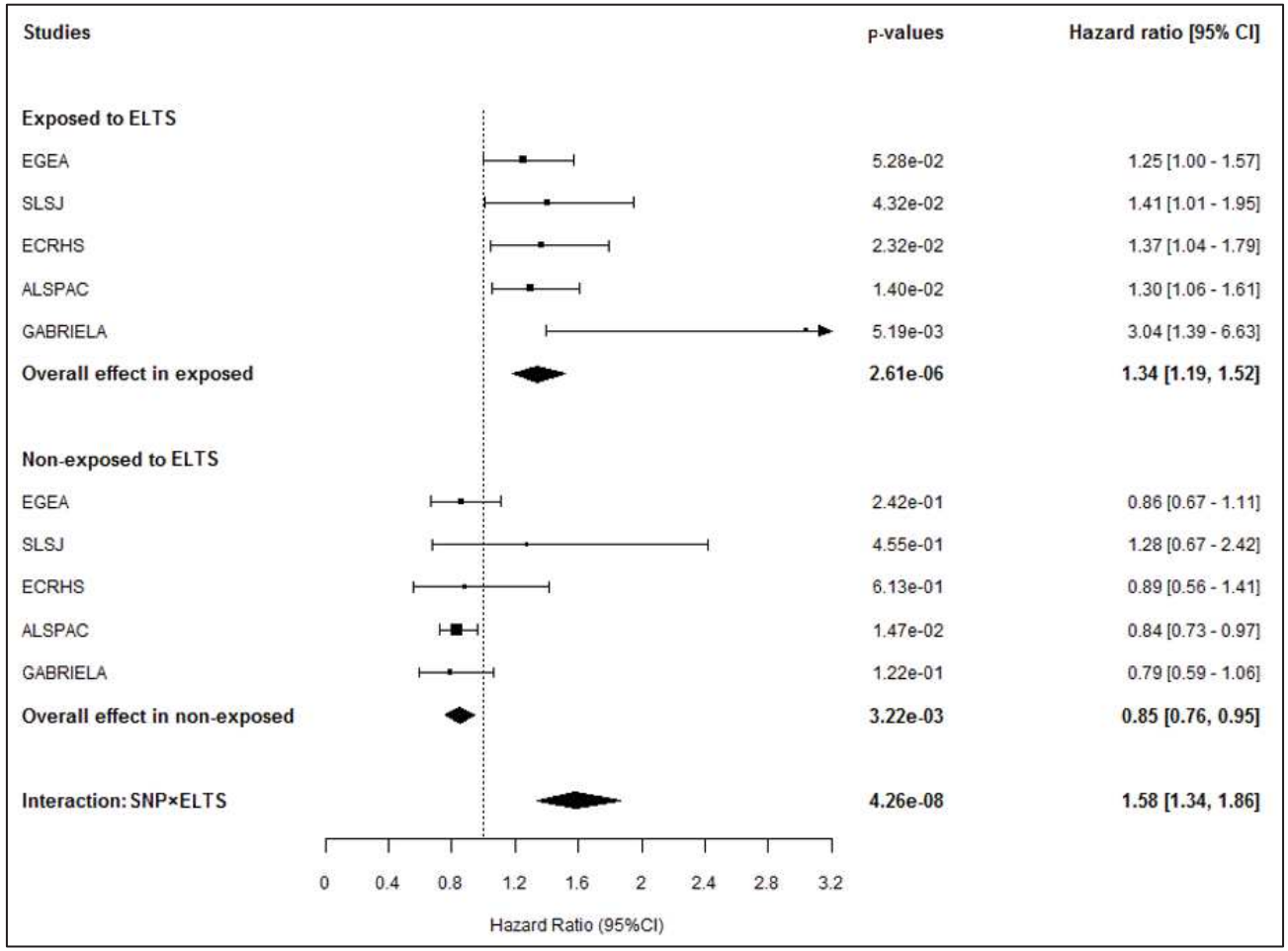


FIGURE 4. Forest Plot of the 13q21 lead SNP (rs7334050) according to early-life tobacco smoke (ELTS) exposure



Supplementary Information

Genome-wide interaction study of early-life smoking exposure on time-to-asthma onset in childhood. Sugier et al.

MATERIAL AND METHODS

Study populations

Five independent studies of European-ancestry were part of the present project. There were two studies with a family-based structure: the Epidemiological study on the Genetics and Environment of Asthma (EGEA) and the Saguenay-Lac-Saint-Jean Familial Collection (SLSJ); and three populations-based studies: the European Community Respiratory Health Survey (ECRHS), the GABRIEL Advanced Surveys (GABRIELA) and the Avon Longitudinal Study of Parents and Children (ALSPAC). All studies had available information on age of asthma onset, age at last examination and early-life tobacco smoke (ELTS) exposure, and imputed genetic data. In the present work, we focused on childhood onset asthma defined as ever asthma with disease onset before 16 years of age. Subjects with an older age of asthma onset or missing age of onset were excluded from the analyses. We included all non-asthmatic subjects whatever their age at their last examination. However, if their age at last examination was greater than 16 years, the age was censored at 16 years of age in the survival analysis model because it was the time of the last event for asthmatics (last age of asthma onset). All adult participants and child's legal guardians provided written informed consent.

After a study-specific quality control process, a total of 8,273 subjects among which 3,187 exposed ($ELTS^+$) and 5,086 non-exposed ($ELTS^-$) subjects were included in the present study. All genome-wide association studies were done centrally at INSERM UMR-946 (Paris, France) except for ALPSAC for which we had only access to summary statistics.

Sample ascertainment and phenotype definition

EGEA

Briefly, the EGEA study is a 20-year longitudinal survey (EGEA1: 1991-1995, EGEA2: 2003-2007 and EGEA3: 2011-2013) which combines a case-control and a family-based study of asthma cases. The whole study population includes 388 families ascertained through at least one asthmatic proband recruited in chest clinics (1,705 probands and first-degree relatives) plus 415 population-based controls (total of 2,120 subjects). All subjects were born in France and were of European ancestry. The protocol of this study has been described elsewhere.[1-3]

Asthma was defined in probands by a positive answer to the following four standardized questions: 1) “Have you ever had attacks of breathlessness at rest with wheezing?”, 2) “Have you ever had asthma attacks?”, 2a) “Was this diagnosis confirmed by a physician?”, and 2b) “Have you had an asthma attack in the last 12 months?” or on a positive self-report to two of the before mentioned items plus a medical record of asthma. The relatives of probands were defined as asthmatics if they answered positively to the items: “Have you ever had attacks of breathlessness at rest with wheezing?” or “Have you ever had asthma attacks?” at EGEA1 or EGEA2 or EGEA3. The subjects that never had asthma up to their last follow-up were considered as non-asthmatics. For individuals who developed asthma, information on asthma age at onset was obtained from adult asthmatics or parents of asthmatic children who answered to the following question: “How old were you when you had your first asthma attack?” or “How old was your child when he (or she) had his (her) first asthma attack?” For non-asthmatic subjects, we considered age at their last examination. Early-life tobacco smoke (ELTS) exposure was defined as follows: 1) for a child, by a positive answer to the question asked to the child’s mother (or father): "Did you or the father of your child smoke when your child was less than 2 years old?" and/or “Did you smoke when you were pregnant?”; 2) for an adult, by a

positive answer to the questions: "Did your mother or your father smoke during your early-childhood?", and/or “Did your mother smoke when she was pregnant with you?”.

Ethical approval was obtained from the relevant institutional review board committees (Cochin Port-Royal Hospital and Necker-Enfants Malades Hospital, Paris: n° 01-07-07, 04-05-03, 04-11-13 and 04-11-18). Written informed consent was signed by all participants. Written informed consent was signed by kin or guardians of the minors/children.

SLSJ

The Saguenay-Lac-Saint-Jean and Quebec City Familial Asthma Collection (SLSJ) consists of a French-Canadian founder population panel of 253 multigenerational families from Saguenay-Lac-Saint-Jean region, ascertained through two asthmatic probands between 1997 and 2002. The SLSJ study protocol has been described elsewhere.[4]

Probands were included in the study if they fulfilled at least two of the following criteria: 1) a minimum of three clinic visits for acute asthma within one year; 2) two or more asthma-related hospital admissions within one year; or 3) steroid dependency, as defined by either six months of oral, or one year of inhaled corticosteroid use. Families were included in the study if at least one parent was available for phenotypic assessment, at least one parent was unaffected, and all four grandparents were of French-Canadian origin. Members of the family were considered asthmatics if a self-reported history of asthma has already been reported and confirmed by a physician, or by clinical evaluation after a positive methacholine test (less or equal to 8mg/ml of methacholine). Age of asthma onset was obtained from answers to the question: “How old were you when you had your first asthma attack?”. When age of onset was defined below 2 years (in 41 cases), a default class of 2 years was adopted to avoid uncertainty. For non-asthmatic subjects, we considered the age at the examination. ELTS exposure was defined by a

passive smoking exposure before two years of age for children or before five years of age for adults.

The SLSJ study was approved by the ethic committees of the academic Integrated Health and Social Services Centres of Saguenay-Lac-Saint-Jean (CIUSSSS) and of UQAC.

ECRHS

Sixteen centres (eight countries) in the European Community Respiratory Health Survey (ECRHS) contributed samples to the GWAS (http://www.ecrhs.org).[5,6] In each centre, a representative community-based sample of at least 3,000 adults aged 20-44 years were invited to complete a brief postal questionnaire asking about respiratory symptoms (ECRHS I - Stage 1) between 1991-1993. A random sample of these (600 per centre) underwent intensive further investigation (ECRHS I - Stage 2 – random sample). Participants who had symptoms highly suggestive of asthma but who had not been selected at random to take part in Stage 2, were also invited to undergo intensive investigations (ECRHS I - Stage 2- enriched sample). About ten years later all adults who had taken part in Stage 2 were re-contacted (ECRHS II) and again asked about respiratory symptoms. Samples suitable for DNA extraction were collected. For the GWAS initiative, all cases of asthma were identified (participants from the random or enriched sample who said yes to the question: “Have you ever had asthma?” at either ECRHS I or ECRHS II) and information on asthma age at onset was obtained from age at first asthma attack at ECRHS I or ECRHS II. Controls were a random sample (of the random sample) who answered “no” to the same question in both surveys. ELTS exposure was defined according to the following questions: “Did your father ever smoke regularly during your childhood?”, “Did your mother ever smoke regularly during your childhood, or before you were born?” (ECRHS I main questionnaire) and/or based on *in utero* smoke exposure defined as any smoking by the mother during pregnancy.

Ethical approval was obtained from the Hospital del Mar/IMIM ethics committee for ECRHS-Spain ((ECRHS II Spain) 98/835/I, (ECRHS III Spain) 2009/3500/1).

GABRIELA Advanced Surveys

The GABRIELA surveys are cross- sectional population- based surveys conducted in rural areas of Austria, Germany, and Switzerland. In total, 135,359 children aged 6- 12 years were addressed through schools. In a first stage in fall/winter 2006, asthma, allergic disease, and contact to farming environments were assessed using a short parental questionnaire (n=79,888). In a second stage in spring/summer 2007, 9,668 children were selected among families consenting in writing to blood sampling, genetic testing and collection of environmental samples by stratified random sampling to ensure representation of children with high exposure to farming environments.[7,8] Genomic DNA and questionnaire data were available for 7,303 children of whom 862 asthmatic cases and 865 controls were selected for genotyping.

An asthmatic case was defined by a parental report of 1) asthma diagnosed by a doctor at least once, or 2) asthmatic bronchitis diagnosed at least twice during lifetime. The subjects with no reported diagnosis of asthma ever and a diagnosis of asthmatic bronchitis no more than once were considered as non-asthmatics. Among asthmatics, the age of disease onset was obtained from the following original question: "How old was your child when the first symptoms of wheezing or whistling in the chest began? At the age of … years – or - if during the first year: At the age of … months". In non-asthmatics, we considered the age at the examination. ELTS exposure was based on the exposure to maternal tobacco smoking at any time during pregnancy. We did not use the information on passive tobacco smoke exposure of the children at the time of the survey (i.e current smoking of the father and/or the mother) because it occurred at 9 years of age, thus after the mean age of asthma onset (2.9 years;

supplementary Table 1A). However, eighty percent of the children that were exposed to ELTS *in utero*, were still exposed to second-hand tobacco at the time of the survey.

The GABRIELA study was approved by the institutional review boards of the Bavarian Medical Association (for Bavaria), Ulm University (for Baden-Württemberg), the cantons Lucerne, Zurich and Thurgau (for Switzerland), Medical University of Innsbruck (for Austria), and Medical University of Wroclaw (for Poland) and informed consent was obtained from the parents.

ALSPAC

The Avon Longitudinal Study of Parents and Children (www.bristol.ac.uk/alspac) is a population-based, longitudinal, birth cohort study that was recruited during pregnancy. Pregnant women resident in Avon, United Kingdom with estimated dates of delivery between 1st April 1991 and 31st December 1992 were recruited through antenatal clinics.[9] Of 14,451 women recruited, there were 14,072 live births and 13,988 children were alive at age one year. Children were followed from birth using self-completion questionnaires sent to their mothers at approximately annual intervals and hands- on assessments at annual dedicated research clinics from age 7 years.

Asthma was defined as a positive response to at least one of the following questions: 1-"Has a doctor ever told you that your child has asthma?" in a questionnaire sent to their mothers at 91 months after birth (approximately 7½ years), and 2-"Did your child had an asthma attack in the last 12 months?" at month 103, 128, 157 and 166. Others with negative complete reports (i.e. negative report at all follow-up) were considered as non-asthmatics. However, non-asthmatic children who experienced wheeze or wheezing and whistling before 6yrs of age, were excluded from the present analysis. Among asthmatics, age of onset asthma was based on the first declaration of whistles. Whistles was defined by a positive response to the question includes in

the annual questionnaire: “Has your child had wheezing, breathlessness or episodes of stopping breathing in past 12 months or since the last questionnaire?”. For non-asthmatics, we considered the follow-up period without any missing visit and took their age at the last examination. ELTS exposure was defined by: 1) maternal active smoking during pregnancy or 2) post-natal exposure to parental smoking (8 months after pregnancy). We used this ELTS exposure definition, so that exposure was most likely to occur before asthma onset which was 3.1 years on average.

Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. A list of the ethics committee/institutional review board(s) that approved aspects of the study are available at http://www.bristol.ac.uk/alspac/researchers/research-ethics/.

ACKNOWLEDGMENTS

EGEA: The authors thank all those who participated to the setting of the study and on the various aspects of the examinations involved: interviewers, technicians for lung function testing and skin prick tests, blood sampling, IgE determinations, coders, those involved in quality control, data and sample management and all those who supervised the study in all EGEA centers. The authors are grateful to the three CIC-Inserm of Necker, Grenoble and Marseille who supported the EGEA study and in which subjects were examined. They are also grateful to the biobanks in Lille (CIC-Inserm) and at Annemasse (Etablissement français du sang) where EGEA biological samples were/are stored. We thank the Epidemiological Study on Genetics and Environment of Asthma (EGEA) cooperative group members as follows.

Coordination: V Siroux (epidemiology, PI since 2013); F Demenais (genetics); I Pin (clinical aspects); R Nadif (biology). F Kauffmann (PI 1992-2012); **Respiratory epidemiology:** Inserm U 700, Paris M Korobaeff (Egea1), F Neukirch (Egea1); Inserm U 707, Paris: I Annesi-

Maesano (Egea1-2); Inserm CESP/U 1018, Villejuif: F Kauffmann, N Le Moual, R Nadif, MP Oryszczyn (Egea1-2), R. Varraso; Inserm U 823, Grenoble: V Siroux. **Genetics:** Inserm U 393, Paris: J Feingold; Inserm U 946, Paris: E Bouzigon, F Demenais, MH Dizier; CNG, Evry: I Gut (now CNAG, Barcelone, Spain), M Lathrop (now Univ McGill, Montreal, Canada). **Clinical centers:** Grenoble: I Pin, C Pison; Lyon: D Ecochard (Egea1), F Gormand, Y Pacheco; Marseille: D Charpin (Egea1), D Vervloet (Egea1-2); Montpellier: J Bousquet; Paris Cochin: A Lockhart (Egea1), R Matran (now in Lille); Paris Necker: E Paty (Egea1-2), P Scheinmann (Egea1-2); Paris-Trousseau: A Grimfeld (Egea1-2), J Just. **Data and quality management:** Inserm ex-U155 (Egea1): J Hochez; Inserm CESP/U 1018, Villejuif: N Le Moual, Inserm ex-U780: C Ravault (Egea1-2); Inserm ex-U794: N Chateigner; Grenoble: J Ferran (Egea1-2)

ALSPAC: We are extremely grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. We would like to acknowledge Asthma UK Grant ref: 06/005 for supporting the collection of the outcome data used in this research project. GWAS data was generated by Sample Logistics and Genotyping Facilities at Wellcome Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe.

REFERENCES

1. Bouzigon E, Nadif R, Le Moual N, et al. [Genetic and environmental factors of asthma and allergy: Results of the EGEA study]. *Rev Mal Respir.* 2015;32(8):822-840.
2. Kauffmann F, Dizier MH, Annesi-Maesano I, et al. EGEA (Epidemiological study on the Genetics and Environment of Asthma, bronchial hyperresponsiveness and atopy)--descriptive characteristics. *Clin Exp Allergy.* 1999;29 Suppl 4:17-21.
3. Kauffmann F, Dizier MH, Pin I, et al. Epidemiological study of the genetics and environment of asthma, bronchial hyperresponsiveness, and atopy: phenotype issues. *Am J Respir Crit Care Med.* 1997;156(4 Pt 2):S123-129.
4. Laprise C. The Saguenay-Lac-Saint-Jean asthma familial collection: the genetics of asthma in a young founder population. *Genes Immun.* 2014;15(4):247-255.
5. The European Community Respiratory Health Survey II. *Eur Respir J.* 2002;20(5):1071-1079.
6. Burney PG, Luczynska C, Chinn S, Jarvis D. The European Community Respiratory Health Survey. *Eur Respir J.* 1994;7(5):954-960.
7. Ege MJ, Strachan DP, Cookson WO, et al. Gene-environment interaction for childhood asthma and exposure to farming in Central Europe. *J Allergy Clin Immunol.* 2011;127(1):138-144, 144 e131-134.
8. Genuneit J, Buchele G, Waser M, et al. The GABRIEL Advanced Surveys: study design, participation and evaluation of bias. *Paediatr Perinat Epidemiol.* 2011;25(5):436-447.
9. Boyd A, Golding J, Macleod J, et al. Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol.* 2013;42(1):111-127.

TABLE S1. Description of the five studies included in the genome-wide interaction analysis of ELTS exposure on time-to-asthma onset in childhood

Part A. Descriptive statistics of the five datasets

	EGEA (N=1 498)		**SLSJ (N=377)**		**ECRHS (N=1 685)**		**GABRIELA (N=1 482)**		**ALSPAC (N=3 231)**	
Descriptive statistics										
ELTS^{+}	Asthmatics	Non-asthmatics	Asthmatics	Non-asthmatics	Asthmatics	Non-asthmatics	Asthmatics	Non-asthmatics	Asthmatics	Non-asthmatics
N	297	643	155	73	175	1,008	82	81	330	343
Sex, men (%)	173 (58.2)	313 (48.7)	70 (45.2)	33 (45.2)	95 (54.3)	491 (48.7)	47 (57.3)	41 (50.6)	197 (59.7)	181 (52.8)
Asthma age-of-onset in years, mean (SD)	6.1 (4.4)	-	5.7 (4.7)	-	7.0 (4.8)	-	2.9 (2.6)	-	3.1 (3.5)	-
ELTS^{-}	Asthmatics	Non-asthmatics	Asthmatics	Non-asthmatics	Asthmatics	Non-asthmatics	Asthmatics	Non-asthmatics	Asthmatics	Non-asthmatics
N	210	348	90	59	84	418	578	741	873	1,685
Sex, men (%)	129 (61.4)	170 (48.9)	53 (58.9)	28 (47.5)	45 (53.6)	209 (50.0)	366 (63.3)	386 (52.1)	484 (55.4)	786 (46.6)
Asthma age-of-onset in years, mean (SD)	5.6 (4.2)	-	5.0 (4.1)	-	6.9 (4.9)	-	2.9 (2.3)	-	3.5 (3.7)	-

EGEA, the Epidemiological study of the Genetics and Environment of Asthma; SLSJ, the Saguenay-Lac-Saint Jean study; ECRHS, the European Community Respiratory Health Survey; GABRIELA: the GABRIEL Advanced Surveys; ALSPAC, the Avon Longitudinal Study of Parents and Children; N, number; SD, standard deviation.

Part B. Information on genotyping, quality control (QC) and imputation

	EGEA (N=1,498)	SLSJ (N=377)	ECRHS (N=1,685)	GABRIELA (N=1,482)	ALSPAC (N=3,231)
Genotyping					
Genotyping platform	Illumina Human610-Quad	Illumina Human610-Quad	Illumina Human610-Quad	Illumina Human610-Quad	Illumina HumanHap550Quad
Genotyping center	Centre National de Génotypage, Evry, France	Centre National de Génotypage, Evry, France	Centre National de Génotypage, Evry, France	Centre National de Génotypage, Evry, France	23andMe subcontracting the Wellcome Trust Sanger Institute, Cambridge, UK, and the LabCorp, Burlington, North Carolina, US
Individual QC					
Call-rate	97%	97%	97%	97%	97%
Heterozygosity	Individuals excluded if <0.30 or >0.33	Individuals excluded if <0.30 or >0.33	Individuals excluded if <0.30 or >0.33	Individuals excluded if <0.30 or >0.33	Individuals excluded if <0.320 or >0.345 for the Sanger data and <0.310 or >0.330 for the LabCorp data
Ethnic outliers	PCA based	PCA based	PCA based	PCA based	PCA based
SNP QC filters before imputation					
Minor Allele Frequency	5%	5%	5%	5%	1%
Hardy Weinberg Equilibrium test p-value	10^{-4}	10^{-4}	10^{-4}	10^{-4}	$5x10^{-7}$
Call-rate	97%	97%	97%	97%	95%
Imputation - Genome					
Software	MACH 1.0	MACH 1.0	MACH 1.0	MACH 1.0	MACH 1.0
HapMap release	Hapmap2 r21	Hapmap2 r21	Hapmap2 r21	Hapmap2 r21	Hapmap2 r22
SNP QC filters	rsq ≥ 0.5 & MAF ≥ 1%	rsq ≥ 0.5 & MAF ≥ 1%	rsq ≥ 0.5 & MAF ≥ 1%	rsq ≥ 0.5 & MAF ≥ 1%	rsq ≥ 0.5 & MAF ≥ 1%

TABLE S2. SNPs showing interaction ($P \leq 5x10^{-6}$) with ELTS exposure on time-to-asthma onset in childhood

CHR	SNP	Position (kb)	Closest Gene (distance of SNP to gene in kb)	E/R	EAF	SNP x ELTS interaction		SNP effect in subjects exposed to ELTS				SNP effect in subjects unexposed to ELTS			
						HR [95% CI]	2-sided *P*	HR_{ELTS+} [95% CI]	2-sided *P*	$P_{Cochran}$	I^2	HR_{ELTS-} [95% CI]	2-sided *P*	$P_{Cochran}$	I^2
2p22	rs232542	38 328	*CYP1B1* (25)	C/T	0.69	1.33[1.18-1.50]	$4.1x10^{-6}$	1.18[1.07-1.30]	$7.0x10^{-4}$	0.88	0	0.89[0.82-0.96]	$1.7x10^{-3}$	0.62	0
2p22	rs232540	38 329	*CYP1B1* (26)	C/T	0.69	1.33[1.18-1.50]	$4.8x10^{-6}$	1.18[1.07-1.30]	$7.5x10^{-4}$	0.87	0	0.89[0.82-0.96]	$2.0x10^{-3}$	0.60	0
2p22	rs151313	38 332	*CYP1B1* (29)	A/T	0.31	0.75[0.67-0.85]	$5.0x10^{-6}$	0.85[0.77-0.93]	$7.8x10^{-4}$	0.87	0	1.13[1.04-1.21]	$2.0x10^{-3}$	0.60	0
2p22	rs232535	38 332	*CYP1B1* (29)	C/T	0.69	1.33[1.18-1.50]	$5.0x10^{-6}$	1.18[1.07-1.30]	$7.9x10^{-4}$	0.87	0	0.89[0.82-0.96]	$2.0x10^{-3}$	0.61	0
13q21	rs17742723	70 635	*KLHL1* (0)	A/T	0.86	0.65[0.55-0.77]	$6.7x10^{-7}$	0.74[0.65-0.83]	$1.6x10^{-6}$	0.45	0	1.12[1.01-1.26]	$3.9x10^{-2}$	0.92	0
13q21	rs1372284	70 639	*KLHL1* (0)	A/G	0.86	0.68[0.58-0.80]	$3.9x10^{-6}$	0.76[0.67-0.86]	$1.3x10^{-5}$	0.84	0	1.11[1.00-1.24]	$4.7x10^{-2}$	0.44	0
13q21	rs2439614	70 641	*KLHL1* (0)	C/T	0.13	1.53[1.30-1.80]	$4.0x10^{-7}$	1.34[1.19-1.52]	$2.9x10^{-6}$	0.51	0	0.88[0.79-0.98]	$1.9x10^{-2}$	0.81	0
13q21	rs7334050	70 645	*KLHL1* (0)	G/T	0.14	1.58[1.34-1.86]	$4.3x10^{-8}$	1.34[1.19-1.52]	$2.6x10^{-6}$	0.31	0	0.85[0.76-0.95]	$3.2x10^{-3}$	0.76	0
13q21	rs9542173	70 646	*KLHL1* (0)	C/T	0.87	0.64[0.54-0.76]	$1.3x10^{-7}$	0.74[0.66-0.84]	$1.9x10^{-6}$	0.42	0	1.16[1.04-1.29]	$1.0x10^{-2}$	0.81	0
13q21	rs299526	70 655	*KLHL1* (0)	C/T	0.13	1.49[1.27-1.76]	$1.6x10^{-6}$	1.33[1.17-1.50]	$8.7x10^{-6}$	0.82	0	0.89[0.80-0.99]	$2.9x10^{-2}$	0.30	0
14q22	rs4898674	51 316	*NIN* (19)	A/C	0.27	1.37[1.20-1.56]	$3.3x10^{-6}$	0.16[1.05-1.29]	$2.9x10^{-3}$	0.77	0	0.85[0.78-0.93]	$2.6x10^{-4}$	0.89	0
14q22	rs7493885	51 317	*NIN* (20)	G/T	0.27	1.37[1.20-1.56]	$2.9x10^{-6}$	1.16[1.05-1.29]	$2.7x10^{-3}$	0.77	0	0.85[0.78-0.93]	$2.5x10^{-4}$	0.90	0
14q22	rs1951474	51 317	*NIN* (20)	A/T	0.73	0.73[0.64-0.83]	$3.1x10^{-6}$	0.86[0.78-0.95]	$2.6x10^{-3}$	0.77	0	1.17[1.08-1.28]	$2.8x10^{-4}$	0.90	0
14q22	rs8020067	51 318	*NIN* (20)	C/G	0.73	0.73[0.64-0.83]	$3.1x10^{-6}$	0.86[0.78-0.95]	$2.6x10^{-3}$	0.77	0	1.17[1.08-1.28]	$2.8x10^{-4}$	0.90	0
20p12	rs13037508	14 928	*MACROD2* (0)	A/T	0.63	1.48[1.27-1.72]	$4.9x10^{-7}$	1.20[1.06-1.35]	$3.6x10^{-3}$	0.15	0.41	0.81[0.74-0.89]	$7.9x10^{-6}$	0.28	0.21

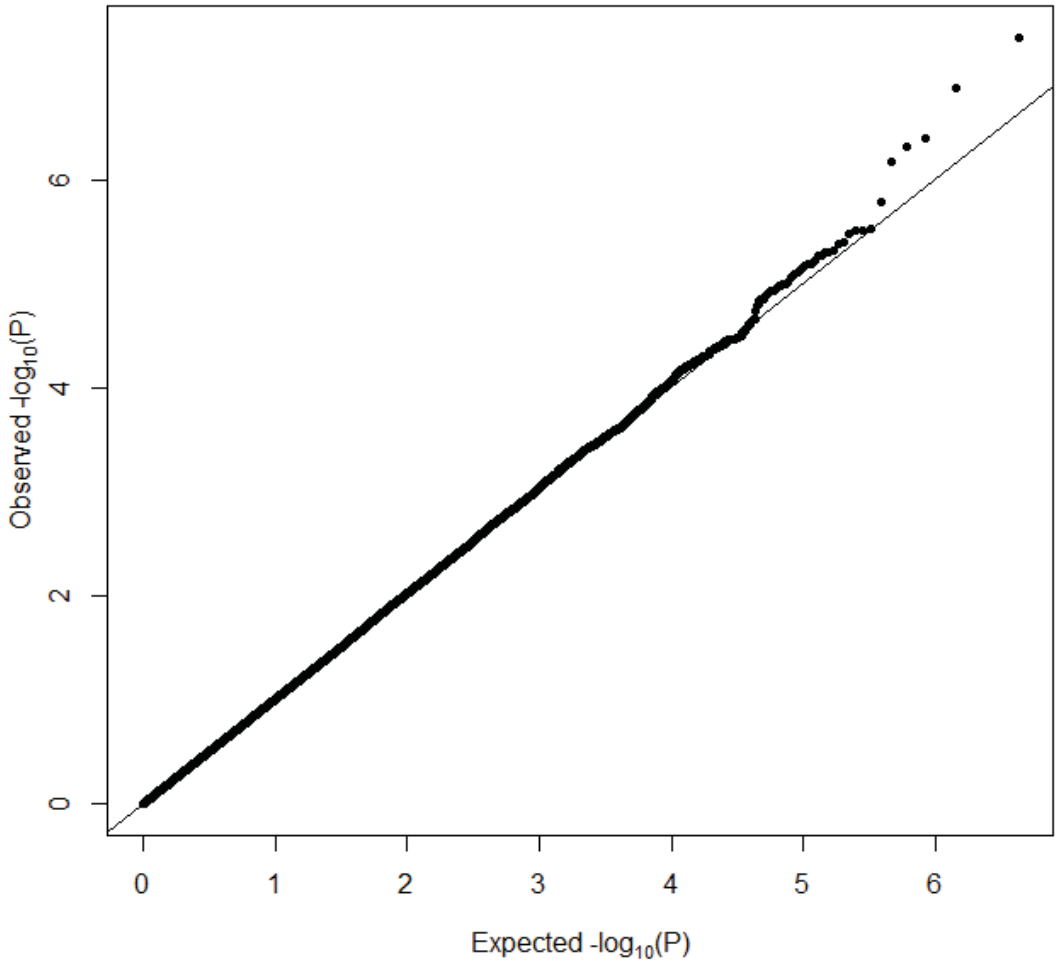
CHR, chromosome; SNP, single nucleotide polymorphism; E, effect allele / R, Reference allele; EAF, effect allele frequency; HR, hazard ratio; CI: confidence interval. $P_{Cochran}$, *P* for Cochran's Q test of heterogeneity; I^2, the I^2 statistic of the fraction of variance that is due to heterogeneity. [a]The interaction effect size between SNP and ELTS exposure was estimated as the difference between the $ELTS^+$ and $ELTS^-$ combined SNP effect sizes obtained from the fixed-effects meta-analyses of the five studies in each $ELTS^{+/}ELTS^-$ stratum.

TABLE S3. Gene Ontology categories showing significant enrichment in genes interacting with early-life tobacco smoke exposure on time-to-asthma onset in childhood.

Gene Ontology (GO)	Type	Gene set size #genes	GO name	Genes driving the enrichment score of each significant GO
GO:0042742	P	67	Defense response to bacterium	*DEFB114, DEFB133, BPI, TLR9, STAB1, LYZ, SPACA3, LBP, STAB2, PRG2, NOD1, DEFB1, RNASE7, BCL3, DEFB110, DEFB111, DEFB127, TLR3, LTF, LALBA, FCER1G, DEFB128, DEFB116, NOD2, DEFA6, DEFB125, DEFB123, DEFB115, DEFA4, IFNB1, PGLYRP1, IL10, DEFB119, WFDC12, DEFB118*
GO:0006119	P	52	Oxidative phosphorylation	*CLCA1, ATP5J, NDUFS4, NDUFA8, ATP5C1, NDUFB1, ATP5O, COX10, NDUFS5, NDUFB2, NDUFV2, NDUFB7, NDUFS2, MON2, NDUFV3*
GO:0016125	P	76	Sterol metabolic process	*SOAT1, FDFT1, SC4MOL, LEPR, OSBPL1A, PON1, CUBN, DHCR7, PCTP, MBTPS1, CYB5R3, ABCA1, ABCG1, CYP7A1, PRKAG2, HDLBP, NPC1L1, LEPROT, HMGCR, CLN6, SORL1, PCSK9, LEP, CYP7B1, IDI2, INSIG2, OSBPL5, APOA4, TRERF1, IDI1, CYB5R1, RXRA, SREBF1, CYP46A1, AKR1D1, TM7SF2, HMGCS2, APOA2, CYB5R2, SC5DL, APOL2, APOL1, SREBF2, PRKAA2, DHCR24, MVK, SCARB1, PPARD*

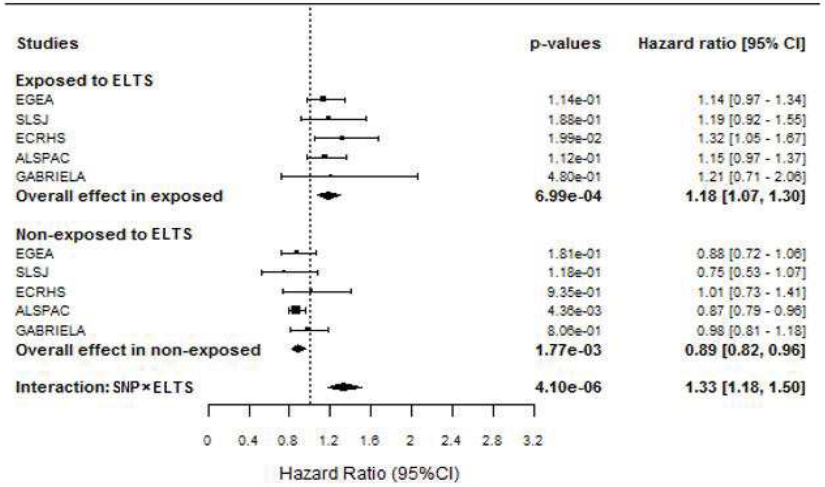
P=biological process; Gene ontology categories showing significant enrichment have FDR < 0.05 and FWER < 0.05 (based on 10,000 permutations of SNP x ELTS interaction test statistics).

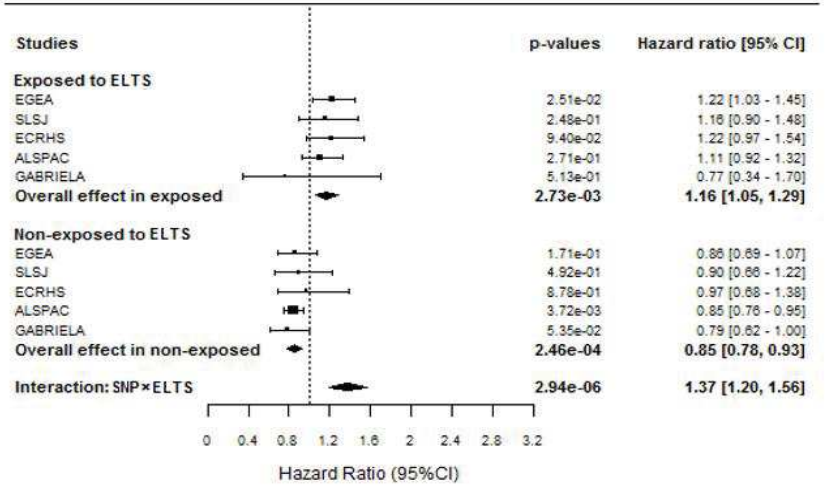
FIGURE S1. Quantile-quantile plot of the genome-wide interaction analysis for time-to-asthma onset in childhood

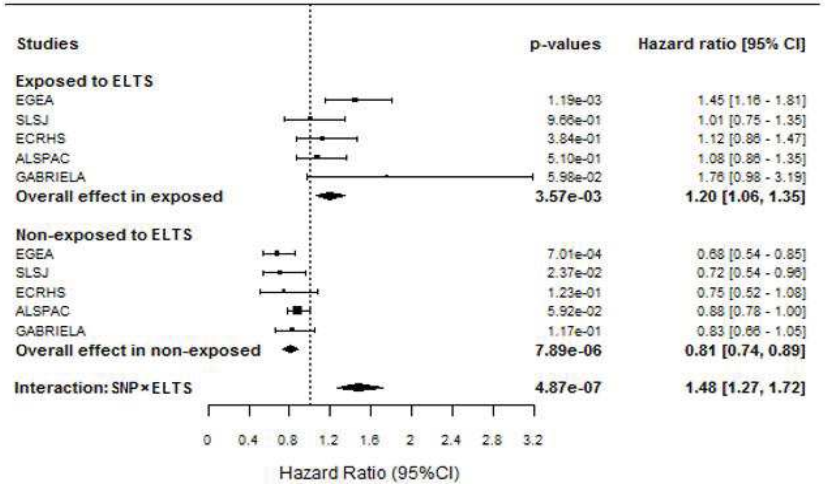


Quantile-quantile (QQ) plot for the tests of interaction between individual SNPs and ELTS exposure on time-to-asthma onset in childhood. The dots represent the distribution of observed $-\log_{10}(P)$ values against the expected $-\log_{10}(P)$ values from a theoretical chi-square distribution with one df. The straight line represents the theoretical distribution of expected $-\log_{10}(P)$ values under the null hypothesis of no association. There was no evidence of any systematic bias: the genomic inflation factor (λ) was equal to 1.003.

FIGURE S2. Forest plots of the lead SNP at the three loci showing suggestive interaction with early-life tobacco smoke (ELTS) exposure on time-to-asthma in childhood ($P \leq 5x10^{-6}$): 2p22 (rs232542) at top-left, 14q22 (rs7493885) at top-right, and 20p12 (rs13037508) at bottom. Hazard ratios (HR) and 95% Confidence Intervals (CI) are plotted by study within each ELTS exposure stratum. The combined HR estimates over all five studies in each stratum and the combined SNP by ELTS exposure interaction HR are plotted as a diamond.







CHAPITRE V – DISCUSSION ET PERSPECTIVES

Les études d'association génétique de l'asthme et de l'atopie conduites jusqu'à ce jour ont identifié des variants fréquents avec des effets relativement modestes qui n'expliquent qu'une part de la composante génétique de ces maladies. Au cours de cette thèse, nous avons identifié de nouveaux variants de susceptibilité génétique à l'asthme et à l'allergie par des analyses d'interactions gène-gène et gène environnement pour mettre en évidence de nouveaux facteurs génétiques associés à l'asthme et à l'atopie. La mise en place de ces analyses a permis de soulever différentes questions qui auront parfois imposé des contraintes à nos analyses, ou simplement à faire des choix méthodologiques. Dans cette partie, je vais discuter certains de ces points : 1) l'homogénéité des données phénotypiques, 2) le choix des méthodologies statistiques, 3) l'apport de connaissances extérieures aux données, 4) l'annotation des SNPs aux gènes, et 5) la validation des résultats.

Homogénéité des données phénotypiques

Lors de cette thèse, j'ai eu l'opportunité d'effectuer mon travail dans le cadre de collaborations internationales, sur des données incluses dans le consortium européen sur l'asthme GABRIEL. Les études qui ont contribué à mes travaux de thèse avaient différents modes de recensement : trois études familiales recensées par des asthmatiques (EGEA, SLSJ, MRC-UK), deux études en population générale (ECRHS, GABRIELA), et une cohorte de naissance (ALSPAC). Ces études disposaient de données phénotypiques et environnementales détaillées. Dans le cadre de mon premier projet de thèse, le phénotype considéré était l'atopie définie à partir de tests cutanés à des aéroallergènes. Nous avons considéré comme atopiques les sujets qui présentaient au moins une réponse positive aux tests cutanés effectués afin d'uniformiser les phénotypes d'intérêts entre les études. Les protocoles pour définir la positivité de ces tests cutanés étaient similaires dans les trois études, mais les allergènes testés n'étaient pas identiques. Certains allergènes parmi les plus courants ont été testés dans les trois échantillons (acariens, chat, fléole des près, *Cladosporium herbarum* et *Aspergillus*), mais le nombre d'allergènes testés était différent : 11 dans EGEA, 24 dans SLSJ, 5 dans MRC. La précision du phénotype n'était donc pas la même à travers les études, et il est notamment possible que certains sujets de MRC considérés comme non atopiques étaient en fait atopiques, entrainant une diminution de la

puissance de détection des SNPs et des interactions gène-gène. Nous avons réalisé des analyses d'association pangénomique simple marqueur afin de voir si le risque de certains variants génétiques était spécifique à des sous-groupes d'allergènes (allergènes d'intérieurs, allergènes d'extérieurs, et moisissures). Les premiers résultats ont montré des tendances (notamment un signal au locus 20p12 spécifique aux moisissures), mais aucun SNP n'atteignait le seuil de correction génome entier.

Dans le cadre de mon second projet visant à identifier des interactions gène-exposition au tabac pendant la petite enfance sur le délai de survenue de l'asthme, j'ai utilisé trois variables (statut asthmatique, l'âge – âge de début de l'asthme ou au dernier examen -, l'exposition au tabagisme passif durant la petite enfance) afin de définir mes phénotypes d'intérêts. Bien qu'il existe des différences entre études dans la définition de l'asthme, ce phénotype était toutefois bien caractérisé. Il est cependant possible que l'âge de début de l'asthme basé sur la déclaration des asthmatiques adultes au moment de l'enquête (avec un âge de début de l'asthme dans l'enfance) des sujets des études EGEA, SLSJ, et ECRHS, soit moins précis que pour les asthmatiques enfants de ces études et des deux études ne comportant que des enfants. Pour ces analyses, il était aussi important de caractériser uniquement l'exposition passive au tabac dans l'enfance survenant avant l'âge de début de l'asthme. Pour cela, nous n'avons pas utilisé dans l'étude GABRIELA, l'information de l'exposition des enfants au moment de l'enquête (c'est-à-dire le tabagisme actuel d'au moins un des deux parents). Cette exposition ayant lieu à l'âge de neuf ans et l'âge moyen de début de l'asthme dans cette étude était de 2,9 ans. Il existait cependant une variabilité dans la définition de l'exposition entre études. Seule l'exposition *in utero* pour la mère a été prise en compte dans GABRIELA. Dans ALSPAC, l'exposition *in utero* ou avant l'âge de huit mois a été considéré, l'âge moyen de début d'asthme étant de 3,1 dans cette étude. Cette définition pouvait aller jusqu'à l'âge de cinq ans pour les adultes de SLSJ, pendant la petite enfance pour les adultes de EGEA, ou pendant l'enfance dans l'étude ECRHS. Dans le cadre de ce projet, il serait également intéressant d'étudier l'âge de début de l'asthme chez les asthmatiques, en utilisant les mêmes méthodes d'analyse de survie, afin d'identifier les interactions entre facteurs génétiques et ELTS influençant directement l'âge d'apparition de la maladie.

Choix des méthodologies utilisées

Dans le cadre de mon premier projet de thèse, nous avons utilisé des filtres statistiques consistant à réduire le nombre de tests d'interactions gène-gène aux SNPs appartenant à des gènes identifiés lors d'une analyse pangénomique simple-marqueur. Parmi un groupe de gènes

sélectionné à un seuil suggestif, nous avons choisi de guider la recherche d'interaction par les gènes détectés par au moins un SNP significatif au seuil génome entier. Cette stratégie a permis de réduire drastiquement le nombre d'interactions possibles, mais il aurait également pu être intéressant de tester toutes les paires de SNPs à l'intérieur de l'ensemble des paires de gènes sélectionnées.

Dans le cadre du second projet d'interaction gène-environnement nous avons choisi d'utiliser le test d'interaction GxE à un degré de liberté. La puissance des méthodes d'interactions gène-environnement dépend très fortement du modèle d'interaction sous-jacent, qui n'est pas connu *a priori*. Le test d'interaction GxE à un degré de liberté a été montré plus puissant que le test d'interaction joint à deux degrés de liberté ou que la plupart des modèles en deux étapes, en cas d'absence d'effet marginal. En faisant ce choix, nous avons favorisé la détection d'interactions de type flip-flop, et dans une moindre mesure d'interactions avec un effet seulement dans l'une des deux strates exposés ou non exposés. Cependant, ce test d'interaction (1ddl) a une puissance réduite pour détecter des interactions à effet synergiques, ou avec un effet marginal du SNP en général. Il est donc possible que d'autres interactions comme des interactions de types synergiques, seraient détectées avec d'autres types de tests d'interactions.

Pour les analyses d'interactions GxE, nous avons aussi intégré la variabilité de l'âge de début dans la définition de la maladie en utilisant des méthodes d'analyse de survie (modèles de Cox). Il a été montré que l'âge de début d'asthme joue un rôle important, l'utilisation de modèles de survies peut apporter un gain de puissance de détection de variants génétiques associés à l'asthme, même s'il est possible que cet apport soit plus important pour l'asthme en général que pour l'asthme apparaissant dans l'enfance. Il serait intéressant d'investiguer différents modèles et notamment de comparer ce modèle à une régression logistique classique dans différents cas. La puissance des méthodes de survie est certainement plus importante dans le cas des interactions pour lesquelles le SNP a également un effet sur l'âge de début.

Apport de connaissances extérieures aux données

Dans le cadre du projet d'analyse d'interactions gène-gène, nous avons utilisés des filtres basé sur des connaissances extérieures en utilisant la méthode de fouille de texte GRAIL afin de sélectionner un sous ensemble de paires de gènes à tester en interaction gène-gène. La méthode GRAIL utilise l'information incluse dans les résumés de PubMed. C'est une méthode basée sur un modèle d'espace vectoriel où les gènes sont représentés par des vecteurs de mots. Comparé à des modèles basés sur la recherche de co-citations de mots clés, les approches basées sur des

vecteurs de mots ont l'avantage de prendre en compte le nombre et la pondération des termes d'index. Cependant, l'ensemble des approches de fouilles de textes restent dépendantes des découvertes déjà publiées dans la littérature. Les relations identifiables entre gènes sont donc biaisées en faveur de gènes déjà présents dans la littérature et peuvent n'offrir que des informations limitées sur de nouvelles découvertes. De plus, l'exploration de textes basée sur la littérature peut toujours présenter un biais par rapport aux concepts existants concernant la maladie. La méthode GRAIL s'est affranchie de ce problème en se basant sur une métrique agnostique à la maladie n'utilisant pas les voies biologiques connues, mais par ce fait augmente le risque de reconstruire des découvertes connues. L'intégration de connaissances venant de multiples sources (littérature, bases de données d'expression, réseaux d'interaction protéine-protéine, ontologie des gènes,…) pourrait permettre, au moins en parti, de pallier le problème du biais lié aux découvertes déjà publiées en enrichissant l'information utilisée. Cependant, la construction d'une métrique basée sur la combinaison de données hétérogènes est très complexe en pratique, et les méthodes basées sur des sources de formats différents sont souvent basées sur des ontologies personnalisées et des correspondances de mots clés. De plus, GRAIL ne permet pas en tant que tel d'identifier les types de relations biologiques qui existent entre les gènes. Par conséquent, les relations prédites entre gènes ne forment pas de voies biologiques car les informations sur les mécanismes putatifs ne sont pas utilisées. Ce problème pourrait également être diminué par l'utilisation de multiples sources de données, bien que la majorité des outils de hiérarchisation actuels [224,225] ne permettent pas de prendre en compte à la fois le type et le sens des relations.

Annotation des SNPs aux gènes

A différentes étapes de mon travail de thèse, il a été nécessaire d'annoter des résultats d'associations aux gènes. Lors du premier projet, l'annotation des résultats de GWAS était nécessaire à la sélection de groupes de gènes pour l'analyse de fouille de texte. Lors du second projet, l'annotation des résultats d'interaction SNPxELTS était nécessaire pour établir une significativité des gènes, puis des pathways, lors de l'analyse de pathway. Classiquement, les SNPs intra-génique sont généralement annotés aux gènes auxquels ils appartiennent. Cependant, la majorité des SNPs identifiés par les études pangénomiques est sont située dans des régions inter-géniques. Afin de considérer des SNPs proches de gènes ou en DL avec des SNPs inclus dans un gène mais qui n'auraient pas été génotypés (ou imputés) ou de prendre en compte des SNPs pouvant réguler la transcription d'un gène, une fenêtre basée sur une distance variable de 10 à 500 kb de part et d'autre des bornes du gène selon les études, ou sur les

structures de DL, est souvent considérée. Dans le cadre de ce travail, nous avons utilisé une définition des gènes étendue d'une distance de 50kbs de part et d'autres des bornes des gènes. De nombreux SNPs n'ont donc pas pu être annotés. Cette stratégie, si elle permet par exemple de réduire le nombre d'interactions GxG testées, implique une perte d'information avec l'impossibilité de détecter des interactions entre des SNPs situés dans des régions inter-géniques éloignées de plus de 50kb des bornes d'un gène.

Validation des résultats.

Le travail issu de cette thèse a permis d'identifier des interactions gène-gène dans l'atopie et des interactions gène-environnement dans le risque de survenue de l'asthme dans l'enfance. Dans le cadre de ces études, la validation des résultats par des études de réplication dans des échantillons indépendants est importante. Cependant, la difficulté d'accès à des données de qualité comprenant la même information sur les phénotypes et les facteurs environnementaux étudiés, rend la réplication de ce types de résultats bien plus compliquée que pour des GWAS. Dans notre second projet, nous avons montré une consistance des résultats entre les cinq différentes études incluses dans la méta-analyse. La poursuite de ce travail par des analyses de cartographies fines des régions identifiées peut permettre de mieux les caractériser. La mise à disposition de panels de référence de taille de plus en plus importante (phase 3 du projet 1000 Genomes, étude UK10K, The Haplotype Reference Consortium) et pouvant être combinés pour réaliser des imputations de données, augmente considérablement les possibilités de capturer des variants potentiellement fonctionnels sans avoir recours au séquençage de ces régions dans nos données. Certaines approches adaptées à l'analyse par cartographie fine, permettant d'intégrer directement des connaissances extérieures issues de bases de données d'annotations fonctionnelles à l'analyse de ces régions pour données plus de poids aux variants potentiellement fonctionnels [226]. Notamment, la méthode PAINTOR [227] estime de manière empirique la contribution de chaque annotation fonctionnelle au trait d'intérêt directement à partir de statistiques d'association résumées. Il existe de nombreuses autres méthodes basées sur des méthodes bayésiennes pour l'analyse de cartographie fine [228].

REFERENCES

1. Balding, D. J. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* **7,** 781–791 (2006).

2. Montana, G. Statistical methods in genetics. *Briefings in Bioinformatics* **7,** 297–308 (2006).

3. Génin, E., Feingold, J. & Clerget-Darpoux, F. Identifying modifier genes of monogenic disease: Strategies and difficulties. *Human Genetics* **124,** 357–368 (2008).

4. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517 (1996).

5. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** 2074–2093 (2006).

6. Haiman, C. A. *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* **39,** 638–644 (2007).

7. Patnala, R., Clements, J. & Batra, J. Candidate gene association studies: A comprehensive guide to useful in silico tools. *BMC Genetics* **14,** (2013).

8. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science (80-.).* **308,** 385–389 (2005).

9. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).

10. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* (2005). doi:10.1038/nrg1521

11. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare Variants Create Synthetic Genome-Wide Associations. *PLoS Biol.* **8,** (2010).

12. Altshuler, D. L. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).

13. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48,** 1279–1283 (2016).

14. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,**

(2009).

15. Roshyara, N. R., Horn, K., Kirsten, H., Ahnert, P. & Scholz, M. Comparing performance of modern genotype imputation methods in different ethnicities. *Sci. Rep.* **6,** (2016).

16. Kruglyak, L. Power tools for human genetics. *Nature Genetics* **37,** 1299–1300 (2005).

17. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb).* **95,** 221–227 (2005).

18. Jannot, A. S., Ehret, G. & Perneger, T. P 5 × 10-8 has emerged as a standard of statistical significance for genome-wide association studies. *J. Clin. Epidemiol.* 460–465 (2015). doi:10.1016/j.jclinepi.2015.01.001

19. Mägi, R. & Morris, A. P. GWAMA: Software for genome-wide association meta-analysis. *BMC Bioinformatics* **11,** (2010).

20. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** (2014).

21. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901 (2017).

22. ENCODE Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).

23. Sloan, C. A. *et al.* ENCODE data at the ENCODE portal. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkv1160

24. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).

25. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics* **11,** 446–450 (2010).

26. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci.* **109,** 1193–1198 (2012).

27. Marquardt, D. W. & Snee, R. D. Ridge regression in practice. *Am. Stat.* **29,** 3–20 (1975).

28. Tibshirani, R. Regression shrinkage and selection via the lasso: A retrospective. *J. R.*

Stat. Soc. Ser. B Stat. Methodol. **73,** 273–282 (2011).

29. Efron, B. *et al.* Least angle regression. *Ann. Stat.* **32,** 407–499 (2004).

30. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* (2005). doi:10.1111/j.1467-9868.2005.00503.x

31. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67,** 91–108 (2005).

32. Bach, F. Bolasso: model consistent Lasso estimation through the bootstrap. *Proc. 25th Int. Conf. Mach. Learn. - ICML '08* (2008). doi:10.1145/1390156.1390161

33. Hoggart, C. J., Whittaker, J. C., De Iorio, M. & Balding, D. J. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.* (2008). doi:10.1371/journal.pgen.1000130

34. Ayers, K. L. & Cordell, H. J. SNP Selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet. Epidemiol.* (2010). doi:10.1002/gepi.20543

35. Abdi, H. & Williams, L. J. Partial least squares methods: Partial least squares correlation and partial least square regression. *Methods Mol. Biol.* **930,** 549–579 (2013).

36. Chung, D. & Keles, S. Sparse partial least squares classification for high dimensional data. *Stat. Appl. Genet. Mol. Biol.* **9,** (2010).

37. Lê Cao, K. A., Rossouw, D., Robert-Granié, C. & Besse, P. A sparse PLS for variable selection when integrating omics data. *Stat. Appl. Genet. Mol. Biol.* **7,** (2008).

38. Liquet, B., De Micheaux, P. L., Hejblum, B. P. & Thiébaut, R. Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics* **32,** 35–42 (2015).

39. Bureau, A. *et al.* Identifying SNPs predictive of phenotype using random forests. *Genet. Epidemiol.* **28,** 171–182 (2005).

40. Liu, J. Z. *et al.* A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **87,** 139–45 (2010).

41. Mishra, A. & Macgregor, S. VEGAS2: Software for more flexible gene-based testing. *Twin Res. Hum. Genet.* **18,** 86–91 (2015).

42. Peng, G. *et al.* Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur. J. Hum. Genet.* **18,** 111–117 (2010).

43. Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H. & Weir, B. S. Truncated product method for combining P-values. *Genet. Epidemiol.* **22,** 170–185 (2002).

44. Dudbridge, F. & Koeleman, B. P. C. Rank Truncated Product of P-Values, with Application to Genomewide Association Scans. *Genet. Epidemiol.* **25,** 360–366 (2003).

45. Yu, K. *et al.* Pathway analysis by adaptive combination of P-values. *Genet. Epidemiol.* **33,** 700–709 (2009).

46. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Comput. Biol.* **11,** (2015).

47. Cabrera, C. P. *et al.* Uncovering Networks from Genome-Wide Association Studies via Circular Genomic Permutation. *G3: Genes|Genomes|Genetics* **2,** 1067–1075 (2012).

48. Liu, Y. *et al.* Network-assisted analysis of GWAS data identifies a functionally-relevant gene module for childhood-onset asthma. *Sci. Rep.* **7,** (2017).

49. Wu, M. C. *et al.* Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *Am. J. Hum. Genet.* **86,** 929–942 (2010).

50. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89,** 82–93 (2011).

51. Chen, H., Meigs, J. B. & Dupuis, J. Sequence Kernel Association Test for Quantitative Traits in Family Samples. *Genet. Epidemiol.* **37,** 196–204 (2013).

52. Oualkacha, K. *et al.* Adjusted Sequence Kernel Association Test for Rare Variants Controlling for Cryptic and Family Relatedness. *Genet. Epidemiol.* **37,** 366–376 (2013).

53. Svishcheva, G. R., Belonogova, N. M. & Axenovich, T. I. FFBSKAT: Fast family-based sequence kernel association test. *PLoS One* **9,** (2014).

54. Bader, G. D. Pathguide: a Pathway Resource List. *Nucleic Acids Res.* **34,** D504–D506 (2006).

55. Holmans, P. *Statistical Methods for Pathway Analysis of Genome-Wide Data for Association with Complex Genetic Traits*. *Advances in Genetics* **72,** (2010).

56. Holmans, P. *et al.* Gene Ontology Analysis of GWA Study Data Sets Provides Insights into the Biology of Bipolar Disorder. *Am. J. Hum. Genet.* **85,** 13–24 (2009).

57. Ayellet, V. S., Groop, L., Mootha, V. K., Daly, M. J. & Altshuler, D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* **6,** (2010).

58. Dennis, G. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4,** R60 (2003).

59. Wang, K., Li, M. & Bucan, M. Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* **81,** 1278–83 (2007).

60. Mishra, A. & MacGregor, S. A Novel Approach for Pathway Analysis of GWAS Data Highlights Role of BMP Signaling and Muscle Cell Differentiation in Colorectal Cancer Susceptibility. *Twin Res. Hum. Genet.* **20,** 1–9 (2017).

61. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Comput. Biol.* **12,** (2016).

62. Way, G. P., Youngstrom, D. W., Hankenson, K. D., Greene, C. S. & Grant, S. F. A. Implicating candidate genes at GWAS signals by leveraging topologically associating domains. *Eur. J. Hum. Genet.* **25,** 1286–1289 (2017).

63. Jia, P. & Zhao, Z. Network-assisted analysis to prioritize GWAS results: Principles, methods and perspectives. *Human Genetics* **133,** 125–138 (2014).

64. Liu, Y. *et al.* SigMod: An exact and efficient method to identify a strongly interconnected diseaseassociated module in a gene network. *Bioinformatics* **33,** 1536–1544 (2017).

65. Wei, W. H., Hemani, G. & Haley, C. S. Detecting epistasis in human complex traits. *Nature Reviews Genetics* **15,** 722–733 (2014).

66. Kam-Thong, T. *et al.* EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. *Eur. J. Hum. Genet.* **19,** 465–471 (2011).

67. Wan, X. *et al.* BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.* **87,** 325–340 (2010).

68. Gyenesei, A., Moody, J., Semple, C. A. M., Haley, C. S. & Wei, W. H. High-

throughput analysis of epistasis in genome-wide association studies with BiForce. *Bioinformatics* **28,** 1957–1964 (2012).

69. Yung, L. S., Yang, C., Wan, X. & Yu, W. GBOOST: A GPU-based tool for detecting gene-gene interactions in genome-wide case control studies. *Bioinformatics* **27,** 1309–1310 (2011).

70. Gauderman, W. J. Sample size requirements for association studies of gene-gene interaction. *Am. J. Epidemiol.* **155,** 478–484 (2002).

71. Sun, X. *et al.* Analysis pipeline for the epistasis search - statistical versus biological filtering. *Frontiers in Genetics* **5,** (2014).

72. Marchini, J., Donnelly, P. & Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **37,** 413–417 (2005).

73. Paré, G., Cook, N. R., Ridker, P. M. & Chasman, D. I. On the use of variance per genotype as a tool to identify quantitative trait interaction effects: A report from the women's genome health study. *PLoS Genet.* **6,** 1–10 (2010).

74. Struchalin, M. V., Dehghan, A., Witteman, J. C. M., van Duijn, C. & Aulchenko, Y. S. Variance heterogeneity analysis for detection of potentially interacting genetic loci: Method and its limitations. *BMC Genet.* **11,** (2010).

75. McKinney, B. A., Reif, D. M., White, B. C., Crowe, J. E. & Moore, J. H. Evaporative cooling feature selection for genotypic data involving interactions. *Bioinformatics* **23,** 2113–2120 (2007).

76. Greene, C. S., Penrod, N. M., Kiralis, J. & Moore, J. H. Spatially Uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions. *BioData Min.* **2,** (2009).

77. Sun, X., Elston, R., Morris, N. & Zhu, X. What is the significance of difference in phenotypic variability across SNP genotypes? *Am. J. Hum. Genet.* **93,** 390–397 (2013).

78. Xu, H.-M. *et al.* GMDR: Versatile Software for Detecting Gene-Gene and Gene-Environment Interactions Underlying Complex Traits. *Curr. Genomics* **17,** 396–402 (2016).

79. Xu, H. M. *et al.* Multivariate dimensionality reduction approaches to identify gene-gene and gene-environment interactions underlying multiple complex traits. *PLoS One*

9, (2014).

80. Ritchie, M. D. *et al.* Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. *Am. J. Hum. Genet.* **69,** 138–147 (2001).

81. Nelson, M. R., Kardia, S. L. R., Ferrell, R. E. & Sing, C. F. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res.* **11,** 458–470 (2001).

82. Calle, M. L., Urrea, V., Malats, N. & van Steen, K. Mbmdr: An R package for exploring gene-gene interactions associated with binary or quantitative traits. *Bioinformatics* **26,** 2198–2199 (2010).

83. Zhao, G., Marceau, R., Zhang, D. & Tzeng, J. Y. Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression. *Genetics* **199,** 695–710 (2015).

84. Marceau, R. *et al.* A Fast Multiple-Kernel Method With Applications to Detect Gene-Environment Interaction. *Genet. Epidemiol.* **39,** 456–468 (2015).

85. Ritchie, M. D. Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies. *Ann. Hum. Genet.* **75,** 172–182 (2011).

86. Ma, L., Clark, A. G. & Keinan, A. Gene-Based Testing of Interactions in Association Studies of Quantitative Traits. *PLoS Genet.* **9,** (2013).

87. Brossard, M. *et al.* Integrated pathway and epistasis analysis reveals interactive effect of genetic variants at TERF1 and AFAP1L2 loci on melanoma risk. *Int. J. Cancer* **137,** 1901–1909 (2015).

88. Vaysse, A. *et al.* A comprehensive genome-wide analysis of melanoma Breslow thickness identifies interaction between CDC42 and SCIN genetic variants. *Int. J. Cancer* **139,** 2012–2020 (2016).

89. Bush, W. S., Dudek, S. M. & Ritchie, M. D. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac. Symp. Biocomput.* (2009). doi:10.1016/j.bbi.2008.05.010

90. Cohen, K. B. & Hunter, L. Natural language processing and systems biology. *Artif. Intell. Methods Tools Syst. Biol.* (2004). doi:10.1007/1-4020-2865-2_9

91. Hunter, L. & Cohen, K. B. Biomedical language processing: What's beyond PubMed? *Mol. Cell* **21,** 589–594 (2006).

92. Luo, Y., Riedlinger, G. & Szolovits, P. Text Mining in Cancer Gene and Pathway Prioritization. *Cancer Inform.* (2014). doi:10.4137/CIN.S13874

93. Chen, J., Bardes, E. E., Aronow, B. J. & Jegga, A. G. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* (2009). doi:10.1093/nar/gkp427

94. Chen, J., Xu, H., Aronow, B. J. & Jegga, A. G. Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics* (2007). doi:10.1186/1471-2105-8-392

95. van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G. & Leunissen, J. A. M. A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.* (2006). doi:10.1038/sj.ejhg.5201585

96. Berger, S. I., Posner, J. M. & Ma'ayan, A. Genes2Networks: Connecting lists of gene symbols using mammalian protein interactions databases. *BMC Bioinformatics* (2007). doi:10.1186/1471-2105-8-372

97. Seelow, D., Schwarz, J. M. & Schuelke, M. GeneDistiller--distilling candidate genes from linkage intervals. *PLoS One* (2008). doi:10.1371/journal.pone.0003874

98. van Driel, M. A. *et al.* GeneSeeker: Extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res.* (2005). doi:10.1093/nar/gki435

99. Perez-Iratxeta, C., Bork, P. & Andrade, M. A. Association of genes to genetically inherited diseases using data mining. *Nat. Genet.* (2002). doi:10.1038/ng895

100. Perez-Iratxeta, C., Bork, P. & Andrade-Navarro, M. A. Update of the G2D tool for prioritization of gene candidates to inherited diseases. *Nucleic Acids Res.* (2007). doi:10.1093/nar/gkm223

101. Perez-Iratxeta, C., Wjst, M., Bork, P. & Andrade, M. A. G2D: A tool for mining genes associated with disease. *BMC Genet.* (2005). doi:10.1186/1471-2156-6-45

102. Yue, P., Melamud, E. & Moult, J. SNPs3D: Candidate gene and SNP selection for association studies. *BMC Bioinformatics* (2006). doi:10.1186/1471-2105-7-166

103. Gaulton, K. J., Mohlke, K. L. & Vision, T. J. A computational system to select candidate genes for complex human traits. *Bioinformatics* (2007). doi:10.1093/bioinformatics/btm001

104. Hamosh, A. *et al.* Onlined Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* (2002). doi:10.1093/nar/gki033

105. Smith, C. L., Goldsmith, C. W. & Eppig, J. T. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.* (2005). doi:10.1186/gb-2004-6-1-r7

106. Kelso, J. *et al.* eVOC: A controlled vocabulary for unifying gene expression data. *Genome Res.* (2003). doi:10.1101/gr.985203

107. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkw1108

108. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: Predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* (2009). doi:10.1371/journal.pgen.1000534

109. Ailem, M., Role, F., Nadif, M. & Demenais, F. Unsupervised text mining for assessing and augmenting GWAS results. *J. Biomed. Inform.* (2016). doi:10.1016/j.jbi.2016.02.008

110. Tranchevent, L. C. *et al.* Candidate gene prioritization with Endeavour. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw365

111. Aerts, S. *et al.* Gene prioritization through genomic data fusion. *Nat. Biotechnol.* (2006). doi:10.1038/nbt1203

112. Zhou, J. & Fu, B. quan. The research on gene-disease association based on text-mining of PubMed. *BMC Bioinformatics* (2018). doi:10.1186/s12859-018-2048-y

113. Johansson, M. *et al.* Using Prior Information from the Medical Literature in GWAS of Oral Cancer Identifies Novel Susceptibility Variant on Chromosome 4 -the AdAPT Method. *PLoS ONE | www.plosone.org 1* **7,** (2012).

114. Wei, P., Tang, H. & Li, D. Functional logistic regression approach to detecting gene by longitudinal environmental exposure interaction in a case-control study. *Genet. Epidemiol.* **38,** 638–651 (2014).

115. Lindström, S., Yen, Y. C., Spiegelman, D. & Kraft, P. The impact of gene-environment dependence and misclassification in genetic association studies incorporating gene-environment interactions. *Hum. Hered.* **68,** 171–181 (2009).

116. Ober, C. & Vercelli, D. Gene-environment interactions in human disease: Nuisance or opportunity? *Trends in Genetics* (2011). doi:10.1016/j.tig.2010.12.004

117. Piegorsch, W. W., Weinberg, C. R. & Taylor, J. A. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat. Med.* **13,** 153–162 (1994).

118. Mukherjee, B. & Chatterjee, N. Exploiting gene-environment independence for analysis of case-control studies: An empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* **64,** 685–694 (2008).

119. Chen, Y. H., Chatterjee, N. & Carroll, R. J. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *J. Am. Stat. Assoc.* **104,** 220–233 (2009).

120. Li, D. & Conti, D. V. Detecting gene-environment interactions using a combined case-only and case-control approach. *Am. J. Epidemiol.* **169,** 497–504 (2009).

121. Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J. & Gauderman, W. J. Exploiting gene-environment interaction to detect genetic associations. *Hum. Hered.* (2007). doi:10.1159/000099183

122. Dai, J. Y. *et al.* Simultaneously Testing for Marginal Genetic Association and Gene-Environment Interaction. *Am. J. Epidemiol.* **176,** 164–173 (2012).

123. Hsu, L. *et al.* Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genet. Epidemiol.* (2012). doi:10.1002/gepi.21610

124. Kazma, R., Babron, M. C. & Génin, E. Genetic association and gene-environment interaction: A new method for overcoming the lack of exposure information in controls. *Am. J. Epidemiol.* **173,** 225–235 (2011).

125. Ionita-Laza, I., McQueen, M. B., Laird, N. M. & Lange, C. Genomewide Weighted Hypothesis Testing in Family-Based Association Studies, with an Application to a 100K Scan. *Am. J. Hum. Genet.* **81,** 607–614 (2007).

126. Roeder, K., Devlin, B. & Wasserman, L. Improving power in genome-wide association

studies: Weights tip the scale. *Genet. Epidemiol.* **31,** 741–747 (2007).

127. Roeder, K. & Wasserman, L. Genome-Wide Significance Levels and Weighted Hypothesis Testing. *Stat Sci.* **24,** 398–413 (2009).

128. Kooperberg, C. & LeBlanc, M. Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genet. Epidemiol.* (2008). doi:10.1002/gepi.20300

129. Murcray, C. E., Lewinger, J. P. & Gauderman, W. J. Gene-environment interaction in genome-wide association studies. *Am. J. Epidemiol.* **169,** 219–226 (2009).

130. Murcray, C. E., Lewinger, J. P., Conti, D. V., Thomas, D. C. & Gauderman, W. J. Sample size requirements to detect gene-environment interactions in genome-wide association studies. *Genet. Epidemiol.* **35,** 201–210 (2011).

131. Ege, M. J. *et al.* Gene-environment interaction for childhood asthma and exposure to farming in Central Europe. *J. Allergy Clin. Immunol.* (2011). doi:10.1016/j.jaci.2010.09.041

132. Jiao, S. *et al.* SBERIA: Set-Based Gene-Environment Interaction Test for Rare and Common Variants in Complex Diseases. *Genet. Epidemiol.* **37,** 452–464 (2013).

133. Liu, Q., Chen, L. S., Nicolae, D. L. & Pierce, B. L. A unified set-based test with adaptive filtering for gene-environment interaction analyses. *Biometrics* **72,** 629–638 (2016).

134. Hüls, A. *et al.* Comparison of weighting approaches for genetic risk scores in gene-environment interaction studies. *BMC Genet.* **18,** 1–12 (2017).

135. Tzeng, J. Y. *et al.* Studying gene and gene-environment effects of uncommon and common variants on continuous traits: A marker-set approach using gene-trait similarity regression. *Am. J. Hum. Genet.* **89,** 277–288 (2011).

136. Lin, X., Lee, S., Christiani, D. C. & Lin, X. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* **14,** 667–681 (2013).

137. Su, Y. R., Di, C. Z. & Hsu, L. A unified powerful set-based test for sequencing data analysis of GxE interactions. *Biostatistics* **18,** 119–131 (2017).

138. Jiao, S. *et al.* Powerful Set-Based Gene-Environment Interaction Testing Framework

for Complex Diseases. *Genet. Epidemiol.* **39,** 609–618 (2015).

139. Sun, J., Zheng, Y. & Hsu, L. A Unified Mixed-Effects Model for Rare-Variant Association in Sequencing Studies. *Genet. Epidemiol.* **37,** 334–344 (2013).

140. Broadaway, K. A. *et al.* Kernel Approach for Modeling Interaction Effects in Genetic Association Studies of Complex Quantitative Traits. *Genet. Epidemiol.* **39,** 366–375 (2015).

141. Gauderman, W. J. *et al.* Update on the State of the Science for Analytical Methods for Gene-Environment Interactions. *Am. J. Epidemiol.* **186,** 762–770 (2017).

142. Ege, M. J. & Strachan, D. P. Comparisons of power of statistical methods for gene-environment interaction analyses. *Eur. J. Epidemiol.* **28,** 785–797 (2013).

143. Higgins, J. P. T. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21,** 1539–1558 (2002).

144. Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F. & Botella, J. Assessing heterogeneity in meta-analysis: Q statistic or I2Index? *Psychol. Methods* **11,** 193–206 (2006).

145. Aschard, H., Hancock, D. B., London, S. J. & Kraft, P. Genome-wide meta-analysis of joint tests for genetic and gene-environment interaction effects. *Hum. Hered.* (2011). doi:10.1159/000323318

146. Manning, A. K. *et al.* Meta-analysis of Gene-Environment interaction: joint estimation of SNP and SNP×Environment regression coefficients. *Genet. Epidemiol.* (2012). doi:10.1002/gepi.20546.Meta-analysis

147. WHO media Center. Asthma fact sheets, from http://www.who.int/mediacentre/factsheets/fs307/en/. (2017).

148. Wenzel, S. E. Asthma phenotypes: The evolution from clinical to molecular approaches. *Nature Medicine* (2012). doi:10.1038/nm.2678

149. The Global Asthma Report 2018. Auckland, New Zealand: Global Asthma Network, 2018.

150. Asher, M. I. *et al.* International Study of Asthma and Allergies in Childhood (ISAAC): rationale and methods. *Eur. Respir. J.* **8,** 483–91 (1995).

151. Burney, P. Variations in the prevalence of respiratory symptoms, self-reported asthma

attacks, and use of asthma medication in the European Community Respiratory Health Survey (ECRHS). *Eur. Respir. J.* **9,** 687–695 (1996).

152. Ober, C. & Yao, T. C. The genetics of asthma and allergic disease: A 21st century perspective. *Immunol. Rev.* **242,** 10–30 (2011).

153. Ding, G., Ji, R. & Bao, Y. Risk and Protective Factors for the Development of Childhood Asthma. *Paediatric Respiratory Reviews* **16,** 133–139 (2015).

154. Abdulrazzaq, Y. M., Bener, A. & DeBuse, P. Association of allergic symptoms in children with those in their parents. *Allergy* **49,** 737–743 (1994).

155. Holberg, C. J., Morgan, W. J., Wright, A. L. & Martinez, F. D. Differences in familial segregation of FEV1between asthmatic and nonasthmatic families: Role of a maternal component. *Am. J. Respir. Crit. Care Med.* **158,** 162–169 (1998).

156. Litonjua, A. A., Carey, V. J., Burge, H. A., Weiss, S. T. & Gold, D. R. Parental history and the risk for childhood asthma: Does mother confer more risk than father? *Am. J. Respir. Crit. Care Med.* **158,** 176–181 (1998).

157. Ober, C. & Vercelli, D. Gene-environment interactions in human disease: Nuisance or opportunity? *Trends in Genetics* **27,** 107–115 (2011).

158. Burke, H. *et al.* Prenatal and Passive Smoke Exposure and Incidence of Asthma and Wheeze: Systematic Review and Meta-analysis. *Pediatrics* **129,** 735–744 (2012).

159. Jaakkola, J. J. K. & Gissler, M. Maternal smoking in pregnancy, fetal development, and childhood asthma. *Am. J. Public Health* (2004). doi:10.2105/AJPH.94.1.136

160. Skorge, T. D., Eagan, T. M. L., Eide, G. E., Gulsvik, A. & Bakke, P. S. The adult incidence of asthma and respiratory symptoms by passive smoking in utero or in childhood. *Am. J. Respir. Crit. Care Med.* **172,** 61–66 (2005).

161. Gilliland, F. D. *et al.* Regular smoking and asthma incidence in adolescents. *Am. J. Respir. Crit. Care Med.* **174,** 1094–1100 (2006).

162. Hedman, L., Bjerg, A., Sundberg, S., Forsberg, B. & Rönmark, E. Both environmental tobacco smoke and personal smoking is related to asthma and wheeze in teenagers. *Thorax* **66,** 20–25 (2011).

163. Vignoud, L. *et al.* Smoking and asthma: Disentangling their mutual influences using a longitudinal approach. *Respir. Med.* **105,** 1805–1814 (2011).

164. Daley, D. The evolution of the hygiene hypothesis: The role of early-life exposures to viruses and microbes and their relationship to asthma and allergic diseases. *Current Opinion in Allergy and Clinical Immunology* **14,** 390–396 (2014).

165. Chen, Y., Dales, R., Tang, M. & Krewski, D. Obesity may increase the incidence of asthma in women but not in men: Longitudinal observations from the Canadian National Population Health Surveys. *Am. J. Epidemiol.* **155,** 191–197 (2002).

166. Gilliland, F. D. *et al.* Obesity and the risk of newly diagnosed asthma in school-age children. *Am. J. Epidemiol.* **158,** 406–415 (2003).

167. Schwartz, J. Air Pollution and Children's Health. *Pediatrics* **113,** 1037–1043 (2004).

168. Okada, H., Kuhn, C., Feillet, H. & Bach, J. F. The 'hygiene hypothesis' for autoimmune and allergic diseases: An update. *Clinical and Experimental Immunology* **160,** 1–9 (2010).

169. Willemsen, G., Van Beijsterveldt, T. C. E. M., Van Baal, C. G. C. M., Postma, D. & Boomsma, D. I. Heritability of self-reported asthma and allergy: A study in adult Dutch twins, siblings and parents. *Twin Res. Hum. Genet.* (2008). doi:10.1375/twin.11.2.132

170. Duffy, D. L., Martin, N. G., Battistutta, D., Hopper, J. L. & Mathews, J. D. Genetics of Asthma and Hay Fever in Australian Twins. *Am. Rev. Respir. Dis.* **142,** 1351–1358 (1990).

171. Los, H., Postmus, P. E. & Boomsma, D. I. Asthma genetics and intermediate phenotypes: A review from twin studies. *Twin Research* **4,** 81–93 (2001).

172. Thomsen, S. F., Ulrik, C. S., Kyvik, K. O., Ferreira, M. A. R. & Backer, V. Multivariate genetic analysis of atopy phenotypes in a selected sample of twins. *Clin. Exp. Allergy* **36,** 1382–1390 (2006).

173. Los, H., Koppelman, G. H. & Postma, D. S. The importance of genetic influences in asthma. *Eur. Respir. J.* (1999). doi:10.1183/09031936.99.14512109

174. Bouzigon, E., Demenais, F. & Kauffmann, F. [Genetics of asthma and atopy: how many genes?]. *Bull Acad Natl Med* **189,** 1435–1448 (2005).

175. Bouzigon, E. *et al.* Meta-analysis of 20 genome-wide linkage studies evidenced new regions linked to asthma and atopy. *Eur. J. Hum. Genet.* (2010). doi:10.1038/ejhg.2009.224

176. March, M. E., Sleiman, P. M. & Hakonarson, H. Genetic polymorphisms and associated susceptibility to asthma. *Int. J. Gen. Med.* (2013). doi:10.2147/IJGM.S28156

177. Halapi, E. & Hakonarson, H. Recent development in genomic and proteomic research for asthma. *Current Opinion in Pulmonary Medicine* **10,** 22–30 (2004).

178. Kabesch, M. Candidate gene association studies and evidence for gene-by-gene interactions. *Immunol. Allergy Clin. North Am.* **25,** 681–708 (2005).

179. Levy, H. *et al.* Association of defensin β-1 gene polymorphisms with asthma. *J. Allergy Clin. Immunol.* **115,** 252–258 (2005).

180. Moffatt, M. F. *et al.* Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448,** 470–473 (2007).

181. Bouzigon, E. *et al.* Effect of 17q21 Variants and Smoking Exposure in Early-Onset Asthma. *N. Engl. J. Med.* **359,** 1985–1994 (2008).

182. Kurukulaaratchy, R. J. *et al.* Characterization of wheezing phenotypes in the first 10 years of life. *Clin. Exp. Allergy* (2003). doi:10.1046/j.1365-2222.2003.01657.x

183. Moffatt, M. F. *et al.* A Large-Scale, Consortium-Based Genomewide Association Study of Asthma. *N. Engl. J. Med.* **363,** 1211–1221 (2010).

184. Torgerson, D. G. *et al.* Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat. Genet.* **43,** 887–892 (2011).

185. Demenais, F. *et al.* Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat. Genet.* **50,** 42–53 (2018).

186. Zhu, Z. *et al.* A genome-wide cross-trait analysis from UK Biobank highlights the shared genetic architecture of asthma and allergic diseases. *Nat. Genet.* **50,** 857–864 (2018).

187. Vercelli, D. Discovering susceptibility genes for asthma and allergy. *Nature Reviews Immunology* **8,** 169–182 (2008).

188. Wan, Y. I. *et al.* A genome-wide association study to identify genetic determinants of atopy in subjects from the United Kingdom. *J. Allergy Clin. Immunol.* (2011). doi:10.1016/j.jaci.2010.10.006

189. Bønnelykke, K. *et al.* Meta-analysis of genome-wide association studies identifies ten

loci influencing allergic sensitization. *Nat. Genet.* **45,** 902–906 (2013).

190. Hinds, D. A. *et al.* A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. *Nat. Genet.* **45,** 907–911 (2013).

191. Waage, J. *et al.* Genome-wide association and HLA fine-mapping studies identify risk loci and genetic pathways underlying allergic rhinitis. *Nature Genetics* **50,** 1072–1080 (2018).

192. Paternoster, L. *et al.* Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat. Genet.* **47,** 1449–1456 (2015).

193. Kontakioti, E., Domvri, K., Papakosta, D. & Daniilidis, M. HLA and asthma phenotypes/endotypes: A review. *Human Immunology* **75,** 930–939 (2014).

194. Ferreira, M. A. *et al.* Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat. Genet.* (2017). doi:10.1038/ng.3985

195. Ferreira, M. A. R. *et al.* Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype. *J. Allergy Clin. Immunol.* **133,** 1564–1571 (2014).

196. Scholtens, S. *et al.* Novel childhood asthma genes interact with in utero and early-life tobacco smoke exposure. *Journal of Allergy and Clinical Immunology* **133,** 885–888 (2014).

197. Vonk, J. M. *et al.* Adult onset asthma and interaction between genes and active tobacco smoking: The GABRIEL consortium. *PLoS One* (2017). doi:10.1371/journal.pone.0172716

198. Gref, A. *et al.* Genome-wide interaction analysis of air pollution exposure and childhood asthma with functional follow-up. *Am. J. Respir. Crit. Care Med.* **195,** 1373–1383 (2017).

199. Ferreira, M. A. R. *et al.* Identification of IL6R and chromosome 11q13.5 as risk loci for asthma. *Lancet* **378,** 1006–1014 (2011).

200. Granada, M. *et al.* A genome-wide association study of plasma total IgE concentrations in the Framingham Heart Study. *J. Allergy Clin. Immunol.* **129,** (2012).

201. Weidinger, S. *et al.* Genome-wide scan on total serum IgE levels identifies FCER1A as novel susceptibility locus. *PLoS Genet.* **4,** (2008).

202. Sleiman, P. M. A. *et al.* Variants of DENND1B Associated with Asthma in Children. *N. Engl. J. Med.* **362,** 36–44 (2010).

203. Forno, E. *et al.* Genome-wide association study of the age of onset of childhood asthma. *J. Allergy Clin. Immunol.* (2012). doi:10.1016/j.jaci.2012.03.020

204. Ding, L. *et al.* Rank-based genome-wide analysis reveals the association of Ryanodine receptor-2 gene variants with childhood asthma among human populations. *Hum. Genomics* **7,** (2013).

205. Himes, B. E. *et al.* Genome-wide Association Analysis Identifies PDE4D as an Asthma-Susceptibility Gene. *Am. J. Hum. Genet.* **84,** 581–593 (2009).

206. Yatagai, Y. *et al.* Genome-wide association study for levels of total serum IgE identifies HLA-C in a Japanese population. *PLoS One* **8,** (2013).

207. Noguchi, E. *et al.* Genome-wide association study identifies HLA-DP as a susceptibility gene for pediatric asthma in Asian populations. *PLoS Genet.* **7,** (2011).

208. Ramasamy, A. *et al.* Genome-Wide Association Studies of Asthma in Population-Based Cohorts Confirm Known and Suggested Loci and Identify an Additional Association near HLA. *PLoS One* **7,** (2012).

209. Lasky-Su, J. *et al.* HLA-DQ strikes again: Genome-wide association study further confirms HLA-DQ in the diagnosis of asthma among adults. *Clin. Exp. Allergy* **42,** 1724–1733 (2012).

210. Galanter, J. M. *et al.* Genome-wide association study and admixture mapping identify different asthma-associated loci in Latinos: The Genes-environments & Admixture in Latino Americans study. *J. Allergy Clin. Immunol.* **134,** 295–305 (2014).

211. Bønnelykke, K. *et al.* A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat. Genet.* **46,** 51–55 (2014).

212. Wan, Y. I. *et al.* Genome-wide association study to identify genetic determinants of severe asthma. *Thorax* **67,** 762–768 (2012).

213. Ramasamy, A. *et al.* A genome-wide meta-analysis of genetic variants associated with

allergic rhinitis and grass sensitization and their interaction with birth order. *J. Allergy Clin. Immunol.* **128,** 996–1005 (2011).

214. Kauffmann, F. *et al.* Epidemiological study of the genetics and Environment of Asthma, bronchial hyperresponsiveness and atopy (EGEA). Protocol and potential selection factors. *Rev Epidemiol Sante Publique* **49,** 343–356 (1997).

215. Laprise, C. The Saguenay-Lac-Saint-Jean asthma familial collection: the genetics of asthma in a young founder population. *Genes Immun.* **15,** 247–255 (2014).

216. Fraser, A. *et al.* Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int. J. Epidemiol.* (2013). doi:10.1093/ije/dys066

217. Golding, J. Children of the nineties. A longitudinal study of pregnancy and childhood based on the population of Avon (ALSPAC). *West Engl Med J* (1990).

218. Burney, P. G. J., Luczynska, C., Chinn, S. & Jarvis, D. The European Community Respiratory Health Survey. *Eur. Respir. J.* (1994). doi:10.1183/09031936.94.07050954

219. The European Community Respiratory Health Survey II Steering Committee. The European Community Respiratory Health Survey II. *Eur. Respir. J.* (2002). doi:10.1183/09031936.02.00046802

220. Genuneit, J. *et al.* The GABRIEL Advanced Surveys: Study design, participation and evaluation of bias. *Paediatric and Perinatal Epidemiology* (2011). doi:10.1111/j.1365-3016.2011.01223.x

221. Loss, G. *et al.* The protective effect of farm milk consumption on childhood asthma and atopy: The GABRIELA study. *J. Allergy Clin. Immunol.* (2011). doi:10.1016/j.jaci.2011.07.048

222. The international HapMap consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).

223. Siroux, V. *et al.* Genetic heterogeneity of asthma phenotypes identified by a clustering approach. *Eur. Respir. J.* (2014). doi:10.1183/09031936.00032713

224. Schlicker, A., Lengauer, T. & Albrecht, M. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. in *Bioinformatics* (2011). doi:10.1093/bioinformatics/btq384

225. Pers, T. H., Dworzyński, P., Thomas, C. E., Lage, K. & Brunak, S. MetaRanker 2.0: a

web server for prioritization of genetic variation data. *Nucleic Acids Res.* (2013). doi:10.1093/nar/gkt387

226. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics* **19,** 491–504 (2018).

227. Kichaev, G. *et al.* Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLoS Genet.* **10,** (2014).

228. Dadaev, T. *et al.* Fine-mapping of prostate cancer susceptibility loci in a large meta-analysis identifies candidate causal variants. *Nat. Commun.* **9,** (2018).

229. Lewontin, R. C. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics* **49,** 49–67 (1964).

230. Hill, W. G. & Robertson, a. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* (1968). doi:10.1007/BF01245622

231. Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics* **15,** 335–346 (2014).

232. Hochberg, Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika* **75,** 800–802 (1988).

233. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57,** 289–300 (1995).

234. Li, M. X., Yeung, J. M. Y., Cherny, S. S. & Sham, P. C. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* (2012). doi:10.1007/s00439-011-1118-2

235. Larsson, K. *et al.* Text mining for improved exposure assessment. *PLoS One* **12,** (2017).

236. Hearst, M. A. Untangling text data mining. in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* - 3–10 (1999). doi:10.3115/1034678.1034679

237. Raja, K. *et al.* A Review of Recent Advancement in Integrating Omics Data with Literature Mining towards Biomedical Discoveries. *International Journal of Genomics*

(2017). doi:10.1155/2017/6213474

238. Baeza-Yates, R. & Ribeiro-Neto, B. Modern Information Retrieval: The Concepts and Technology behind Search. *Inf. Retr. Boston.* **82,** 944 (2011).

239. Lin, Y., Li, W., Chen, K. & Liu, Y. A Document Clustering and Ranking System for Exploring MEDLINE Citations. *J. Am. Med. Informatics Assoc.* (2007). doi:10.1197/jamia.M2215

240. Darmoni, S. J. *et al.* Improving information retrieval using Medical Subject Headings Concepts: a test case on rare and chronic diseases. *J. Med. Libr. Assoc.* (2012). doi:10.3163/1536-5050.100.3.007

241. Petrova, M., Sutcliffe, P., Fulford, K. W. M. & Dale, J. Search terms and a validated brief search filter to retrieve publications on health-related values in Medline: A word frequency analysis study. *J. Am. Med. Informatics Assoc.* (2012). doi:10.1136/amiajnl-2011-000243

242. Leaman, R. & Gonzalez, G. BANNER: an executable survey of advances in biomedical named entity recognition. in *Biocomputing 2008* doi:10.1142/9789812776136_0062

243. Raja, K., Subramani, S. & Natarajan, J. PPInterFinder - A mining tool for extracting causal relations on human proteins from literature. *Database* (2013). doi:10.1093/database/bas052

ANNEXE

Identification of a new locus at 16q12 associated with time to asthma onset

Chloé Sarnowski, PhD,[a,b] Pierre-Emmanuel Sugier, MSc,[a,b] Raquel Granell, PhD,[c] Debbie Jarvis, MD,[d,e] Marie-Hélène Dizier, PhD,[a,b] Markus Ege, MD,[f,g] Medea Imboden, PhD,[h,i] Catherine Laprise, PhD,[j] Elza K. Khusnutdinova, PhD,[k,l] Maxim B. Freidin, PhD,[m] William O. C. Cookson, MD, DPhil,[n] Miriam Moffatt, DPhil,[n] Mark Lathrop, PhD,[o] Valérie Siroux, PhD,[p,q,r] Ludmila M. Ogorodova, MD, PhD,[s] Alexandra S. Karunas, MD, PhD,[k,l] Alan James, MD,[t] Nicole M. Probst-Hensch, PhD,[h,i] Erika von Mutius, MD,[f,g] Isabelle Pin, MD, PhD,[p,q,u] Manolis Kogevinas, MD, PhD,[v,w,x,y] A. John Henderson, MD,[c] Florence Demenais, MD,[a,b] and Emmanuelle Bouzigon, MD, PhD[a,b] *Paris and Grenoble, France; Bristol and London, United Kingdom; Munich, Germany; Basel, Switzerland; Saguenay and Montreal, Quebec, Canada; Ufa and Tomsk, Russia; Nedlands and Crawley, Australia; and Barcelona and Madrid, Spain*

Background: Asthma is a heterogeneous disease in which age of onset plays an important role.
Objective: We sought to identify the genetic variants associated with time to asthma onset (TAO).
Methods: We conducted a large-scale meta-analysis of 9 genome-wide association studies of TAO (total of 5462 asthmatic patients with a broad range of age of asthma onset and 8424 control subjects of European ancestry) performed by using survival analysis techniques.
Results: We detected 5 regions associated with TAO at the genome-wide significant level ($P < 5 \times 10^{-8}$). We evidenced a new locus in the 16q12 region (near cylindromatosis turban tumor syndrome gene *[CYLD]*) and confirmed 4 asthma risk regions: 2q12 (IL-1 receptor–like 1 *[IL1RL1]*), 6p21 *(HLA-DQA1)*, 9p24 *(IL33)*, and 17q12-q21 (zona pellucida binding protein 2 *[ZPBP2]*–gasdermin A *[GSDMA]*). Conditional analyses identified 2 distinct signals at 9p24 (both upstream of *IL33*) and 17q12-q21 (near *ZPBP2* and within *GSDMA*). Together, these 7 distinct loci explained 6.0% of the variance in TAO. In addition, we showed that genetic variants at 9p24 and 17q12-q21 were strongly associated with an earlier onset of childhood asthma ($P \leq .002$), whereas the 16q12 single nucleotide polymorphism was associated with later asthma onset ($P = .04$). A high burden of disease risk alleles at these loci was associated with earlier age of asthma onset (4 vs 9-12 years, $P = 10^{-4}$).
Conclusion: The new susceptibility region for TAO at 16q12 harbors variants that correlate with the expression of *CYLD* and nucleotide-binding oligomerization domain 2 *(NOD2)*, 2 strong candidates for asthma. This study demonstrates that incorporating the variability of age of asthma onset in asthma modeling is a helpful approach in the search for disease susceptibility genes. (J Allergy Clin Immunol 2016;138:1071-80.)

Key words: *Asthma, age of onset, genetics, genome-wide association study, survival analysis, conditional analysis,* CYLD, NOD2

From [a]Inserm, UMR-946, Paris; [b]Université Paris Diderot, Sorbonne Paris Cité, Institut Universitaire d'Hématologie, Paris; [c]the School of Social and Community Medicine, University of Bristol; [d]Respiratory Epidemiology, Occupational Medicine and Public Health, National Heart and Lung Institute, Imperial College, London; [e]MRC-PHE Centre for Environment & Health, London; [f]Dr von Hauner Children's Hospital, Ludwig Maximilian University, Munich; [g]Comprehensive Pneumology Center Munich (CPC-M), German Center for Lung Research, Munich; [h]the Swiss Tropical and Public Health Institute, Basel; [i]the University of Basel; [j]Département des sciences fondamentales, Université du Québec à Chicoutimi, Saguenay; [k]the Institute of Biochemistry and Genetics, Ufa Scientific Centre, Russian Academy of Sciences, Ufa; [l]the Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa; [m]the Research Institute for Medical Genetics, Tomsk; [n]the National Heart Lung Institute, Imperial College London; [o]the McGill University and Génome Québec Innovation Centre, Montreal; [p]Université Grenoble Alpes, IAB, Team of Environmental Epidemiology applied to Reproduction and Respiratory Health, Grenoble; [q]Inserm and [r]CHU de Grenoble, IAB, Team of Environmental Epidemiology applied to Reproduction and Respiratory Health, Grenoble; [s]Siberian State Medical University, Tomsk; [t]the Busselton Population Medical Research Institute, Department of Pulmonary Physiology and Sleep Medicine, Sir Charles Gairdner Hospital, Nedlands, and the School of Population Health, University of Western Australia, Crawley; [u]CHU de Grenoble, Pediatrics, Grenoble; [v]the Centre for Research in Environmental Epidemiology (CREAL), Barcelona; [w]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid; [x]IMIM (Hospital del Mar Medical Research Institute), Barcelona; and [y]Universitat Pompeu Fabra, Barcelona.

Supported by the French National Agency for Research (ANR-CES-2009, ANR-11-BSV1-027-GWIS-AM), Région Ile-de-France (DIM-SEnT grant), "Fonds de dotation Recherche en Santé Respiratoire," the Russian Foundation for Basic Research (grants 13-04-01397 and 01-04-48213a), Healthway and the Departments of Science and Health of the Government of Western Australia, the UK Medical Research Council, the Wellcome Trust (grant 102215/2/13/2), the University of Bristol, and the Swiss National Science Foundation (current grants no 33CS30-148470/1). The Canada Research Chair held by C.L. and the funding supports from Canadian Institutes of Health Research (CIHR) enabled the maintenance and continuation of the SLSJ asthma study. Genotyping was supported by grants from the European Commission (no. LSHB-CT-2006-018996-GABRIEL) and the Wellcome Trust (WT084703MA).

Disclosure of potential conflict of interest: D. Jarvis has received a grant from the European Commission. M. Ege declares receiving a grant from the European Commission. E. K. Khusnutdinova declares receiving a grant, support for travel, and provision of writing support, medicines, equipment, or administrative support from the Commission of the European Communities, Integrated Project GABRIEL. V. Siroux declares providing consultancy to Edimark Santé and TEVA. A. S. Karunas declares receiving a grant, travel support, and provision of writing assistance, medicines, equipment, or administrative support from the Commission of the European Communities and a grant, travel support, and provision of writing assistance, medicines, equipment, or administrative support from the Russian Federation for Basic Research. E. von Mutius declares receiving a grant from the European Commission, European Research Council. I. Pin declares receiving payment for lectures and travel/accommodations/meeting expenses from GlaxoSmithKline and Novartis. A. J. Henderson declares receiving a grant from the Medical Research Council and a grant from Wellcome Trust. The rest of the authors declare that they have no relevant conflicts of interest.

Received for publication August 14, 2015; revised February 5, 2016; accepted for publication March 16, 2016.

Available online April 6, 2016.

Corresponding author: Emmanuelle Bouzigon, MD, PhD, UMR-946, INSERM/Université Paris-Diderot, Institut de Génétique Moléculaire, 27 rue Juliette Dodu, 75010 Paris, France. E-mail: emmanuelle.bouzigon@inserm.fr.

The CrossMark symbol notifies online readers when updates have been made to the article such as errata or minor corrections

0091-6749/$36.00

http://dx.doi.org/10.1016/j.jaci.2016.03.018

Abbreviations used
CYLD: Cylindromatosis (turban tumor syndrome)
eQTL: Expression quantitative trait locus
GSDMA: Gasdermin A
GWAS: Genome-wide association study
IL1RL1: IL-1 receptor–like 1
LCL: Lymphoblastoid cell line
NFkB1: Nuclear factor of kappa light polypeptide gene enhancer in B cells 1
NOD2: Nucleotide-binding oligomerization domain containing 2
QC: Quality control
SNP: Single nucleotide polymorphism
TAO: Time to asthma onset
ZPBP2: Zona pellucida binding protein 2

The prevalence of asthma has dramatically increased over the past decades in high-income countries, affecting 5% to 16% of persons worldwide.[1] It is the most common chronic disease among children, and a decrease in age of asthma onset has been documented recently.[2]

Asthma is a complex and heterogeneous disease with variable clinical expression over the lifespan.[1] It is now well recognized that asthma is not a single disease but rather a collection of different phenotypes that might represent different manifestations of a common underlying pathologic process or might be separate disease entities.[3] One of the simplest characteristics that can be used to differentiate disease phenotypes is age at onset.[4,5] Indeed, asthma displays different characteristics according to the lifetime period during which it occurs.[6] Early age of onset is more frequently associated with a family history of asthma, allergy sensitization, and clinical response to triggers, whereas late-onset disease is associated with eosinophilic inflammation and obesity, more common in women, and generally less allergic.[3]

The risk of asthma has a strong genetic component, with estimated heritability ranging from 35% to 95%.[7] Genome-wide association studies (GWASs) have been successful in identifying more than 20 loci associated with asthma.[8] However, the genetic factors identified to date account only for a small part of the genetic component of the disease.[1] This hidden heritability might be linked to the phenotypic heterogeneity of asthma.[9] The vast majority of GWASs conducted until now have analyzed asthma as a binary phenotype. A few genetic studies have considered a more specific definition of asthma incorporating the age of disease onset. A genome-wide linkage screen conducted for time to asthma onset (TAO) in French families revealed 2 regions, 1p31 and 5q13, potentially linked to this phenotype.[10] A single GWAS has been performed on age of asthma onset in asthmatic children and led to the identification of 2 loci not found by the previous asthma GWASs; these loci on chromosomes 3p26 and 11q24 were associated with an earlier onset of childhood asthma.[11] Moreover, the effect of 17q12-q21 genetic variants identified by the first GWAS of asthma[12] was found to be restricted to early-onset asthma.[13,14]

Instead of stratifying the data according to age of disease onset with an arbitrary threshold, one can integrate the age of onset in modeling asthma risk by using survival analytic methodologies applied to both asthmatic and nonasthmatic subjects. The goal of the present study was to identify the genetic determinants underlying TAO in a large meta-analysis of 5462 asthmatic patients and 8424 control subjects from 9 independent European-ancestry populations.

METHODS

Populations

We studied 13,886 subjects of European ancestry from 9 independent studies (1 birth cohort, 5 population-based studies, and 3 family studies) that were part of the GABRIEL European consortium on the genetics of asthma.[14] A brief description of these studies with appropriate references is provided in the Methods section and Table E1 in this article's Online Repository at www.jacionline.org. All of these studies had age of asthma onset and imputed genetic data available.

For all studies, ethical approval was obtained from the appropriate institutional ethic committees, and all subjects or children's legal guardians provided written informed consent.

TAO definition

The definition of asthma was based on report of doctor's diagnosis, on standardized questionnaires, or both (see the Methods section in this article's Online Repository). To model TAO, we used age of onset or age at first wheeze for patients with asthma, whereas in subjects who were free of disease on examination, we used age at last examination.

Genotyping

Genotyping, the single nucleotide polymorphism (SNP) imputation process, and quality control (QC) criteria (for subjects and SNPs) for each study are described in Table E1. All data sets were genotyped at Centre National de Génotypage (Evry, France) as part of the European GABRIEL asthma consortium.[14] QC and imputations were performed independently for each study. Genome-wide imputations were conducted with MACH 1.0 software,[15] with reference haplotype panels from HapMap2. SNPs with imputation quality scores (R^2) of 0.5 or greater and minor allele frequencies of 1% or greater were kept for analysis. Then, to further investigate the regions associated with TAO at the genome-wide significant level, we used imputed data from the 1000 Genomes Project and applied the same SNP QC criteria.

Statistical analysis and strategy of analysis

After the study-specific QC, a total of 13,886 subjects from the 9 cohorts were included in the present study. In each data set association between TAO and individual SNPs was investigated under an additive genetic model by using a Cox proportional hazards regression model adjusted for sex and the first 4 principal components to account for population structure. A robust sandwich estimation of variance[16] was used in family data to take into account familial dependencies. Moreover, because of the complex sampling design of the GABRIELA study, survey regression techniques were used for this study to estimate robust SEs (*svy* command in Stata software). Proportional hazard assumptions for the main SNP effect were tested and never rejected. GWASs of TAO were first conducted in each of the 9 data sets separately and then combined through a meta-analysis to increase power and obtain more robust findings. Meta-analyzed hazard ratios and 95% CIs were calculated by using a fixed-effect (inverse variance) model. The Cochran Q statistic was calculated to assess the heterogeneity of the SNP effect across studies. If heterogeneity was evidenced, a random-effect model was fitted. All analyses were performed with Stata software (version 13.1; StataCorp, College Station, Tex). After the meta-analysis, we only kept meta-analysis summary statistics of SNPs included in at least 66% of the studies (>6 of the 9 studies) to reduce the rate of false-positive findings. The meta-analysis results were obtained for a total of 2,387,926 SNPs. We used the classical threshold of a P value of 5×10^{-8} or less to declare a meta-analyzed SNP effect as genome-wide significant.

Conditional analysis to uncover distinct signals at TAO-associated loci

To identify distinct TAO-associated SNPs in each region harboring genome-wide significant signals, we reanalyzed separately these regions in each of the 9 studies. For that purpose, we added the region's top SNP into the primary Cox model as a covariate and tested the effect of each other SNP of that region. Then the results were meta-analyzed by using the same strategy as

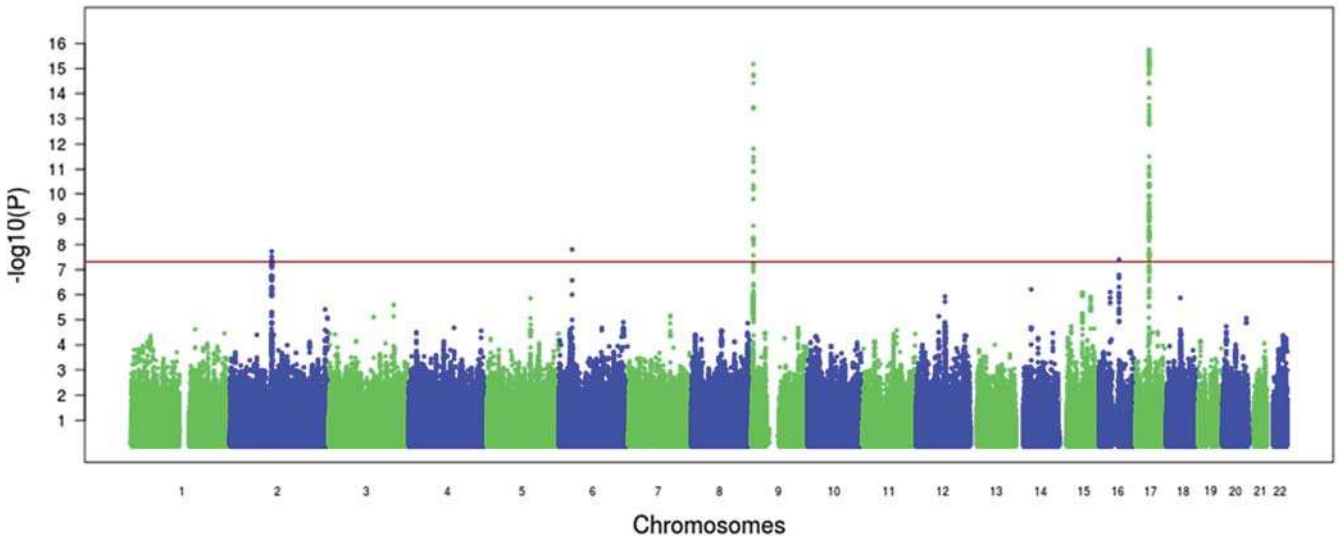


FIG 1. Manhattan plot showing association *P* values of the genome-wide association results for TAO from the meta-analysis. The $-\log_{10}$ of the *P* value for each of 2,387,926 SNPs *(y-axis)* is plotted against the genomic position *(x-axis)*. The *solid red line* indicates the genome-wide significance threshold of a *P* value of 5×10^{-8}.

the primary GWASs. If a secondary signal was detected in a region, a second run of conditional analyses was performed to check for a third distinct signal in that region. The length of the explored regions was based on regional association plots and ranged from 200 to 500 kb depending on recombination hotspots.

Expression quantitative trait locus analysis and functional annotations

We queried whether significant SNPs (or their proxies) associated with TAOs at a *P* value of 5×10^{-8} or less and potentially secondary signals from conditional analysis were expression quantitative trait loci (eQTLs). We used existing eQTL databases in multiple tissues (especially blood and lung) for populations of European ancestry (see the Methods section in this article's Online Repository).[17-23]

Functional annotations of significant SNPs (or their proxies) were obtained by using Encyclopedia of DNA Elements data[24] provided by the HaploReg tool.[25]

Relationship of TAO-associated loci with age of asthma onset

In a first step we investigated in asthmatic patients whether each of the SNPs associated with TAO were also associated with age of asthma onset by using a nonparametric rank test, followed by a nonparametric equality of medians test. In a second step we assessed the cumulative effect of risk alleles of SNPs found to be associated with the age of asthma onset at step 1. For that purpose, we used either the number of risk alleles or the quintiles of a polygenic score distribution. The polygenic risk score is the weighted sum of the number of age of asthma onset–associated alleles, with weight being the log of the adjusted hazard ratio estimated in asthmatic patients only. The associations were tested in 8 studies for which we had access to raw data (all data sets except the Avon Longitudinal Study of Parents and Children) by using a cox proportional hazard model adjusted on sex and principal components.

RESULTS

Description of populations

A total of 13,886 subjects were included in the present study (5,462 asthmatic patients and 8,424 nonasthmatic subjects). Asthmatic patients had a mean age of asthma onset of 12.5 years (range, 0.5-75 years; see Fig E1 in this article's Online Repository at www.jacionline.org) and a mean age of 26.8 years at examination (mean per study ranging from 9.1-51.3 years), and 52.6% were male. Nonasthmatic subjects had a mean age of 32.4 years at examination (mean per study ranging from 8.9-55.8 years), and 49% were male (see Table E1).

Genetic variants associated with TAO

The Manhattan and quantile-quantile plots of the meta-analysis of TAO GWAS results are shown in Fig 1 and Fig E2 in this article's Online Repository at www.jacionline.org, respectively. A total of 155 SNPs were associated with TAO at a genome-wide significance level of a *P* value of less than 5×10^{-8}. These SNPs clustered into 5 distinct chromosomal regions (Table I) that included a new risk locus on 16q12 (near *CYLD*, 1 SNP) and 4 established risk loci for asthma: 2q12 (IL-1 receptor–like 1 *[IL1RL1]–IL18R1*, 7 SNPs), 6p21 (near *HLA-DQA1*, 1 SNP), 9p24 (flanking *IL33*, 25 SNPs), and 17q12-q21 (121 SNPs spanning 389 kb, with the main signal located near zona pellucida binding protein 2 *[ZPBP2]*). The regional association plots for these genome-wide associated loci are shown in Fig 2[26] and Fig E3 in this article's Online Repository at www.jacionline.org, and the forest plots for the top signal in each region are shown in Fig E4 in this article's Online Repository at www.jacionline.org. Three additional loci were associated with TAO at a suggestive significance threshold ($5 \times 10^{-8} < P < 10^{-6}$, Table I): mitogen-activated protein kinase kinase kinase kinase 4 (*MAP4K4*; 2q11-q12), RAR-related orphan receptor A (*RORA*; 15q22), and IL-4 receptor (*IL4R*; 16p12-p11).

To determine whether any of the 5 TAO loci harbored additional association signals, we performed conditional association analysis in each region. For this analysis, a threshold *P* value of 2.1×10^{-5} or less was used to declare significance, corresponding to a Bonferroni threshold for 2382 independent tests. These analyses evidenced 2 secondary signals (Table II and see Fig E5 in this article's Online Repository at www.jacionline.org): (1) rs413382 in the 9p24 region at 73 kb of *IL33* ($P = 9.7 \times 10^{-6}$ after conditioning on the top SNP and

TABLE I. Top SNPs in main loci associated with TAO at genome-wide ($P \leq 5 \times 10^{-8}$) and suggestive significance levels ($5 \times 10^{-8} < P < 10^{-6}$)

Chromosome	Marker	Position*	Nearest gene or genes (kb distance)	Effect/reference alleles†	Effect frequency	Time to asthma onset: n = 13,886		
						Hazard ratio (95% CI)	*P* value‡	P_{Het} value§
Loci with genome-wide significance ($P \leq 5 \times 10^{-8}$)								
2q12	rs10208293	102,966,310	*IL1RL1*‖	G/A	0.73	1.14 (1.08-1.19)	3.1×10^{-8}	.26
6p21	rs9272346	32,604,372	*HLA-DQA1 (0.8)*	A/G	0.59	1.13 (1.08-1.17)	1.6×10^{-8}	.12
9p24	rs928413	6,213,387	*IL33 (2)*	G/A	0.25	1.19 (1.13-1.25)	6.5×10^{-16}	.15
16q12	rs1861760	50,857,693	*CYLD (22)*	A/C	0.04	1.28 (1.17-1.40)	4.2×10^{-8}	.11
17q12-q21	rs9901146	38,043,343	*ZPBP2 (9)* *GSDMB (17)*	G/A	0.51	1.18 (1.13-1.22)	1.9×10^{-16}	.17
Suggestive loci ($5 \times 10^{-8} < P < 10^{-6}$)								
2q11-q12	rs12468899	102,426,140	*MAP4K4*‖	G/A	0.69	1.12 (1.09-1.16)	1.7×10^{-7}	.89
15q22	rs11071559	61,069,988	*RORA*‖	C/T	0.85	1.16 (1.10-1.24)	8.3×10^{-7}	.96
16p12-p11	rs1805013	27,373,980	*IL4R*‖	T/C	0.05	1.22 (1.13-1.32)	8.0×10^{-7}	.37

*Position in base pairs: build 37.3, National Center for Biotechnology Information.
†For the calculation of hazard ratios, effect alleles were designated as risk alleles. *Effect frequency* denotes the frequency of the effect allele.
‡*P* values obtained from the single-SNP Cox model for TAO adjusted for sex and principal components (fixed-effect model when there was no significant evidence of heterogeneity or random-effect model otherwise).
§P_{Het} reflects the *P* value of the Cochran Q statistic across studies.
‖The SNP is located within the reported gene.

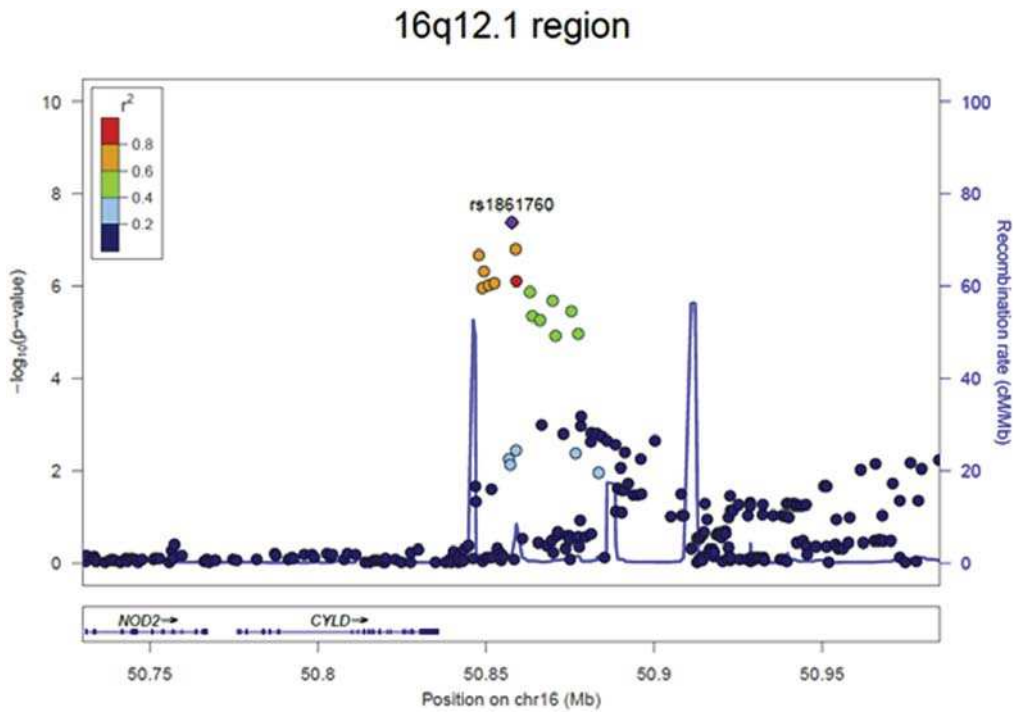


FIG 2. Regional association plot of the 16q12 region using Locuzoom software.[26] SNPs are plotted with their *P* values ($-\log_{10}$ values, *left y-axis*) as a function of genomic position *(x-axis)*. Estimated recombination rates *(right y-axis)* taken from the 1000 Genomes Project (EUR) are plotted to reflect the local linkage disequilibrium structure around the top associated SNP *(purple circle)* and correlated proxies (according to a blue to red scale from and r^2 value of 0-1).

$P = 5.9 \times 10^{-8}$ in the primary meta-analysis) and (2) rs3859192 in the 17q12-q21 region within gasdermin A (*GSDMA*; $P = 4.0 \times 10^{-6}$ after conditioning on the top SNP and $P = 1.5 \times 10^{-13}$ in the primary meta-analysis). In contrast, at the 2q12, 6p21, and 16q12 regions, inclusion of the most significant TAO GWAS SNP as a covariate in association analysis resulted in nearly complete reduction of the association signal in these regions, suggesting that there was no evidence for a second distinct genetic factor in these regions.

To obtain a denser map of the new TAO 16q12 locus, we repeated association analyses using 1000 Genomes Project–imputed SNPs. These analyses strengthened our original finding with additional signals ($3.8 \times 10^{-8} \leq P \leq 2.6 \times 10^{-7}$) located in an intergenic region encompassing the lead SNP rs1861760 (see Table E2 and Fig E6 in this article's Online Repository at www.jacionline.org). These SNPs were in moderate to high linkage disequilibrium with rs1861760 ($0.71 \leq r^2 \leq 0.81$) and thus did not represent independent signals from that top hit. Similar

TABLE II. Secondary signals associated with TAO after stepwise conditional analysis in 9p24 and 17q12-q21 regions

Chromosome	Marker	Nearest gene (kb distance)	Position*	Effect/ reference alleles†	Effect frequency	Single-SNP analysis			Fitted SNP(s)		
						Hazard ratio (95% CI)	*P* value‡	P_{Het}§	Hazard ratio (95% CI)	*P* value‡	P_{Het}§
9p24 region										**rs928413**	
9	rs413382	*IL33 (73)*	6,142,948	A/C	0.80	1.15 (1.08-1.22)	5.9×10^{-8}	.84	1.13 (1.06-1.20)	9.7×10^{-6}	.80
9	**rs928413**	*IL33 (2)*	6,213,387	G/A	0.25	1.19 (1.13-1.25)	6.5×10^{-16}	.15	—	—	—
17q12-q21 region										**rs9901146**	
17	**rs9901146**	*ZPBP2 (9)*	38,043,343	G/A	0.51	1.18 (1.13-1.22)	1.9×10^{-16}	.17	—	—	—
17	rs3859192	*GSDMA*‖	38,128,648	T/C	0.48	1.16 (1.12-1.21)	1.5×10^{-13}	.90	1.11 (1.06-1.15)	4.0×10^{-6}	.74

For these 2 regions, this table contains the top TAO SNP in boldface (rs928413 and rs9901146 respectively) and the most significant SNP in the conditional analysis after fitting the lead SNP in the region.
*Position: Position in base pairs: build 37.3, National Center for Biotechnology Information.
†For calculation of the hazard ratio, effect alleles were designated as risk alleles. *Effect frequency* denotes frequency of the effect allele.
‡*P* values are obtained from the Cox model of TAO adjusted for sex and principal components.
§P_{Het} reflects the *P* value of the Cochran Q statistic across studies.
‖The SNP is located within the reported gene.

TABLE III. Main cis-eQTL results for the top SNPs in genome-wide associated regions from the meta-analysis of TAO

Locus	SNP* (LD with top SNP)	Alleles (reference/ effect)	Gene(s)	Range of *P* values	Tissue	Source‡
2q12	**rs10208293**	G/A	*IL18RAP, IL18R1*	2.5×10^{-13} to 9.8×10^{-198}	Blood, LCLs	Blood eQTLs, eQTL Browser
6p21	**rs9272346**	G/A	*HLA-DQA1/DQA2/DQAS1/ DQB1/DQB2, HLA-DRA/ DRB1/DRB5/DRB6, TAP2*	1.3×10^{-6} to 2.1×10^{-121}	LCLs, lung, blood	eQTL_Chicago,GTEx, blood eQTLs
16q12	**rs1861760**	C/A	*NOD2*	3.6×10^{-11}	Blood	Blood eQTLs
	rs5743266† ($D' = 1$, $r^2 = 0.02$)		*CYLD, NOD2*	5.0×10^{-9} to 3.2×10^{-120}	Blood	Blood eQTLs
	rs7205760† ($D' = 1$, $r^2 = 0.005$)		*CYLD, NOD2*	2.8×10^{-6} to 4.0×10^{-15}	Lung, blood	Lung eQTLs, blood eQTLs
17q12-q21	**rs9901146**	A/G	*GSDMB, ORMDL3*	3.8×10^{-6} to 9.8×10^{-198}	Blood, LCLs	Blood eQTLs, GTEx, eQTL Browser, eQTL_Chicago
	rs3859192	C/T	*GSDMA, GSDMB, ORMDL3*	1.1×10^{-7} to 2.5×10^{-12}	Lung, LCLs	GTEx, eQTL Browser

We focused on eQTLs measured in blood, lymphoblastoid cell lines, and lung tissue.
LCL, Lymphoblastoid cell line; *LD*, linkage disequilibrium.
*Top genome-wide significant SNPs in TAO meta-analysis and secondary associations identified by conditional analyses are indicated in boldface.
†Haplotype reconstruction was done with Haploview; the effect allele of the top SNP (A-rs1861760) is always transmitted with the effect allele of its proxy (G-rs5743266 and G-rs7205760).
‡Interrogated databases: eQTL Browser (LCLs of British subjects with asthma or eczema),[18] Blood eQTL Browser (nontransformed peripheral blood samples),[20] Lung eQTLs (lung tissue),[17] GTEx eQTL Browser v4 (several tissues, among which were blood and lung tissue),[23] and eQTL Chicago Browser (LCLs).[19,21,22]

analyses conducted in the 4 other TAO-associated regions also supported our original findings and did not find evidence for any additional independent signal in these regions.

Overall, the 7 distinct SNPs (5 top SNPs and 2 secondary SNPs) associated with TAO showed low heterogeneity between studies ($P > .11$) and together explained 6.0% of the variance in TAO.

Functional annotations and effect on gene expression

To provide some insights into the potential molecular mechanisms underlying the TAO-associated variants, we queried whether the 5 top SNPs and 2 secondary signals (and their proxies) were (1) tagging potentially deleterious SNPs, (2) located in regulatory elements, and (3) reported to influence the expression of 1 or more of the nearby genes (eQTLs at $P < 5 \times 10^{-5}$). We focused on the new TAO risk locus at the 16q12 region. Functional annotations for the remaining 6 loci are presented in the Results section in this article's Online Repository at www.jacionline.org, and eQTL data are presented in Table III[17-23] and Table E3 in this article's Online Repository at www.jacionline.org.

The 16q12 TAO-associated variants are located in an intergenic region delimited by 2 recombination hotspots on each side near *CYLD* (22 kb downstream). rs1861760 maps to the FOXJ1 and SOX binding sites. This SNP and/or its proxies correlate with the expression of *CYLD* in both blood and human lung tissues and the expression of nucleotide-binding oligomerization domain 2 *(NOD2)* in blood (Table III and see Table E3).[17,20]

Relationship between TAO-associated variants and age of asthma onset

To investigate whether TAO-associated SNPs influence age of asthma onset, in asthmatic patients we compared the distribution of age of asthma onset according to the number of risk alleles at

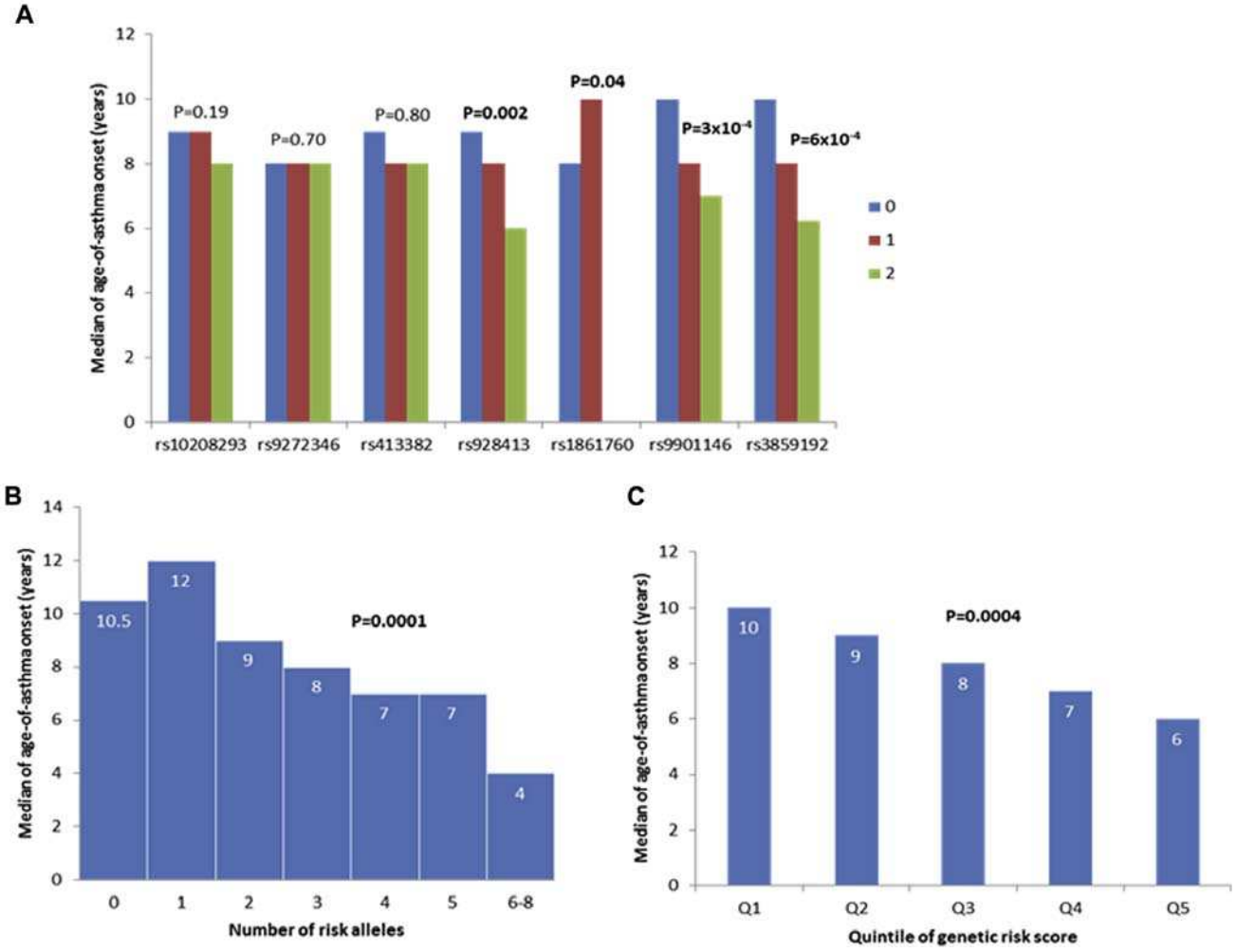


FIG 3. Relationship between TAO-associated SNPs and age of asthma onset. **A,** Association between age of asthma onset and genotypes at individual loci. **B,** Median of age of asthma onset as a function of the total number of risk alleles of SNPs found associated with the age-of-asthma onset and carried by asthmatic subjects. **C,** Median of age of asthma onset by quintile of genetic risk score.

each of the 7 main and secondary TAO-associated SNPs (Fig 3). Asthmatic patients carrying 1 or 2 copies of the risk allele at 17q12-q21 SNPs (rs9901146 and rs3859192) or at 9p24 rs928413 had a younger age of asthma onset than noncarriers (median of 6-8 vs 10 years [$P \leq 6 \times 10^{-4}$] and 6-8 vs 9 years [$P = .002$], respectively), whereas those having at least 1 copy of the rs1861760 risk allele at 16q12 had a later age of asthma onset than noncarriers (median of 10 vs 8 years, $P = .04$). No significant difference was found for the other 3 SNPs. We evidenced that an increased number of risk alleles at these 4 SNPs was associated with a younger age of asthma onset (median of 12 years for carrying 1 risk allele to 4 years for carrying 6-8 risk alleles, $P = 10^{-4}$). Finally, we detected a strong association between age of asthma onset and the polygenic risk score (from a median of 10 years in the first quintile to 6 years in the last quintile, $P = 4 \times 10^{-4}$).

Comparison of TAO GWAS results with previous asthma GWASs

To investigate the effect of taking into account the age of asthma onset in disease modeling through survival analysis, we explored whether the top TAO SNPs were associated with asthma modeled as a binary trait in the 9 cohorts included in the present study (see Table E4). We also investigated the GABRIEL top SNPs in our TAO meta-analysis (see Table E4).[14] We observed a strong decrease in heterogeneity of the SNP effect across studies in our TAO analysis ($P_{Het} \geq .11$) compared with the asthma binary trait analyzed in the same data sets ($P_{Het} \geq .004$), as well as in all GABRIEL data sets ($P_{Het} \geq .0009$), especially in the 9p24 and 17q12-q21 regions. The association signals were always more significant in TAO analysis compared with the binary trait analysis in the same data sets. This increase in significance level was very high: 100-fold for 2q12 and 16q12 and 10^4- to 10^6-fold for 9p24 and 17q12-q21. In fact, the asthma binary trait analysis only detected 2 loci (*HLA* and *GSDMA*) at the genome-wide significance level 7 TAO-associated loci. Conversely, at the genome-wide significance level, the present TAO analysis identified 4 of the 6 main published GABRIEL regions[14] and events at higher significance for the 9p24 and 17q12-q21 regions (100- to 10^4-fold) compared with GABRIEL significance levels. The 2 remaining GABRIEL loci not detected by our TAO analysis were those with weaker effects (odds ratio, 1.12 for rs744910 in 15q22 and rs2284033 in 22q13) in the GABRIEL meta-analysis.[14]

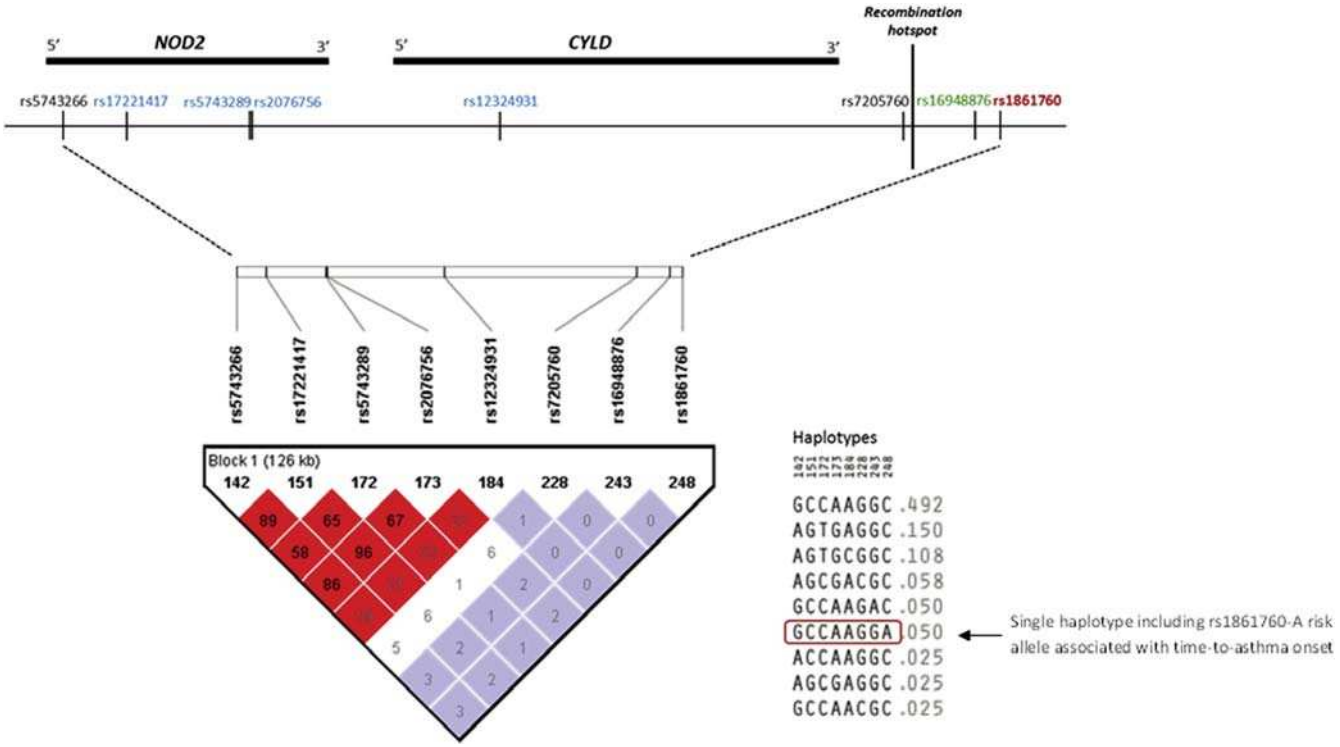


FIG 4. Map of the 16q12 region (build 37.3 position: 50,723,355 to 50,860,722) and haplotype reconstruction for SNPs found to be associated with inflammatory bowel disease (among which was Crohn disease, *blue*), leprosy *(green)*, or asthma *(red)* or with expression of *CYLD* or *NOD2 (black)*. The linkage disequilibrium plot was obtained by using the Hapmap2 CEU reference sample from Haploview[37] (values and colors reflect r^2 and D′ values, respectively). The 16q12 top SNP (rs1861760) associated with TAO is indicated in boldface.

Finally, we evaluated whether previously reported susceptibility loci for asthma[27] were associated with TAO in our meta-analysis (see Table E5 in this article's Online Repository at www.jacionline.org). Among the 21 loci detected in European populations, 12 were replicated at 5% in our TAO meta-analysis, with the same direction of effects. Among the 9 nonreplicated signals, 3 SNPs (or some proxies) were not available in our data, and the remaining 6 loci had been reported for specific phenotypes: asthma exacerbation, age of asthma onset *per se* in asthmatic children only (quantitative trait), or childhood asthma (binary trait).[11,28,29]

DISCUSSION

By taking into account age of asthma onset in an asthma association analysis, in this large meta-analysis including both asthmatic and nonasthmatic subjects (adults and children), we identified a new susceptibility locus at 16q12 associated with TAO and confirmed the involvement of 6 other distinct loci belonging to 4 regions in asthma pathogenesis (2q12, 6p21, 9p24, and 17q12-q21). Genetic variants at 9p24 and 17q12-q21 were strongly associated with an earlier onset of childhood asthma, whereas the 16q12 lead SNP was associated with a risk of later-onset asthma.

The most significant 16q12 genetic variant (rs1861760) is located near *CYLD* and *NOD2* and also maps to a binding site of FOXJ1, a transcription factor associated with allergic rhinitis.[30] Genetic variants located in a 130-kb region around rs1861760 were reported to be associated with immune-related diseases: inflammatory bowel diseases (Crohn disease) and leprosy.[31-36] Interestingly, haplotype reconstruction (Fig 4[37]) showed that the TAO rs1861760-A risk allele was always associated with SNP alleles that conferred a decreased risk of Crohn disease (rs17221417-C, rs5743289-C, and rs2076756-A located in *NOD2* and rs12324931-A located in *CYLD)*[31-33,36,38] and of leprosy (rs16948876-G located in intergenic region at 2 kb from rs1861760).[34] Indeed, GWASs revealed common genetic susceptibility loci for asthma and other immune-related disorders, suggesting shared molecular pathways involved in their cause; however, opposite alleles appear to be at risk.[39] Interestingly, an opposite effect of the rs1861760-A allele is also observed at the gene expression level. Thus the TAO risk allele at rs1861760 correlated with both expression of *CYLD* and *NOD2* in blood, although with an opposite effect.[20] However, this TAO risk allele was only associated with increased *CYLD* expression in lung tissue.[17] *CYLD* encodes a deubiquitinating enzyme that regulates diverse physiologic processes, including immune response and inflammation.[40] *CYLD* mainly acts as a negative regulator of nuclear factor-κB *(NFkB1)* to protect the host from an overreactive inflammatory response.[40] Conversely, *NOD2*, which plays an important role in the innate immune response to intracellular bacterial LPSs, activates the *NFkB1* pathway.[41] NFkB1 is a pleiotropic transcription factor that acts as a key regulator of immune and inflammatory genes, and activation of the NFkB1 pathway has been implicated in airway inflammation and asthma.[42,43] Moreover, the FOXJ1 transcription factor that binds to the genomic region encompassing the 16q12 TAO-associated SNP (rs1861760) was described to inhibit *NFkB1* activity.[44] Recently, *CYLD* has been shown to regulate lung fibrosis in mice by inhibiting TGF-β signaling through a decrease of SMAD3 protein stability.[45] Of interest, *SMAD3* has

been reported to be associated with asthma in previous GWASs.[14]

Defining the phenotype is an important consideration because phenotypic heterogeneity can reduce the power of GWASs.[46] In the present analyses we studied the variability of TAO in both asthmatic and nonasthmatic subjects based on survival analysis methods. The information used for such analysis was the age of onset in asthmatic patients and the age at last examination or death in nonasthmatic subjects. In such a model unaffected subjects represent censored observations because they are still at risk for disease, being perhaps too young to exhibit the trait. This approach, which allowed combining the age of asthma onset and disease status (affected/unaffected), led to a decrease in genetic heterogeneity across studies and an increase in the power to detect association signals (on a 10^6-fold increase compared with the disease status–only analysis). More specifically, increased evidence of association was observed in regions in which age of asthma onset explained at least in part the genetic heterogeneity, such as the 17q12-q21 locus, for which a restricted SNP effect to a particular group of age of onset (early childhood–onset asthma) was demonstrated.[13] Moreover, this analysis led to the identification of a new locus at 16q12 near *CYLD* and of an additional signal in the 9p24 region. These results support the hypothesis that a better consideration of the phenotypic heterogeneity of asthma might help disentangle the genetic heterogeneity of asthma.

Our study included both children and adults with asthma. Age of disease onset might be subject to recall bias, especially among subjects who are furthest from the time of first symptoms (eg, adults with asthma in childhood), because it is often defined in a retrospective manner. However, high accuracy of the self-reported year of asthma onset by adult subjects has been shown by 2 independent studies, including the European Community Respiratory Health Survey, which was part of the present study.[47,48] Erroneous recall of age of asthma onset is unlikely to have significantly affected the results because we observed little genetic heterogeneity across studies (eg, childhood-onset asthma reported by either adults or children).

It was suggested that some genetic variants can influence asthma in an age-specific manner. Among TAO-associated SNPs, we confirmed the association of 17q12-q21 SNPs with an early age of asthma onset[13,14] and evidenced for the first time that the top 9p24 genetic variant near *IL33* was also associated with early childhood–onset asthma (median age of onset of 6-8 years in risk allele carriers). Indeed, in the GABRIEL meta-analysis 9p24 SNPs were more strongly associated with early-onset (before age 16 years) than late-onset (after age 16 years) asthma, but this difference was not significant.[14] Conversely, genetic variants at the new susceptibility locus, 16q12, conferred a risk of later-onset asthma (median age of onset of 10 years in risk allele carriers). Moreover, we evidenced that a high burden of disease risk alleles at these loci is associated with earlier age of asthma onset (4 vs 9-12 years). This difference in asthma onset might reflect the difference in patterns of onset of disease.[49] Indeed, we evidenced in the GABRIELA study that subjects with persistent early wheezing carried more risk alleles than subjects with transient early wheezing, and we confirmed the previous association between persistent early wheezing and 9p24 and 17q12-q21 loci (data not shown). The 17q12-q21 genetic variants were reported to be associated with the persistent childhood wheeze phenotype, whereas 9p24 variants were mostly associated with intermediate-onset wheeze but also with persistent early wheeze.[50,51] Moreover, 17q12-q21 SNPs were associated with fraction of exhaled nitric oxide levels in children but not adults, childhood severe asthma, and allergic rhinitis, and 9p24 SNPs were associated with childhood severe asthma, asthma plus rhinitis, atopic asthma, allergy, and eosinophil counts.[51-57]

In summary, we identified 5 regions harboring 7 distinct signals associated with TAO, including the 16q12 region, which is reported for the first time. Several lines of evidence suggest that *CYLD* and *NOD2*, which are located in that region, are strong candidate genes for asthma. This study demonstrates that incorporating the variability of age of asthma onset in disease modeling is a useful strategy to uncover new disease genes.

Epidemiological Study on Genetics and Environment of Asthma (EGEA): We thank all those who participated in the setting of the study and on the various aspects of the examinations involved: interviewers; technicians for lung function testing and skin prick tests, blood sampling, and IgE determinations; coders; those involved in QC, data, and sample management; and all those who supervised the study in all EGEA centers. We thank the 3 CIC-Inserm facilities of Necker, Grenoble, and Marseille who supported the EGEA study and in which subjects were examined. We also thank the biobanks in Lille (CIC-Inserm) and at Annemasse (Etablissement français du sang), where EGEA biological samples are stored. Finally, we thank the EGEA cooperative group members as follows.

Coordination: V. Siroux (epidemiology, PI since 2013); F. Demenais (genetics); I. Pin (clinical aspects); R. Nadif (biology). F. Kauffmann (PI 1992-2012); *Respiratory epidemiology*: Inserm U 700, Paris—M. Korobaeff (Egea1), F. Neukirch (Egea1); Inserm U 707, Paris—I. Annesi-Maesano (Egea1-2); Inserm CESP/U 1018, Villejuif—F. Kauffmann, N. Le Moual, R. Nadif, M. P. Oryszczyn (Egea1-2), and R. Varraso; Inserm U 823, Grenoble—V. Siroux. *Genetics*: Inserm U 393, Paris—J. Feingold; Inserm U 946, Paris—E. Bouzigon, .F Demenais, M. H. Dizier; Centre National de Génotypage, Evry—I. Gut (now CNAG, Barcelone, Spain), M. Lathrop (now University of McGill, Montreal, Canada). *Clinical centers*: Grenoble—I. Pin, C. Pison; Lyon—D. Ecochard (Egea1), F. Gormand, Y. Pacheco; Marseille—D. Charpin (Egea1), D. Vervloet (Egea1-2); Montpellier—J. Bousquet; Paris Cochin—A. Lockhart (Egea1), R. Matran (now in Lille); Paris Necker—E. Paty (Egea1-2), P. Scheinmann (Egea1-2); Paris-Trousseau—A. Grimfeld (Egea1-2), J. Just. *Data and quality management*: Inserm ex-U155 (Egea1)—J. Hochez; Inserm CESP/U 1018, Villejuif—N. Le Moual, Inserm ex-U780—C. Ravault (Egea1-2); Inserm ex-U794—N. Chateigner; Grenoble—J. Ferran (Egea1-2).

Saguenay-Lac-Saint-Jean Familial Collection (SLSJ): We thank all participants included in the SLSJ asthma familial collection. Catherine Laprise built, coordinates, and manages the SLSJ study. Drs Paul Bégin and Charles Morin confirmed the respiratory diagnosis for the adults and children, respectively. We also thank the laboratory technicians (Nadia Mior and Denise Morin), research professional (Anne-Marie Madore), and nurses (from the ECOGENE-21 clinical research group). Catherine Laprise is the Canada Research Chair in Environment and Genetics of Respiratory Disorders and Allergy, Director of the Asthma Strategic Group of the Respiratory Health Network (RHN) of Fonds de la recherche en santé du Québec (FRSQ), and researcher of the AllerGen NCE.

Avon Longitudinal Study of Parents and Children (ALSPAC): We thank all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

UFA: We thank the staff members of the Departments of Pediatrics and Propaedeutics of internal diseases of Bashkir Medical State University.

Swiss study on Air Pollution and Lung and Heart Disease In Adults (SAPALDIA): The study could not have been done without the help of the study participants, technical and administrative support, and medical teams and field workers at the local study sites. *Study directorate*:

N. M. Probst-Hensch (PI; e/g); T. Rochat (p), C. Schindler (s), N. Künzli (e/exp), J. M. Gaspoz (c). *Scientific team*: J. C. Barthélémy (c), W. Berger (g), R. Bettschart (p), A. Bircher (a), C. Brombach (n), P. O. Bridevaux (p), L. Burdet (p), Felber Dietrich (e), M. Frey (p), U. Frey (pd), M. W. Gerbase (p), D. Gold (e), E. de Groot (c), W. Karrer (p), F. Kronenberg (g), B. Martin (pa), A. Mehta (e), D. Miedinger (o), M. Pons (p), F. Roche (c), T. Rothe (p), P. Schmid-Grendelmeyer (a), D. Stolz (p), A. Schmidt-Trucksäss (pa), J. Schwartz (e), A. Turk (p), A. von Eckardstein (cc), E. Zemp Stutz (e). *Scientific team at coordinating centers*: M. Adam (e), I. Aguilera (exp), S. Brunner (s), D. Carballo (c), S. Caviezel (pa), I. Curjuric (e), A. Di Pascale (s), J. Dratva (e), R. Ducret (s), E. Dupuis Lozeron (s), M. Eeftens (exp), I. Eze (e), E. Fischer (g), M. Foraster (e), M. Germond (s), L. Grize (s), S. Hansen (e), A. Hensel (s), M. Imboden (g), A. Ineichen (exp), A. Jeong (g), D. Keidel (s), A. Kumar (g), N. Maire (s), A. Mehta (e), R. Meier (exp), E. Schaffner (s), T. Schikowski (e), M. Tsai (exp). *a*, Allergology; *c*, cardiology; *cc*, clinical chemistry; *e*, epidemiology; *exp*, exposure; *g*, genetic and molecular biology; *m*, meteorology; *n*, nutrition; *o*, occupational health; *p*, pneumology; *pa*, physical activity; *pd*, pediatrics; *s*, statistics.

Key messages

- **16q12 genetic variants are associated with TAO and correlate with *CYLD* and *NOD2* expression.**
- **Genetic variants at 9p24 (upstream of *IL33*) and 17q12-q21 (nearby *ZPBP2* and within *GSDMA*) are associated with an earlier asthma onset, whereas variants at 16q12 are associated with later asthma onset.**
- **Taking into account the variability of age of asthma onset in disease modeling can increase the power of identifying new genes involved in asthma physiopathology.**

REFERENCES

1. Martinez FD, Vercelli D. Asthma. Lancet 2013;382:1360-72.
2. Radhakrishnan DK, Dell SD, Guttmann A, Shariff SZ, Liu K, To T. Trends in the age of diagnosis of childhood asthma. J Allergy Clin Immunol 2014;134:1057-62.e5.
3. Wenzel SE. Asthma phenotypes: the evolution from clinical to molecular approaches. Nat Med 2012;18:716-25.
4. Siroux V, Basagana X, Boudier A, Pin I, Garcia-Aymerich J, Vesin A, et al. Identifying adult asthma phenotypes using a clustering approach. Eur Respir J 2011;38:310-7.
5. Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. Am J Respir Crit Care Med 2010;181:315-23.
6. Szefler SJ, Chmiel JF, Fitzpatrick AM, Giacoia G, Green TP, Jackson DJ, et al. Asthma across the ages: knowledge gaps in childhood asthma. J Allergy Clin Immunol 2014;133:3-14.
7. Ober C, Yao TC. The genetics of asthma and allergic disease: a 21st century perspective. Immunol Rev 2011;242:10-30.
8. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 2014;42:D1001-6.
9. Siroux V, Gonzalez JR, Bouzigon E, Curjuric I, Boudier A, Imboden M, et al. Genetic heterogeneity of asthma phenotypes identified by a clustering approach. Eur Respir J 2014;43:439-52.
10. Bouzigon E, Ulgen A, Dizier MH, Siroux V, Lathrop M, Kauffmann F, et al. Evidence for a pleiotropic QTL on chromosome 5q13 influencing both time to asthma onset and asthma score in French EGEA families. Hum Genet 2007;121:711-9.
11. Forno E, Lasky-Su J, Himes B, Howrylak J, Ramsey C, Brehm J, et al. Genome-wide association study of the age of onset of childhood asthma. J Allergy Clin Immunol 2012;130:83-90.e4.
12. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 2007;448:470-3.
13. Bouzigon E, Corda E, Aschard H, Dizier MH, Boland A, Bousquet J, et al. Effect of 17q21 variants and smoking exposure in early-onset asthma. N Engl J Med 2008;359:1985-94.
14. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, et al. A large-scale, consortium-based genomewide association study of asthma. N Engl J Med 2010;363:1211-21.
15. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet 2009;10:387-406.
16. Williams RL. A note on robust variance estimation for cluster-correlated data. Biometrics 2000;56:645-6.
17. Hao K, Bosse Y, Nickle DC, Pare PD, Postma DS, Laviolette M, et al. Lung eQTLs to help reveal the molecular underpinnings of asthma. PLoS Genet 2012;8:e1003029.
18. Liang L, Morar N, Dixon AL, Lathrop GM, Abecasis GR, Moffatt MF, et al. A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. Genome Res 2013;23:716-26.
19. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 2010;464:773-7.
20. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet 2013;45:1238-43.
21. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, et al. Population genomics of human gene expression. Nat Genet 2007;39:1217-24.
22. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, et al. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 2008;4:e1000214.
23. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45:580-5.
24. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57-74.
25. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 2012;40:D930-4.
26. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 2010;26:2336-7.
27. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 2009;106:9362-7.
28. Himes BE, Hunninghake GM, Baurley JW, Rafaels NM, Sleiman P, Strachan DP, et al. Genome-wide association analysis identifies PDE4D as an asthma-susceptibility gene. Am J Hum Genet 2009;84:581-93.
29. McGeachie MJ, Wu AC, Tse SM, Clemmer GL, Sordillo J, Himes BE, et al. CTNNA3 and SEMA3D: Promising loci for asthma exacerbation identified through multiple genome-wide association studies. J Allergy Clin Immunol 2015;136:1503-10.
30. Li CS, Chae SC, Lee JH, Zhang Q, Chung HT. Identification of single nucleotide polymorphisms in FOXJ1 and their association with allergic rhinitis. J Hum Genet 2006;51:292-7.
31. Kenny EE, Pe'er I, Karban A, Ozelius L, Mitchell AA, Ng SM, et al. A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. PLoS Genet 2012;8:e1002559.
32. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661-78.
33. Cleynen I, Vazeille E, Artieda M, Verspaget HW, Szczypiorska M, Bringer MA, et al. Genetic and microbial factors modulating the ubiquitin proteasome system in inflammatory bowel disease. Gut 2014;63:1265-74.
34. Zhang F, Liu H, Chen S, Low H, Sun L, Cui Y, et al. Identification of two new loci at IL23R and RAB32 that influence susceptibility to leprosy. Nat Genet 2011;43:1247-51.
35. Liu H, Irwanto A, Fu X, Yu G, Yu Y, Sun Y, et al. Discovery of six new susceptibility loci and analysis of pleiotropic effects in leprosy. Nat Genet 2015;47:267-71.
36. Kugathasan S, Baldassano RN, Bradfield JP, Sleiman PM, Imielinski M, Guthery SL, et al. Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. Nat Genet 2008;40:1211-5.
37. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263-5.
38. Elding H, Lau W, Swallow DM, Maniatis N. Dissecting the genetics of complex inheritance: linkage disequilibrium mapping provides insight into Crohn disease. Am J Hum Genet 2011;89:798-805.

39. Li X, Ampleford EJ, Howard TD, Moore WC, Torgerson DG, Li H, et al. Genome-wide association studies of asthma indicate opposite immunopathogenesis direction from autoimmune diseases. J Allergy Clin Immunol 2012;130:861-8.e7.
40. Sun SC. CYLD: a tumor suppressor deubiquitinase regulating NF-kappaB activation and diverse biological processes. Cell Death Differ 2010;17:25-34.
41. Ogura Y, Inohara N, Benito A, Chen FF, Yamaoka S, Nunez G. Nod2, a Nod1/Apaf-1 family member that is restricted to monocytes and activates NF-kappaB. J Biol Chem 2001;276:4812-8.
42. Poynter ME, Cloots R, van Woerkom T, Butnor KJ, Vacek P, Taatjes DJ, et al. NF-kappa B activation in airways modulates allergic inflammation but not hyperresponsiveness. J Immunol 2004;173:7003-9.
43. Kurakula K, Vos M, Logiantara A, Roelofs JJ, Nieuwenhuis MA, Koppelman GH, et al. Nuclear receptor Nur77 attenuates airway inflammation in mice by suppressing NF-kappaB activity in lung epithelial cells. J Immunol 2015;195:1388-98.
44. Lin L, Spoor MS, Gerth AJ, Brody SL, Peng SL. Modulation of Th1 activation and inflammation by the NF-kappaB repressor Foxj1. Science 2004;303:1017-20.
45. Lim JH, Jono H, Komatsu K, Woo CH, Lee J, Miyata M, et al. CYLD negatively regulates transforming growth factor-beta-signalling via deubiquitinating Akt. Nat Commun 2012;3:771.
46. Ioannidis JP, Thomas G, Daly MJ. Validating, augmenting and refining genome-wide association signals. Nat Rev Genet 2009;10:318-29.
47. Pattaro C, Locatelli F, Sunyer J, de Marco R. Using the age at onset may increase the reliability of longitudinal asthma assessment. J Clin Epidemiol 2007;60:704-11.
48. Toren K, Palmqvist M, Lowhagen O, Balder B, Tunsater A. Self-reported asthma was biased in relation to disease severity while reported year of asthma onset was accurate. J Clin Epidemiol 2006;59:90-3.
49. Dijk FN, de Jongste JC, Postma DS, Koppelman GH. Genetics of onset of asthma. Curr Opin Allergy Clin Immunol 2013;13:193-202.
50. Granell R, Henderson AJ, Timpson N, St Pourcain B, Kemp JP, Ring SM, et al. Examination of the relationship between variation at 17q21 and childhood wheeze phenotypes. J Allergy Clin Immunol 2013;131:685-94.
51. Savenije OE, Mahachie John JM, Granell R, Kerkhof M, Dijk FN, de Jongste JC, et al. Association of IL33-IL-1 receptor-like 1 (IL1RL1) pathway polymorphisms with wheezing phenotypes and asthma in childhood. J Allergy Clin Immunol 2014;134:170-7.
52. van der Valk RJ, Duijts L, Timpson NJ, Salam MT, Standl M, Curtin JA, et al. Fraction of exhaled nitric oxide values in childhood are associated with 17q11.2-q12 and 17q12-q21 variants. J Allergy Clin Immunol 2014;134:46-55.
53. Bonnelykke K, Sleiman P, Nielsen K, Kreiner-Moller E, Mercader JM, Belgrave D, et al. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. Nat Genet 2014;46:51-5.
54. Fuertes E, Soderhall C, Acevedo N, Becker A, Brauer M, Chan-Yeung M, et al. Associations between the 17q21 region and allergic rhinitis in 5 birth cohorts. J Allergy Clin Immunol 2015;135:573-6.
55. Ferreira MA, Matheson MC, Tang CS, Granell R, Ang W, Hui J, et al. Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype. J Allergy Clin Immunol 2014;133:1564-71.
56. Hinds DA, McMahon G, Kiefer AK, Do CB, Eriksson N, Evans DM, et al. A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. Nat Genet 2013;45:907-11.
57. Gudbjartsson DF, Bjornsdottir US, Halapi E, Helgadottir A, Sulem P, Jonsdottir GM, et al. Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. Nat Genet 2009;41:342-7.

Online Repository

Identification of a new locus at 16q12 associated with time-to-asthma onset

C Sarnowski *et al*.

METHODS

Study populations

The present project includes nine independent studies (Table E1) among which five population-based studies: the European Community Respiratory Health Survey (ECRHS), the Swiss study on Air Pollution and Lung and Heart Disease In Adults (SAPALDIA), the Busselton Health Study (BUSSELTON), the GABRIEL Advanced Surveys (GABRIELA) and the UFA study (UFA); three familial studies: the Epidemiological study on the Genetics and Environment of Asthma (EGEA); the Saguenay-Lac-Saint-Jean Familial Collection (SLSJ) and the TOMSK study (TOMSK); and one birth cohort: the Avon Longitudinal Study of Parents and Children (ALSPAC). All of these studies were part of the GABRIEL European consortium on asthma,[1] and had the information on age of asthma onset, age at last examination and imputed genetic data available. A total of 13,886 subjects from European ancestry were included in the time-to-asthma onset GWAS meta-analysis (5,462 asthmatics and 8,424 non-asthmatics).

Study populations and Phenotype definition

ALSPAC

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a population-based birth cohort initially comprising of 14,541 mothers and their children recruited in the former County of Avon, UK between 1991 and 1992.[2]

Asthmatics were defined by a positive response to the question: “Did you child had asthma in past 12 months?” at 81, 91, 103, 128, 157 or 166 months. Non-asthmatics were those who answered no at all surveys. In asthmatics, age of onset was defined by the first time they declared wheeze or wheezing and whistling. Wheeze was defined by a positive response to the question: “Has your child had wheezing, breathlessness or episodes of stopping breathing in past 12 months or since he was (age at last Q)?”. Wheezing and whistling were defined by a positive response to the question: “Has your child had any periods when there was wheezing with whistling on his chest when he breathed in past 12 months or since he was (age at last Q)?”. In non-asthmatics, we considered age at the last examination without any missing visit at the preceding surveys. Thus, for non-asthmatics with negative complete reports, we considered age at last examination and for non-asthmatics with negative incomplete and/or discontinued reports, we considered age until the last visit before the first missing visit. We did not include in the present analysis non-asthmatics who experienced wheeze or wheezing and whistling before 6yrs of age.

ECRHS

The ECRHS study is a European population-based study of young adults with a 8-year follow-up (ECRHS I: 1991-1993, ECRHS II: 1999-2002 and ECRHS III: 2010-ongoing).[3,4] The time-to-asthma onset GWAS is based on the two first survey data. Participants included in the meta-analysis were derived from the nested asthma case/control sample subjected to genome-wide genotyping in the context of the GABRIEL asthma GWAS.

Asthma cases were identified by participants from the random or enriched sample who said yes to the question ‘Have you ever had asthma?’ at either ECRHS I Stage 2 or at ECRHS II. Controls were a random sample (of the random sample) who answered ‘no’ to the same question in both surveys. For individuals who developed asthma, information on asthma age at onset

was obtained from age at first asthma attack at ECRHS I or II. For individuals who were free of disease upon examination, we considered age at last examination.

EGEA

Briefly, the EGEA study combines a case-control and a family-based study of asthma cases (N=2,120 subjects) with three surveys over 20 years (EGEA1: 1991-1995, EGEA2: 2003-2007 and EGEA3: 2011-2013). The whole study population included 388 asthmatic probands recruited in chest clinics and their 1,317 family members (probands' parents and/or siblings) plus 415 population-based controls.[5]

Asthma was defined in probands by a positive answer to the following four items "Have you ever had attacks of breathlessness at rest with wheezing?", "Have you ever had asthma attacks?", "Was this diagnosis confirmed by a physician?", and "Have you had an asthma attack in the last 12 months?" or on a positive self-report to two of the before mentioned items plus a medical record of asthma. Individuals were considered free of disease if they answered no at all items. Relatives of probands were defined as asthmatics if they answered positively at either survey to "Have you ever had attacks of breathlessness at rest with wheezing?" or "Have you ever had asthma attacks?" at EGEA1, EGEA2 or EGEA3.

For individuals who developed asthma, information on asthma age at onset was obtained from adult asthmatics or parents of asthmatic children who answered to the following question: "How old were you when you had your first asthma attack?" or "How old was your child when he (or she) had his (her) first asthma attack?". For individuals who were free of disease upon last examination, we considered age at last examination.

GABRIEL Advanced Surveys

GABRIELA are cross-sectional population-based surveys conducted in rural areas of Austria, Germany, and Switzerland during fall/winter 2006 and spring/summer 2007.[6]

A case was defined as a parental report of asthma diagnosed by a doctor at least once or asthmatic bronchitis diagnosed at least twice during lifetime. The reference category for asthma was no reported diagnosis of asthma ever and a diagnosis of asthmatic bronchitis no more than once. The original question on age of onset was: "How old was your child when the first symptoms of wheezing or whistling in the chest began? At the age of … years. If during the first year: At the age of … months." The variable is coded for years. Months were transferred to years. In non-asthmatics, we considered age at examination. To define atopy, we used an atopic sensitization to mite, cat, or birch upper or equal to 0.7 kU/L.

SAPALDIA

The SAPALDIA study is a cohort study with integrated biobank in the Swiss population initiated in 1991 (SAPALDIA 1: n=9,651; age 18-60 at baseline) with two follow-up assessments in 2001-2003 (SAPALDIA 2: n=8,047) and in 2010-2011 (SAPALDIA 3: n=6,200).[7,8] The time-to-asthma onset GWAS is based on the first two survey data. Participants included in the meta-analysis were derived from the nested asthma case/control sample subjected to genome-wide genotyping in the context of the GABRIEL asthma GWAS.

Asthma status was defined by an affirmative answer to the question "Have you ever had asthma" at baseline and/or follow-up interview. Controls were defined by a negative answer to the same question. Age of onset was self-reported by study participants. For individuals who were free of disease upon examination, we considered age at last examination.

BUSSELTON

The Busselton Health Study is a population-based, nested, case-control panel of 1,549 individuals of European Caucasian descent from Australia[9,10] with seven cross-sectional respiratory health surveys of adults conducted between 1966 and 2005-2007 and five cross-sectional respiratory health surveys of all school children conducted between 1967 and 1983. Asthma cases were defined as those who reported doctor-diagnosed asthma at any survey that they attended from 1966 to 1994 (answer 'Yes' to 'Has your doctor ever told you that you had asthma?'). Controls are those who have consistently answered 'No' to 'Has your doctor ever told you that you had asthma?' at all previous surveys that they have attended from 1966 to 1994. Age of onset was obtained from answer to the following question "How old were you when you first developed symptoms of asthma?". For individuals who were free of disease upon examination, we considered age at last examination.

SLSJ

The Saguenay-Lac-Saint-Jean and Quebec City Familial Asthma Collection (SLSJ) consisting of a French-Canadian founder population panel of 253 multigenerational families from Saguenay-Lac-Saint-Jean region, ascertained through two asthmatic probands between 1997 and 2002.[11]

Probands were included in the study if they fulfill at least two of the following criteria: 1) a minimum of three clinic visits for acute asthma within one year; 2) two or more asthma-related hospital admissions within one year; or 3) steroid dependency, as defined by either six months of oral, or one year of inhaled corticosteroid use. Families were included in the study if at least one parent was available for phenotypic assessment, at least one parent was unaffected, and all four grandparents were of French-Canadian origin. For family members, they were considered as asthmatic: (1) if they had a reported history of asthma (validated by a physician), or (2) if

they presented asthma-related symptoms and positive PC20 (less or equal to 8mg/ml of methacholine) at recruitment. If individuals had at least one positive response on skin prick tests (wheal diameter X 3mm at 10min), they were defined as atopic. Age of onset was obtained from answers to the following questions "Have you ever had asthma attacks? How old were you when you had your first asthma attack?". When age of onset was defined below 2 years (in 41 cases), a default class of 2 years was adopted to avoid uncertainty. For non-asthmatics, we considered the age at examination.

UFA

UFA is a population-based case-control study of asthma cases and controls matched on age and sex and recruited between 1999 and the year 2007.[12] Subjects are of different ethnic origins (Russians, Tatars and Bashkirs) from Volga-Ural region of Russian Federation.

Cases are unrelated patients with physician-diagnosed asthma and controls are free of disease. Asthma patients were diagnosed by pulmonologists on the basis of clinical examination, family and medication history, objective tests of lung function. The controls were healthy subjects who met all the following criteria: (1) no symptoms or history of asthma or other pulmonary diseases; (2) no symptoms or history of atopy; and (3) absence of first-degree relatives with a history of asthma or atopy. The age of asthma onset was obtained from answer to the following question "How old were you when you had your first attack of asthma?". For non-asthmatics, we considered the age at examination.

TOMSK

TOMSK is a population-based family study conducted by the Research Institute of Medical Genetics and Siberian State Medical University (TOMSK, Russia) from 1998 onwards.[13,14]

Both nuclear families and extended pedigrees were recruited through atopic bronchial asthmatic probands. Both probands and their relatives were clinically examined to establish diagnosis of asthma and atopy using the GINA criteria (Global Initiative for Asthma: Global Strategy for Asthma Management and Prevention. http://www.ginasthma.org). The age of onset was set as the age when asthma was first diagnosed by a doctor. For newly identified cases, it was established through their physical examination, while for other cases, it was established through the reply to a question: "What age were you when doctor first time told you that you have asthma?". For non-asthmatics, we considered the age at examination.

eQTL analysis and functional annotations

We assessed whether significant SNPs associated with time-to-asthma onset (or their proxies) were expression quantitative trait loci (eQTLs) by using publically available databases: eQTL Browser (lymphoblastoid cell lines (LCLs) from British asthma (MRCA) and eczema (MRCE) family subjects),[15] Blood eQTL Browser (non-transformed peripheral blood samples),[16] Lung eQTLs (lung tissue),[17] GTEx eQTL Browser release v4 (uploaded in July 2015 for multiple tissues including blood and lung)[18] and eQTL Chicago Browser that includes eQTL results from many sources among which Montgomery *et al*,[19] Stranger *et al*[20] and Veyrieras *et al*[21] that were performed in human LCLs.

Replication of prior asthma GWAS results

Finally, we evaluated whether previously reported susceptibility loci for asthma and asthma-related phenotypes (asthma exacerbation, asthma-plus-rhinitis comorbidity, age of asthma onset and bronchial-hyper-responsiveness) at genome-wide significance level were associated with time-to-asthma onset in our meta-analysis using data from the NHGRI Catalog of Published GWASs (last update June 2015).[22] A total of 15 GWASs mainly conducted on European

populations have reported 57 SNPs belonging to 28 independent loci (including seven loci specific to Japanese subjects or African-Americans and Latinos) associated with asthma and/or selected asthma-related phenotypes at a genome-wide significant level.

RESULTS

Functional annotations and effect on gene expression of main time-to-asthma onset associated SNPs

To provide some insights into the potential molecular mechanisms underlying the TAO associated variants, we queried whether the seven top SNPs (or their proxies) were 1) tagging potentially deleterious SNPs, 2) located in regulatory elements, and 3) reported to influence the expression of one or more of nearby genes (eQTLs at $P<5x10^{-5}$; see Table III and Table E3 for summary and complete results respectively).

The 2q12 region top SNP (rs10208293) is located within *interleukin 1 receptor-like 1* gene (*IL1RL1*) in a binding site of the nuclear factor-κB (NFkB1) and is in strong LD (D'=1 and r^2=0.57) with three *IL1RL1* missense SNPs (rs4988956, rs10192157 and rs10206753). This SNP and one of its proxy correlate with the expression of *interleukin 18 receptor 1* (*IL18R1*) and *IL18 receptor accessory protein* (*IL18RAP*) genes in blood.[16]

In 6p21 region, the strongest association was with rs9272346 located near *HLA-DQA1*. This SNP lies within an enhancer histone mark in B cells and is in strong LD (r^2>0.8) with SNPs in a predicted promoter in B cells. This SNP (or its proxies) correlates with the expression of 17 HLA class II genes in blood tissue and lymphoblastoid cell lines (LCL).[15,16,18-20] However, due to the extensive LD within HLA region, some of these associations might reflect signal inter-correlations rather than true pleiotropic effects on gene expression.

The two distinct SNPs (rs928413 and rs413382) associated with TAO at 9p24 are located in intergenic region upstream of *IL33*. We did not find any evidence for eQTL for these SNPs or their proxies.

The span of the SNPs associated with TAO at genome-wide significance level in 17q12-q21 region was approximately 389 kb. The strongest association was with a SNP (rs9901146) nearby *zona pellucida binding protein 2* gene (*ZPBP2*) that is in strong LD with several SNPs tagging and/or belonging to other genes in the region: *IKAROS family zinc finger 3* (*IKZF3*), *ZPBP2 (including rs11557467 missense SNP)*, *gasdermin B* (*GSDMB*; among which two missense SNPs rs2305480 and rs2305479) and *ORMDL sphingolipid biosynthesis regulator 3* (*ORMDL3*). The rs9901146-G risk allele was positively correlated with the expression of *GSDMB* and *ORMDL3*, and negatively with *IKZF3* expression in LCL and blood tissue.[15,16,18,19,21] The second distinct signal at 17q12-q21 (rs3859192) is located within the *gasdermin A* gene (*GSDMA*), in a GABPA (GA binding protein transcription factor, alpha subunit) binding site. Moreover, this SNP is in strong LD with rs56030650 missense SNP (*GSDMA*) and SNPs lying in a predicted promoter in B cells. The rs3859192-T risk allele was correlated with decreased expression of *GSDMA* in the lung[17] and increased expression of both *GSDMB* and *ORMDL3* in LCLs.[15]

REFERENCES

1. Moffatt MF,Gut IG,Demenais F,Strachan DP,Bouzigon E,Heath S, et al. A large-scale, consortium-based genomewide association study of asthma. N Engl J Med 2010; 363:1211-21.
2. Fraser A,Macdonald-Wallis C,Tilling K,Boyd A,Golding J,Davey Smith G, et al. Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. Int J Epidemiol 2013; 42:97-110.

3. The European Community Respiratory Health Survey II. Eur Respir J 2002; 20:1071-9.
4. Burney PG,Luczynska C,Chinn S,Jarvis D. The European Community Respiratory Health Survey. Eur Respir J 1994; 7:954-60.
5. Kauffmann F,Dizier MH,Pin I,Paty E,Gormand F,Vervloet D, et al. Epidemiological study of the genetics and environment of asthma, bronchial hyperresponsiveness, and atopy: phenotype issues. Am J Respir Crit Care Med 1997; 156:S123-9.
6. Genuneit J,Buchele G,Waser M,Kovacs K,Debinska A,Boznanski A, et al. The GABRIEL Advanced Surveys: study design, participation and evaluation of bias. Paediatr Perinat Epidemiol 2011; 25:436-47.
7. Downs SH,Schindler C,Liu LJ,Keidel D,Bayer-Oglesby L,Brutsche MH, et al. Reduced exposure to PM10 and attenuated age-related decline in lung function. N Engl J Med 2007; 357:2338-47.
8. Imboden M,Schwartz J,Schindler C,Curjuric I,Berger W,Liu SL, et al. Decreased PM10 exposure attenuates age-related lung function decline: genetic variants in p53, p21, and CCND1 modify this effect. Environ Health Perspect 2009; 117:1420-7.
9. James AL,Knuiman MW,Bartholomew HC,Musk AB. What can Busselton population health surveys tell us about asthma in older people? Med J Aust 2005; 183:S17-9.
10. James AL,Palmer LJ,Kicic E,Maxwell PS,Lagan SE,Ryan GF, et al. Decline in lung function in the Busselton Health Study: the effects of asthma and cigarette smoking. Am J Respir Crit Care Med 2005; 171:109-14.
11. Laprise C. The Saguenay-Lac-Saint-Jean asthma familial collection: the genetics of asthma in a young founder population. Genes Immun 2014; 15:247-55.
12. Karunas AS,Iunusbaev BB,Fedorova I,Gimalova GF,Ramazanova NN,Gur'eva LL, et al. [Genome-wide association study of bronchial asthma in the Volga-Ural region of Russia]. Mol Biol (Mosk) 2011; 45:992-1003.

13. Freidin MB,Kobyakova OS,Ogorodova LM,Puzyrev VP. Association of polymorphisms in the human IL4 and IL5 genes with atopic bronchial asthma and severity of the disease. Comp Funct Genomics 2003; 4:346-50.
14. Freidin MB,Puzyrev VP,Ogorodova LM,Kobiakova OS,Kulmanakova IM. [Polymorphism of interleukins and interleukin receptor genes: population distribution and association with atopic bronchial asthma]. Genetika 2002; 38:1710-8.
15. Liang L,Morar N,Dixon AL,Lathrop GM,Abecasis GR,Moffatt MF, et al. A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. Genome Res 2013; 23:716-26.
16. Westra HJ,Peters MJ,Esko T,Yaghootkar H,Schurmann C,Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet 2013; 45:1238-43.
17. Hao K,Bosse Y,Nickle DC,Pare PD,Postma DS,Laviolette M, et al. Lung eQTLs to help reveal the molecular underpinnings of asthma. PLoS Genet 2012; 8:e1003029.
18. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013; 45:580-5.
19. Montgomery SB,Sammeth M,Gutierrez-Arcelus M,Lach RP,Ingle C,Nisbett J, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 2010; 464:773-7.
20. Stranger BE,Nica AC,Forrest MS,Dimas A,Bird CP,Beazley C, et al. Population genomics of human gene expression. Nat Genet 2007; 39:1217-24.
21. Veyrieras JB,Kudaravalli S,Kim SY,Dermitzakis ET,Gilad Y,Stephens M, et al. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 2008; 4:e1000214.

22. Hindorff LA,Sethupathy P,Junkins HA,Ramos EM,Mehta JP,Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 2009; 106:9362-7.
23. Barrett JC,Fry B,Maller J,Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005; 21:263-5.
24. Ferreira MA,Matheson MC,Duffy DL,Marks GB,Hui J,Le Souef P, et al. Identification of IL6R and chromosome 11q13.5 as risk loci for asthma. Lancet 2011; 378:1006-14.
25. Torgerson DG,Ampleford EJ,Chiu GY,Gauderman WJ,Gignoux CR,Graves PE, et al. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. Nat Genet 2011; 43:887-92.
26. Sleiman PM,Flory J,Imielinski M,Bradfield JP,Annaiah K,Willis-Owen SA, et al. Variants of DENND1B associated with asthma in children. N Engl J Med 2010; 362:36-44.
27. Ramasamy A,Kuokkanen M,Vedantam S,Gajdos ZK,Couto Alves A,Lyon HN, et al. Genome-wide association studies of asthma in population-based cohorts confirm known and suggested loci and identify an additional association near HLA. PLoS One 2012; 7:e44008.
28. Ferreira MA,Matheson MC,Tang CS,Granell R,Ang W,Hui J, et al. Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype. J Allergy Clin Immunol 2014; 133:1564-71.
29. Forno E,Lasky-Su J,Himes B,Howrylak J,Ramsey C,Brehm J, et al. Genome-wide association study of the age of onset of childhood asthma. J Allergy Clin Immunol 2012; 130:83-90 e4.
30. Ding L,Abebe T,Beyene J,Wilke RA,Goldberg A,Woo JG, et al. Rank-based genome-wide analysis reveals the association of ryanodine receptor-2 gene variants with childhood asthma among human populations. Hum Genomics 2013; 7:16.

31. Hirota T,Takahashi A,Kubo M,Tsunoda T,Tomita K,Doi S, et al. Genome-wide association study identifies three new susceptibility loci for adult asthma in the Japanese population. Nat Genet 2011; 43:893-6.
32. Himes BE,Hunninghake GM,Baurley JW,Rafaels NM,Sleiman P,Strachan DP, et al. Genome-wide association analysis identifies PDE4D as an asthma-susceptibility gene. Am J Hum Genet 2009; 84:581-93.
33. Lasky-Su J,Himes BE,Raby BA,Klanderman BJ,Sylvia JS,Lange C, et al. HLA-DQ strikes again: genome-wide association study further confirms HLA-DQ in the diagnosis of asthma among adults. Clin Exp Allergy 2012; 42:1724-33.
34. Noguchi E,Sakamoto H,Hirota T,Ochiai K,Imoto Y,Sakashita M, et al. Genome-wide association study identifies HLA-DP as a susceptibility gene for pediatric asthma in Asian populations. PLoS Genet 2011; 7:e1002170.
35. Bonnelykke K,Sleiman P,Nielsen K,Kreiner-Moller E,Mercader JM,Belgrave D, et al. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. Nat Genet 2014; 46:51-5.
36. McGeachie MJ,Wu AC,Tse SM,Clemmer GL,Sordillo J,Himes BE, et al. CTNNA3 and SEMA3D: Promising loci for asthma exacerbation identified through multiple genome-wide association studies. J Allergy Clin Immunol 2015.
37. Moffatt MF,Kabesch M,Liang L,Dixon AL,Strachan D,Heath S, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 2007; 448:470-3.
38. Wan YI,Shrine NR,Soler Artigas M,Wain LV,Blakey JD,Moffatt MF, et al. Genome-wide association study to identify genetic determinants of severe asthma. Thorax 2012; 67:762-8.

39. Pruim RJ,Welch RP,Sanna S,Teslovich TM,Chines PS,Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 2010; 26:2336-7.

Table E1. Main characteristics of the nine studies included in the meta-analysis of time-to-asthma onset GWAS

	ALSPAC	BUSSELTON	ECRHS	EGEA	GABRIELA	SAPALDIA	SLSJ	TOMSK	UFA
Cohort information									
Country	United-Kingdom	Australia	Europe	France	Austria, Germany, Switzerland	Switzerland	Canada	Russia	Russia
Study collection type	Birth Cohort	Population-based Case/Control	Population-based case-control study	Longitudinal Case/Control and Family study	Population-based Case/Control	Population-based case-control study	Population-based family study	Population-based family study	Population-based case-control study
Sample Size									
	3,420	1,191	2,085	1,835	1,503	1,435	1,127	622	669
Main characteristics									
Sex, men (%)	1,760 (51.5)	510 (42.8)	981 (47.1)	937 (51.1)	849 (56.5)	698 (48.6)	515 (45.7)	341 (54.8)	413 (61.7)
Age in years*, mean (SD)	13.9 (0.1)	53.4 (17.3)	42.8 (7.1)	31.3 (17.0)	9.0 (1.6)	52.3 (11.2)	38.4 (21.6)	27.2 (16.7)	19.2 (13.3)
Asthma, n (%)	1,336 (39.1)	391 (32.8)	618 (29.7)	793 (43.2)	664 (44.2)	557 (38.8)	534 (47.4)	240 (38.6)	329 (49.2)
Asthma age-of-onset, median [25-75%]	1.6 [0.5-5.6]	18 [6-35]	20 [7-30]	9 [3-25]	2 [0.8-4]	21 [7-37]	9 [3-25]	5 [3-12]	5 [2-13]
Atopy**, n (%)	652 (25.1)	507 (42.6)	864 (41.5)	1,063 (57.9)	812 (54.0)	-	618 (54.8)	381 (61.3)	281 (42.0)***
IgE (UI/mL), mean (SD)	267.4 (555.6)	NA	135.9 (296.9)	282.6 (632.4)	NA	119.8 (263.8)	182.8 (202.8)	226.3 (261.4)	238.7 (373.5)
Genotyping									
Genotyping platform and SNP panel	Illumina HumanHap550Quad	Illumina 610K	Illumina 610K	Illumina 610K	Illumina 610K	Illumina 610K	Illumina 610K	Illumina 610K	Illumina 610K

	ALSPAC	BUSSELTON	ECRHS	EGEA	GABRIELA	SAPALDIA	SLSJ	TOMSK	UFA
Genotyping center	23andMe subcontracting the Wellcome Trust Sanger Institute, Cambridge, UK, and the LabCorp, Burlington, North Carolina, US	Centre National de Génotypage, Evry, France	Centre National de Génotypage, Evry, France	Centre National de Génotypage, Evry, France	Centre National de Génotypage, Evry, France	Centre National de Génotypage, Evry, France	Centre National de Génotypage, Evry, France	Centre National de Génotypage, Evry, France	Centre National de Génotypage, Evry, France
Individual QC									
Call-rate	97%	97%	97%	97%	97%	97%	97%	97%	97%
Heterozygosity	Individuals excluded if <0.320 or >0.345 for the Sanger data and <0.310 or >0.330 for the LabCorp data	Individuals excluded if <0.30 or >0.33	Individuals excluded if <0.30 or >0.33	Individuals excluded if <0.30 or >0.33	Individuals excluded if <0.30 or >0.33	Individuals excluded if <0.30 or >0.33	Individuals excluded if <0.30 or >0.33	Individuals excluded if <0.30 or >0.33	Individuals excluded if <0.30 or >0.33
Ethnic outliers	PCA based	PCA based	PCA based	PCA based	PCA based	PCA based	PCA based	PCA based	PCA based
SNP QC filters before imputation									
MAF	1%	5%	5%	5%	5%	5%	5%	5%	5%
HWE p-value	$5x10^{-7}$	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}
Call-rate	95%	97%	97%	97%	97%	97%	97%	97%	97%
Imputation - Genome									
Software	MACH 1.0	MACH 1.0	MACH 1.0	MACH 1.0	MACH 1.0	MACH 1.0	MACH 1.0	MACH 1.0	MACH 1.0
Hapmap release	Hapmap2 r22	Hapmap2 r21	Hapmap2 r21	Hapmap2 r21	Hapmap2 r21	Hapmap2 r21	Hapmap2 r21	Hapmap2 r24	Hapmap2 r21

	ALSPAC	BUSSELTON	ECRHS	EGEA	GABRIELA	SAPALDIA	SLSJ	TOMSK	UFA
SNP QC filters	Rsq ≥ 0.5 & MAF ≥ 1%	Rsq ≥ 0.5 & MAF ≥ 1%	Rsq ≥ 0.5 & MAF ≥ 1%	Rsq ≥ 0.5 & MAF ≥ 1%	Rsq ≥ 0.5 & MAF ≥ 1%	Rsq ≥ 0.5 & MAF ≥ 1%	Rsq ≥ 0.5 & MAF ≥ 1%	Rsq ≥ 0.5 & MAF ≥ 1%	Rsq ≥ 0.5 & MAF ≥ 1%
Imputation - Region									
Software	MINIMAC	IMPUTE2 v2.1.2	IMPUTE2 v2.1.2	IMPUTE2 v2.1.2	IMPUTE2 v2.1.2	IMPUTE2 v2.1.2	IMPUTE2 v2.1.2	IMPUTE2 v2.1.2	IMPUTE2 v2.1.2
1000G release	November 2010	June 2014	June 2014	June 2014	June 2014	June 2014	June 2014	June 2014	June 2014
SNP QC filters	Rsq ≥ 0.5	Info ≥ 0.5	Info ≥ 0.5	Info ≥ 0.5	Info ≥ 0.5	Info ≥ 0.5	Info ≥ 0.5	Info ≥ 0.5	Info ≥ 0.5

*Age at last examination

**Atopy defined by a positive skin prick test response to at least one aeroallergen

***Available only in asthmatics

Table E2. Results of the analyses conducted in 16q12 region using 1000G CEU reference sample.

Marker	Position*	Alleles Effect/Ref†	Effect Freq	Hazard Ratio [95% CI]	P-value‡	P-Het**
rs11867101	50 847 368	T/C	0.03	1.32 [1.19-1.45]	$1.1x10^{-7}$	0.16
rs11863019	50 847 819	C/A	0.04	1.32 [1.19-1.47]	$6.3x10^{-8}$	0.12
rs4785228	50 848 914	A/G	0.03	1.32 [1.19-1.47]	$1.4x10^{-7}$	0.17
rs7199870	50 850 082	T/C	0.03	1.32 [1.19-1.47]	$9.2x10^{-8}$	0.13
rs2032688	50 850 847	T/C	0.03	1.32 [1.19-1.47]	$5.8x10^{-8}$	0.14
rs7195092	50 852 366	G/C	0.03	1.32 [1.19-1.47]	**$3.8x10^{-8}$**	0.11
rs2032687	50 852 432	T/C	0.03	1.32 [1.19-1.47]	$1.2x10^{-7}$	0.11
rs4785458	50 856 194	A/G	0.03	1.32 [1.19-1.47]	$6.2x10^{-8}$	0.10
rs1861760	50 857 693	A/C	0.04	1.28 [1.16-1.41]	$2.6x10^{-7}$	0.10

*Position in base pairs (bp) – build 37.3 NCBI.

†For the calculation of the hazard ratios, effect alleles (Effect) were designated as risk alleles. Effect Freq denotes effect allele frequency, CI confidence interval, and Ref reference allele.

‡P-values are obtained from meta-analysis of single-SNP Cox model of time-to-asthma onset adjusted for sex and principal components.

**P-Het value reflect test of heterogeneity across studies using Cochran's Q test.

Table E3. Cis-eQTLs results for the top SNPs (and their proxies) in genome-wide associated regions from the meta-analysis of time-to-asthma onset. We focused our search on eQTLs measured in blood, LCLs and lung tissue.

Chr	SNP	Position (build 37.3)	LD (D'/r^2) with main SNP	Alleles† (Ref/Effect)	Z score / LOD	P-value	Gene	FDR	Source	Tissue	Reference
2	**rs10208293***	102 966 310	-	G/A	34.06	$9.8x10^{-198}$	*IL18RAP*	$<10^{-5}$	Blood eQTLs	Blood	Westra *et al*, 2013
				G/A	11.65	$2.5x10^{-13}$	*IL18R1*	NA	eQTL Browser	LCLs (eczema)	Liang *et al*, 2013
	rs3771167	102 986 188	rs10208293 (D'=1, r^2=0.56)	G/A	7.34	$2.1x10^{-13}$	*IL18R1*	$<10^{-5}$	Blood eQTLs	Blood	Westra *et al*, 2013
6	**rs9272346***	32 604 372	-	NA	NA	$2.1x10^{-21}$	*HLA-DQA1*	NA	eQTL_Chicago	LCLs	Stranger et al, 2007
				NA	NA	$4.6x10^{-21}$	*HLA-DQA1*	NA	eQTL_Chicago	LCLs	Montgomery *et al*, 2010
				NA	NA	$8.3x10^{-18}$	*HLA-DQB1*	NA	eQTL_Chicago	LCLs	Montgomery *et al*, 2010
				NA	NA	$1.1x10^{-7}$	*HLA-DRB1*	NA	eQTL_Chicago	LCLs	Montgomery *et al*, 2010
				G/A	-10.61	$1.4x10^{-26}$	*HLA-DQA1*	NA	GTEx	Lung	GTEx consortium, 2013
				G/A	-12.57	$1.6x10^{-36}$	*HLA-DQA1*	NA	GTEx	Whole_Blood	GTEx consortium, 2013
				G/A	8.98	$1.4x10^{-19}$	*HLA-DQA2*	NA	GTEx	Whole_Blood	GTEx consortium, 2013
				G/A	6.63	$1.7x10^{-11}$	*HLA-DQA2*	NA	GTEx	Lung	GTEx consortium, 2013
				G/A	-8.84	$4.6x10^{-19}$	*HLA-DQB1*	NA	GTEx	Lung	GTEx consortium, 2013
				G/A	-12.16	$2.4x10^{-34}$	*HLA-DQB1*	NA	GTEx	Whole_Blood	GTEx consortium, 2013

Chr	SNP	Position (build 37.3)	LD (D'/r^2) with main SNP	Alleles† (Ref/Effect)	Z score / LOD	P-value	Gene	FDR	Source	Tissue	Reference
				G/A	-8.72	$1.4x10^{-18}$	*HLA-DQB1-AS1*	NA	GTEx	Lung	GTEx consortium, 2013
				G/A	-11.34	$4x10^{-30}$	*HLA-DQB1-AS1*	NA	GTEx	Whole_Blood	GTEx consortium, 2013
				G/A	10.06	$4x10^{-24}$	*HLA-DQB2*	NA	GTEx	Whole_Blood	GTEx consortium, 2013
				G/A	-6.85	$7.5x10^{-12}$	*HLA-DRA*	$<10^{-5}$	Blood eQTLs	Blood	Westra *et al*, 2013
				G/A	-5.85	$2.5x10^{-9}$	*HLA-DRB1*	NA	GTEx	Lung	GTEx consortium, 2013
				G/A	-7.43	$5.3x10^{-14}$	*HLA-DRB1*	NA	GTEx	Whole_Blood	GTEx consortium, 2013
				G/A	-23.43	$2.1x10^{-121}$	*HLA-DRB5*	$<10^{-5}$	Blood eQTLs	Blood	Westra *et al*, 2013
				G/A	-5.15	$1.3x10^{-7}$	*HLA-DRB5*	NA	GTEx	Whole_Blood	GTEx consortium, 2013
				G/A	4.70	$1.3x10^{-6}$	*HLA-DRB6*	NA	GTEx	Whole_Blood	GTEx consortium, 2013
				G/A	-6.60	$4.1x10^{-11}$	*TAP2*	$<10^{-5}$	Blood eQTLs	Blood	Westra *et al*, 2013
	rs3129889	32 413 545	rs9272346 (D'=1, r^2=0.41)	G/A	-6.68	$2.9x10^{-8}$	*HLA-DRB1*	NA	eQTL Browser	LCLs (asthma)	Liang *et al*, 2013
	rs9272723	32 609 427	rs9272346 (D'=1, r^2=0.97)	T/C	-10.10	$5.7x10^{-24}$	*TAP2*	$<10^{-5}$	Blood eQTLs	Blood	Westra *et al*, 2013
					7.23	$4.8x10^{-13}$	*HLA-DOB*	$<10^{-5}$			
					-25.04	$2.3x10^{-138}$	*HLA-DRB5*	$<10^{-5}$			

Chr	SNP	Position (build 37.3)	LD (D'/r^2) with main SNP	Alleles† (Ref/Effect)	Z score / LOD	P-value	Gene	FDR	Source	Tissue	Reference
	rs9273325	32 623 193	rs9272346 (D'=1, r^2=0.02)	A/G	-4.50	$6.9x10^{-6}$	*TAP1*	NA	eQTLs_Lung	Lung	Hao *et al*, 2012
	rs2859579	32 784 073	rs9272346 (D'=1, r^2=0.007)	T/G	19.10	$6.8x10^{-21}$	*TAP2*	NA	eQTL Browser	LCLs (asthma)	Liang *et al*, 2013
	rs9277725	33 091 543	rs9272346 (D'=1, r^2=0.05)	T/A	15.60	$2.3x10^{-17}$	*HLA-DPB2*	NA	eQTL Browser	LCLs (asthma)	Liang *et al*, 2013
	rs2395357	33 101 006	rs9272346 (D'=1, r^2=0.05)	A/G	7.80	$6.2x10^{-15}$	*HSD17B8*	$<10^{-5}$	Blood eQTLs	Blood	Westra *et al*, 2013
					-6.66	$2.8x10^{-11}$	*HLA-DPB1*	$<10^{-5}$			
					-5.12	$3.1x10^{-7}$	*HLA-DMA*	10^{-4}			
16	**rs1861760***	50 857 693	-	C/A	6.62	$3.6x10^{-11}$	*NOD2*	$<10^{-5}$	Blood eQTLs	Blood	Westra *et al*, 2013
	rs5743266 (now rs2076752)	50 731 096	rs1861760 (D'=1, r^2=0.02)	A/G	-5.85	$5.0x10^{-9}$	*CYLD*	$<10^{-5}$	Blood eQTLs	Blood	Westra *et al*, 2013
					23.31	$3.2x10^{-120}$	*NOD2*	$<10^{-5}$			
	rs7205760	50 844 773	rs1861760 (D'=1, r^2=0.005)	C/G	4.69	$2.8x10^{-6}$	*CYLD*	NA	eQTLs_Lung	Lung	Hao *et al*, 2012
					7.85	$4.0x10^{-15}$	*NOD2*	$<10^{-5}$	Blood eQTLs	Blood	Westra *et al*, 2013
17	**rs9901146***	38 043 343	-	A/G	36.54	$9.8x10^{-198}$	*GSDMB*	$<10^{-5}$	Blood eQTLs	Blood	Westra *et al*, 2013

Chr	SNP	Position (build 37.3)	LD (D'/r^2) with main SNP	Alleles† (Ref/Effect)	Z score / LOD	P-value	Gene	FDR	Source	Tissue	Reference
				A/G	6.00	$9.9x10^{-10}$	*GSDMB*	NA	GTEx	Whole_Blood	GTEx consortium, 2013
				A/G	8.55	$1.3x10^{-17}$	*GSDMB*	NA	eQTL Browser	LCLs (asthma)	Liang *et al*, 2013
				A/G	33.50	$2.2x10^{-35}$	*GSDMB*	NA	eQTL Browser	LCLs (eczema)	Liang *et al*, 2013
				A/G	36.41	$9.8x10^{-198}$	*ORMDL3*	$<10^{-5}$	Blood eQTLs	Blood	Westra *et al*, 2013
				A/G	11.16	$6.2x10^{-29}$	*ORMDL3*	NA	eQTL Browser	LCLs (asthma)	Liang *et al*, 2013
				A/G	42.30	$3.3x10^{-44}$	*ORMDL3*	NA	eQTL Browser	LCLs (eczema)	Liang *et al*, 2013
				NA	NA	$1.3x10^{-10}$	*ORMDL3*	NA	eQTL_Chicago	LCLs	Veyrieras *et al*, 2008
				NA	NA	$3.8x10^{-6}$	*ORMDL3*	NA	eQTL_Chicago	LCLs	Montgomery *et al*, 2010
				A/G	4.57	$2.4x10^{-6}$	*ORMDL3*	NA	GTEx	Whole_Blood	GTEx consortium, 2013
	rs9896940	37 895 975	rs9901146 (D'=1, r^2=0.07)	G/A	-15.81	$2.6x10^{-56}$	*IKZF3*	$<10^{-5}$	Blood eQTLs	Blood	Westra *et al*, 2013
17	**rs3859192***	38 128 648	-	C/T	-6.91	$2.5x10^{-12}$	*GSDMA*	NA	GTEx	Lung	GTEx consortium, 2013
				C/T	6.10	$1.1x10^{-7}$	*GSDMB*	NA	eQTL Browser	LCLs (eczema)	Liang *et al*, 2013
				C/T	8.10	$1.1x10^{-9}$	*ORMDL3*	NA	eQTL Browser	LCLs (eczema)	Liang *et al*, 2013

*Top Genome-wide significant SNPs in time-to-asthma onset meta-analysis and secondary associations identified by conditional analyses are indicated in bold

†Haplotype reconstruction was done using Haploview.[23] The effect allele of the top SNP is always transmitted with the indicated effect allele of its proxy

Table E4. Comparison of the main results of time-to-asthma onset (TAO, in bold) GWAS meta-analysis ($P \leq 5x10^{-8}$) with asthma (binary trait) GWAS meta-analysis results obtained in the same nine studies and in the whole GABRIEL dataset (25 studies, N=26,475)[1]

						TAO meta-analysis 9 studies			**AST (binary) meta-analysis 9 studies**														**AST (binary) meta-analysis All GABRIEL**			
									ALL					**Childhood-onset**					**Adult-onset**								
Chr	**Marker**	**Position***	**Closest Gene (kb distance)**	**Effect allele Freq**	**Effect/ Ref Alleles†**	**HR**	**P‡**	**Phet****	**OR fix**	**P fix**	**OR ran**	**P ran**	**Phet**	**OR fix**	**P fix**	**OR ran**	**P ran**	**Phet**	**OR fix**	**P fix**	**OR ran**	**P ran**	**P het**	**P Het**	**OR**	**P ran**	**P fix**	**P het**
2	**rs10208293**	102.97	*IL1RL1*	0.27	A/G	0.88	$3.1x10^{-8}$	0.26	0.88	$4.0x10^{-5}$	0.88	$4.7x10^{-3}$	0.03	0.84	$1.4x10^{-5}$	0.84	$1.4x10^{-5}$	0.84	0.94	$2.7x10^{-1}$	0.97	$7.8x10^{-1}$	0.004	0.07				
2	rs3771166	102.99	*IL18R1*	0.38	A/G	0.89	$5.0x10^{-8}$	0.57	0.88	$5.5x10^{-6}$	0.88	$4.2x10^{-4}$	0.10	0.84	$6.0x10^{-7}$	0.84	$6.0x10^{-7}$	0.53	0.96	$3.4x10^{-1}$	0.97	$6.1x10^{-1}$	0.13	0.02	1.15	$3.4x10^{-9}$	$3.5x10^{-12}$	0.18
6	**rs9272346††**	32.60	*HLA-DQA1 (0.8)*	0.58	A/G	1.13	$1.6x10^{-8}$	0.12	1.17	$4.9x10^{-8}$	1.17	$1.2x10^{-7}$	0.38	1.13	$6.1x10^{-4}$	1.14	$2.5x10^{-3}$	0.21	1.25	$6.3x10^{-6}$	1.25	$6.3x10^{-6}$	0.83	0.12	1.18	$7.0x10^{-14}$	$7.0x10^{-14}$	0.50
9	**rs413382**	6.14	*IL33 (73)*	0.80	A/C	1.16	$5.9x10^{-8}$	0.84	1.20	$3.3x10^{-7}$	1.22	$1.9x10^{-4}$	0.01	1.19	$2.2x10^{-4}$	1.19	$5.8x10^{-3}$	0.06	1.23	$3.7x10^{-4}$	1.26	$1.8x10^{-2}$	0.02	0.63				
9	rs1342326	6.19	*IL33 (26)*	0.84	A/C	0.84	$1.6x10^{-12}$	0.43	0.80	$2.1x10^{-9}$	0.80	$6.9x10^{-6}$	0.05	0.74	$4.8x10^{-11}$	0.73	$7.1x10^{-9}$	0.26	0.93	$1.9x10^{-1}$	0.92	$1.9x10^{-1}$	0.41	0.003	1.20	$9.2x10^{-10}$	$8.7x10^{-12}$	0.22
9	**rs928413**	6.21	*IL33 (2)*	0.76	A/G	0.84	$6.5x10^{-16}$	0.15	0.80	$2.2x10^{-12}$	0.80	$2.5x10^{-7}$	0.04	0.75	$5.4x10^{-13}$	0.75	$4.7x10^{-9}$	0.16	0.90	$3.0x10^{-2}$	0.90	$4.7x10^{-2}$	0.33	0.006				
15	rs744910	6.74	*SMAD3*	0.51	A/G	0.93	$3.2x10^{-4}$	0.60	0.92	$1.9x10^{-3}$	0.92	$1.9x10^{-3}$	0.87	0.90	$2.0x10^{-3}$	0.90	$2.0x10^{-3}$	0.81	0.95	$2.9x10^{-1}$	0.95	$2.9x10^{-1}$	0.74	0.31	1.12	$3.9x10^{-9}$	$3.9x10^{-9}$	0.85
16	**rs1861760**	50.86	*CYLD (22)*	0.04	A/C	1.28	$4.2x10^{-8}$	0.11	1.36	$3.8x10^{-6}$	1.37	$3.3x10^{-5}$	0.22	1.35	$3.8x10^{-4}$	1.36	$1.1x10^{-2}$	0.05	1.37	$3.1x10^{-3}$	1.37	$3.1x10^{-3}$	0.79	0.93				
17	**rs9901146**	38.04	*ZPBP2 (9)*	0.48	A/G	0.85	$1.9x10^{-16}$	0.17	0.85	$4.0x10^{-9}$	0.85	$7.3x10^{-5}$	0.01	0.78	$3.0x10^{-12}$	0.78	$3.0x10^{-12}$	0.61	0.98	$5.9x10^{-1}$	0.96	$5.4x10^{-1}$	0.15	0.0002				
17	rs2305480	38.06	*GSDMB*	0.42	A/G	0.85	$8.1x10^{-16}$	0.14	0.85	$1.7x10^{-8}$	0.86	$3.7x10^{-4}$	0.004	0.77	$4.1x10^{-13}$	0.77	$4.1x10^{-13}$	0.83	1.01	$8.3x10^{-1}$	1.00	$9.7x10^{-1}$	0.20	$4.9x10^{-6}$	1.18	$9.6x10^{-8}$	-	0.0009
17	rs3894194	38.12	*GSDMA*	0.47	A/G	1.16	$1.4x10^{-13}$	0.89	1.17	$1.7x10^{-8}$	1.17	$1.4x10^{-6}$	0.19	1.25	$8.6x10^{-11}$	1.25	$8.6x10^{-11}$	0.88	1.04	$4.3x10^{-1}$	1.04	$4.3x10^{-1}$	0.55	$9.3x10^{-4}$	1.17	$4.6x10^{-9}$	-	0.02
17	**rs3859192**	38.13	*GSDMA*	0.54	C/T	0.86	$1.5x10^{-13}$	0.90	0.86	$2.7x10^{-8}$	0.86	$2.7x10^{-8}$	0.64	0.81	$1.7x10^{-9}$	0.81	$1.7x10^{-9}$	0.95	0.95	$2.2x10^{-1}$	0.95	$2.2x10^{-1}$	0.82	0.009				
22	rs2284033	37.53	*IL2RB*	0.47	A/G	0.94	$2.1x10^{-3}$	0.36	0.91	$1.1x10^{-3}$	0.91	$1.1x10^{-3}$	0.66	0.94	$7.9x10^{-2}$	0.94	$7.9x10^{-2}$	0.46	0.87	$2.0x10^{-3}$	0.87	$2.0x10^{-3}$	0.86	0.16	1.12	$1.2x10^{-8}$	$1.2x10^{-8}$	0.92

LD between TAO main SNPs and GABRIEL main SNPs in genome-wide associated regions:

2q12: rs10208293 & rs3771166 (r^2=0.53, D'=0.95); **6p21**: rs9272346 & rs9273349 (r^2=D'=1.0); **9p24**: rs928413 & rs1342326 (r^2=0.51, D'=1.0), rs413382 - rs1342326 (r^2=0.0, D'=0.23); **17q12-21**: rs9901146 & rs2305480 (r^2=0.82, D'=1.0), rs3859192 & rs3894194 (r^2=0.43, D'=0.71)

*Position in megabases (Mb) – build 37.3 NCBI

†For the calculation of the hazard ratios, effect alleles were designated as risk alleles. Effect Freq denotes frequency of the effect allele, CI confidence interval, and Ref reference allele.

‡P-value obtained from single-SNP Cox model for time-to-asthma onset adjusted for sex and principal components (fixed-effect model when there was no significant evidence of heterogeneity or random-effect model otherwise)

**P-Het reflects P-value of Cochran's Q statistic across studies

†† main GABRIEL SNP in 6p21 (rs9273349) was not imputable

‡‡ Additional distinct SNPs detected with conditional analyses

Table E5. Published genome-wide associations with asthma compared with time-to-asthma onset GWAS meta-analysis results

				Meta-analysis of time-to-asthma onset GWASs				**GW results reported in GWAS for asthma and asthma-related traits** **NCBI GWAS Catalog, June 2015**							
Chr	**SNP or proxy***	**Pos (Mb)**	**Gene Symbol**	**Effect/Ref Alleles†**	**Effect Freq**	**HR [95% CI]**	**P‡**	**Mapped Genes**	**Effect Allele**	**Effect Freq in ctrls**	**P**	**OR [95% CI]**	**Trait**	**References**	**Pop**
1	rs4129267	154.43	*IL6R*	C/T	0.57	0.97 [0.94-1.01]	0.19	*IL6R*	T	0.37	$2.0x10^{-8}$	1.09 [1.06-1.12]	Asthma	Ferreira, Lancet, 2011[24]	European
1	rs1101999	158.93	*PYHIN1*	NA	NA	NA	NA	*PYHIN1*	NA	NA	$4.0x10^{-9}$	NA	Asthma	Torgerson, Nat Genet, 2011[25]	African American & Latinos
1	rs2786098	197.33	*CRB1*	G/T	0.78	1.03 [0.98-1.08]	0.25	*CRB1-DENND1B*	NA	0.85	$2.0x10^{-13}$	1.43 [NA]	Asthma	Sleiman, N Engl J Med, 2010[26]	European
2	rs3771180	102.95	*IL1RL1*	G/T	0.86	1.16 [1.1-1.23]	**$5.9x10^{-7}$**	*IL1RL1*	NA	NA	$2.0x10^{-15}$	NA	Asthma	Torgerson, Nat Genet, 2011[25]	Multi-ethnic
2	rs13408661	102.96	*IL1RL1*	A/G	0.14	0.86 [0.81-0.91]	**$5.9x10^{-7}$**	*IL1RL1*	G	0.84	$1.0x10^{-9}$	1.23 [1.15-1.31]	Asthma	Ramasamy, PLoS One, 2012[27]	European

				Meta-analysis of time-to-asthma onset GWASs				GW results reported in GWAS for asthma and asthma-related traits NCBI GWAS Catalog, June 2015							
Chr	**SNP or proxy***	**Pos (Mb)**	**Gene Symbol**	**Effect/Ref Alleles†**	**Effect Freq**	**HR [95% CI]**	**P‡**	**Mapped Genes**	**Effect Allele**	**Effect Freq in ctrls**	**P**	**OR [95% CI]**	**Trait**	**References**	**Pop**
2	rs10197862	102.97	*IL1RL1*	A/G	0.86	1.16 [1.09-1.23]	**$9.8x10^{-7}$**	*IL1RL1*	A	0.85	$4.0x10^{-11}$	1.24 [1.16-1.32]	Asthma & hay fever	Ferreira, J Allergy Clin Immunol, 2014[28]	European
2	rs3771166	102.99	*IL18R1*	A/G	0.35	0.89 [0.86-0.93]	**$5.0x10^{-8}$**	*IL18R1*	G	0.62	$3.0x10^{-9}$	1.15 [1.10-1.20]	Asthma	Moffatt, N Engl J Med, 2010[1]	European
3	rs9815663	3.61	.	C/T	0.82	1 [0.95-1.05]	0.96	*CRBN - LRRN1*	T	0.182	$2.0x10^{-8}$	0.84 [NA]	Childhood Asthma	**Forno, J Allergy Clin Immunol, 2012[29]**	European
4	rs4833095	38.80	*TLR1*	C/T	0.27	0.95 [0.9-0.99]	0.02	*TLR1*	T	0.74	$5.0x10^{-12}$	1.2 [1.14-1.26]	Asthma & hay fever	Ferreira, J Allergy Clin Immunol, 2014[28]	European
4	rs17218161	59.21	.	NA	NA	NA	NA	*SRIP1 - MIR548AG1*	NA	NA	$2.0x10^{-8}$	NA	Childhood Asthma	Ding, Hum Genomics, 2013[30]	European
4	rs7686660	144.00	.	G/T	0.24	0.99 [0.94-1.03]	0.50	*FLJ44477 - USP38*	T	0.27	$2.0x10^{-12}$	1.16 [1.11-1.21]	Asthma	Hirota, Nat Genet, 2011[31]	Japanese
4	rs3805236	144.36	*GAB1*	A/G	0.30	0.99 [0.95-1.04]	0.80	*GAB1*	G	0.25	$7.0x10^{-8}$	1.20 [1.14-1.26]	Asthma	Hirota, Nat Genet, 2011[31]	Japanese
5	rs1588265	59.37	.	A/G	0.70	0.99 [0.94-1.03]	0.50	*PDE4D*	C	0.29	$3.0x10^{-8}$	1.18 [1.08-1.30]	Asthma	Himes, Am J Hum Genet, 2009[32]	European
5	rs1837253	110.40	.	C/T	0.77	1.13 [1.05-1.2]	**$5.6x10^{-4}$**	*SLC25A46 - TSLP*	C	0.35	$1.0x10^{-16}$	1.17 [1.13-1.22]	Asthma	Hirota, Nat Genet, 2011[31]	Japanese

				Meta-analysis of time-to-asthma onset GWASs				GW results reported in GWAS for asthma and asthma-related traits NCBI GWAS Catalog, June 2015							
Chr	**SNP or proxy***	**Pos (Mb)**	**Gene Symbol**	**Effect/Ref Alleles†**	**Effect Freq**	**HR [95% CI]**	**P‡**	**Mapped Genes**	**Effect Allele**	**Effect Freq in ctrls**	**P**	**OR [95% CI]**	**Trait**	**References**	**Pop**
			.	C/T	0.77	1.13 [1.05-1.2]	**$5.6x10^{-4}$**	*SLC25A46 - TSLP*	NA	NA	$1.0x10^{-14}$	NA	Asthma	Torgerson, Nat Genet, 2011[25]	Multi-ethnic
			.	C/T	0.77	1.13 [1.05-1.2]	**$5.6x10^{-4}$**	*SLC25A46 - TSLP*	C	0.71	$1.0x10^{-9}$	1.17 [1.12-1.24]	Asthma & hay fever	Ferreira, J Allergy Clin Immunol, 2014[28]	European
5	rs1438673	110.47	.	C/T	0.54	1.08 [1.04-1.13]	**$4.5x10^{-5}$**	*WDR36 - RPS3AP21*	C	0.49	$3.0x10^{-11}$	1.16 [1.11-1.21]	Asthma & hay fever	Ferreira, J Allergy Clin Immunol, 2014[28]	European
6	rs204993	32.16	*PBX2*	A/G	0.76	0.96 [0.92-1.01]	0.08	*PBX2*	A	0.58	$2.0x10^{-15}$	1.17 [1.12-1.21]	Asthma	Hirota, Nat Genet, 2011[31]	Japanese
6	rs404860	32.18	*NOTCH4*	C/T	0.18	1 [0.91-1.11]	0.95	*NOTCH4*	A	0.5	$4.0x10^{-23}$	1.21 [1.16-1.25]	Asthma	Hirota, Nat Genet, 2011[31]	Japanese
6	rs3129943	32.34	*C6orf10*	A/G	0.75	1.01 [0.97-1.06]	0.62	*C6orf10*	T	0.62	$3.0x10^{-15}$	1.17 [1.12-1.21]	Asthma	Hirota, Nat Genet, 2011[31]	Japanese
6	rs3117098	32.36	.	A/G	0.66	1.02 [0.98-1.07]	0.29	*HNRNPA1P2 - BTNL2*	G	0.25	$5.0x10^{-12}$	1.16 [1.11-1.21]	Asthma	Hirota, Nat Genet, 2011[31]	Japanese
6	rs9268516	32.38	.	C/T	0.68	0.93 [0.87-0.99]	0.02	*BTNL2-HLA-DRA*	T	0.24	$1.0x10^{-8}$	1.15 [1.10-1.21]	Asthma	Ramasamy, PLoS One, 2012[27]	European
6	rs3129890	32.43	.	A/C	0.22	0.97	0.15	*HLA-DRA - HLA-DRB9*	T	0.61	$5.0x10^{-13}$	1.15	Asthma	Hirota, Nat Genet, 2011[31]	Japanese

				Meta-analysis of time-to-asthma onset GWASs				GW results reported in GWAS for asthma and asthma-related traits NCBI GWAS Catalog, June 2015							
Chr	**SNP or proxy***	**Pos (Mb)**	**Gene Symbol**	**Effect/Ref Alleles†**	**Effect Freq**	**HR [95% CI]**	**P‡**	**Mapped Genes**	**Effect Allele**	**Effect Freq in ctrls**	**P**	**OR [95% CI]**	**Trait**	**References**	**Pop**
	rs9268856*					[0.92-1.01]						[1.11-1.20]			
6	rs9272346	32.60	.	A/G	0.61	1.13 [1.08-1.17]	**$1.6x10^{-8}$**	*HLA-DQA1*	NA	NA	$2.0x10^{-8}$	NA	Asthma	Lasky-Su, Clin Exp Allergy, 2012[33]	European
6	rs9273349 rs9272346*	32.60	.	A/G	0.61	1.13 [1.08-1.17]	**$1.6x10^{-8}$**	*HLA-DQA1 - HLA-DQB1*	C	0.58	$7.0x10^{-14}$	1.18 [1.13-1.24]	Asthma	Moffatt, N Engl J Med, 2010[1]	European
6	rs9273373	32.63	.	NA	NA	NA	NA	*HLA-DQA1 - HLA-DQB1*	G	0.54	$4.0x10^{-14}$	1.24 [1.17-1.30]	Asthma & hay fever	Ferreira, J Allergy Clin Immunol, 2014[28]	European
6	rs7775228	32.66	.	C/T	0.16	1.05 [0.98-1.13]	0.18	*HLA-DQB1 - HLA-DQA2*	A	0.63	$5.0x10^{-15}$	1.17 [1.12-1.21]	Asthma	Hirota, Nat Genet, 2011[31]	Japanese
6	rs9275698	32.69	.	A/G	0.60	1.02 [0.98-1.07]	0.29	*HLA-DQB1 - HLA-DQA2*	T	0.79	$5.0x10^{-12}$	1.18 [1.12-1.24]	Asthma	Hirota, Nat Genet, 201 [31]	Japanese
6	rs9500927	32.96	.	A/G	0.16	1 [0.95-1.05]	0.98	*BRD2 - HLA-DOA*	T	0.26	$4.0x10^{-9}$	1.13 [1.09-1.18]	Asthma	Hirota, Nat Genet, 2011[31]	Japanese
6	rs987870	33.04	.	A/G	0.82	0.97 [0.92-1.02]	0.28	*HLA-DPA1; HLA-DPB1*	C	0.14	$2.0x10^{-10}$	1.4 [1.26-1.55]	Asthma	Noguchi, PLoS Genet, 2011[34]	Japanese
7	rs6967330	105.66	*FLJ23834*	A/G	0.18	1.13 [1.07-1.19]	**$1.4x10^{-5}$**	*CDHR3*	A	0.19	$3.0x10^{-14}$	1.26 [1.18-1.33]	Childhood Asthma	Bonnelykke, Nat Genet, 2014[35]	European

				Meta-analysis of time-to-asthma onset GWASs				GW results reported in GWAS for asthma and asthma-related traits NCBI GWAS Catalog, June 2015							
Chr	**SNP or proxy***	**Pos (Mb)**	**Gene Symbol**	**Effect/Ref Alleles†**	**Effect Freq**	**HR [95% CI]**	**P‡**	**Mapped Genes**	**Effect Allele**	**Effect Freq in ctrls**	**P**	**OR [95% CI]**	**Trait**	**References**	**Pop**
8	rs7009110	81.29	.	C/T	0.59	0.93 [0.9-0.97]	**$7.6x10^{-4}$**	*RPS5P5 - ZBTB10*	T	0.36	$4.0x10^{-9}$	1.14 [1.09-1.19]	Asthma & hay fever	Ferreira, J Allergy Clin Immunol, 2014[28]	European
8	rs3019885	118.03	.	G/T	0.48	1.02 [0.98-1.06]	0.40	*SLC30A8*	G	0.31	$5.0x10^{-13}$	1.34 [1.24-1.45]	Asthma	Noguchi, PLoS Genet, 2011[34]	Japanese
9	rs72699186	6.18	.	NA	NA	NA	NA	*RANBP6 - IL33*	T	0.15	$2.0x10^{-9}$	1.26 [1.16-1.35]	Asthma & hay fever	Ferreira, J Allergy Clin Immunol, 2014[28]	European
9	rs1342326	6.19	.	A/C	0.79	0.84 [0.8-0.88]	**$1.6x10^{-12}$**	*RANBP6 - IL33*	C	0.16	$9.0x10^{-10}$	1.2 [1.13-1.28]	Asthma	Moffatt, N Engl J Med, 2010[1]	European
9	rs2381416	6.19	.	A/C	0.68	0.85 [0.81-0.89]	**$3.6x10^{-14}$**	*RANBP6 - IL33*	NA	NA	$2.0x10^{-12}$	NA	Asthma	Torgerson, Nat Genet, 2011[25]	Multi-ethnic
9	rs928413	6.21	.	A/G	0.70	0.84 [0.8-0.88]	**$6.5x10^{-16}$**	*IL33*	G	0.28	$9.0x10^{-13}$	1.24 [1.17-1.32]	Childhood severe Asthma	Bonnelykke, Nat Genet, 2014[35]	European
9	rs16929097	12.52	.	A/G	0.04	1.04 [0.91-1.19]	0.55	*PTPRD - TYRP1*	NA	NA	$8.0x10^{-9}$	NA	Childhood Asthma	Ding, Hum Genomics, 2013[30]	European
10	rs7915695	68.44	*CTNNA3*	NA	NA	NA	NA	*CTNNA3*	C	0.09	$2.2x10^{-8}$	NA	Asthma exacerbations	McGeachie, J Allergy Clin Immunol,2015[36]	European
10	rs12570188	100.86	*HPSE2*	NA	NA	NA	NA	*HPSE2*	NA	NA	$5.0x10^{-8}$	NA	Childhood Asthma	Ding, Hum Genomics, 2013[30]	European

				Meta-analysis of time-to-asthma onset GWASs				GW results reported in GWAS for asthma and asthma-related traits NCBI GWAS Catalog, June 2015							
Chr	**SNP or proxy***	**Pos (Mb)**	**Gene Symbol**	**Effect/Ref Alleles†**	**Effect Freq**	**HR [95% CI]**	**P‡**	**Mapped Genes**	**Effect Allele**	**Effect Freq in ctrls**	**P**	**OR [95% CI]**	**Trait**	**References**	**Pop**
10	rs10508372	8.97	.	A/G	0.08	0.95 [0.88-1.02]	0.16	*KRT8P16 - TCEB1P3*	C	0.433	$2.0x10^{-15}$	1.16 [1.12-1.21]	Asthma	Hirota, Nat Genet, 2011[31]	Japanese
11	rs7130588	76.27	.	A/G	0.63	0.95 [0.91-0.99]	$9.2x10^{-3}$	*C11orf30 - LRRC32*	G	0.34	$2.0x10^{-8}$	1.09 [1.06-1.13]	Asthma	Ferreira, Lancet, 2011[24]	European
11	rs215521	76.59		G/T	0.51	0.93 [0.90-0.97]	$7.6x10^{-4}$	*C11orf30 - LRRC32*	T	0.48	$5.0x10^{-11}$	1.16 [1.11-1.21]	Asthma & hay fever	Ferreira, J Allergy Clin Immunol, 2014[28]	European
11	rs7927044	127.76	.	A/G	0.01	1.09 [0.89-1.33]	0.43	*NCRNA00288 - ETS1*	A	0.0134	$7.0x10^{-9}$	0.85 [NA]	Childhood Asthma	Forno, J Allergy Clin Immunol, 2012[29]	European
12	rs2069408	56.36	*CDK2*	A/G	0.66	0.98 [0.94-1.02]	0.31	*CDK2*	C	0.23	$1.0x10^{-10}$	1.15 [1.10-1.20]	Asthma	Hirota, Nat Genet, 2011[31]	Japanese
12	rs1701704	56.41	.	G/T	0.35	1.03 [0.99-1.08]	0.10	*SUOx - IKZF4*	G	0.18	$2.0x10^{-13}$	1.19 [1.14-1.25]	Asthma	Hirota, Nat Genet, 2011[31]	Japanese
15	rs744910	67.45	*SMAD3*	A/G	0.49	0.93 [0.9-0.97]	**$3.2x10^{-4}$**	*SMAD3*	G	0.49	$4.0x10^{-9}$	1.12 [1.09-1.16]	Asthma	Moffatt, N Engl J Med, 2010[1]	European
15	rs17294280	67.47	*SMAD3*	A/G	0.71	0.93 [0.88-0.98]	$4.9x10^{-3}$	*SMAD3*	G	0.23	$4.0x10^{-9}$	1.18 [1.11-1.25]	Asthma & hay fever	Ferreira, J Allergy Clin Immunol, 2014[28]	European
16	rs62026376	11.23	*CLEC16A*	NA	NA	NA	NA	*CLEC16A*	C	0.72	$1.0x10^{-8}$	1.17	Asthma & hay fever	Ferreira, J Allergy Clin Immunol, 2014[28]	European

				Meta-analysis of time-to-asthma onset GWASs				GW results reported in GWAS for asthma and asthma-related traits NCBI GWAS Catalog, June 2015							
Chr	**SNP or proxy***	**Pos (Mb)**	**Gene Symbol**	**Effect/Ref Alleles†**	**Effect Freq**	**HR [95% CI]**	**P‡**	**Mapped Genes**	**Effect Allele**	**Effect Freq in ctrls**	**P**	**OR [95% CI]**	**Trait**	**References**	**Pop**
												[1.11-1.24]			
17	rs2305480	38.06	*GSDMB*	A/G	0.41	0.85 [0.82-0.88]	**$8.1x10^{-16}$**	*GSDMB*	G	0.60	$6.0x10^{-23}$	1.32 [1.23-1.39]	Childhood severe Asthma	Bonnelykke, Nat Genet, 2014[35]	European
				A/G	0.41	0.85 [0.82-0.88]	**$8.1x10^{-16}$**	*GSDMB*	A	0.45		0.85 [0.81-0.90]	Asthma	Moffatt, N Engl J Med, 2010[1]	European
17	rs11078927	38.06	*GSDMB*	C/T	0.59	1.18 [1.13-1.22]	**$6.8x10^{-16}$**	*GSDMB*	NA	NA	$2.0x10^{-16}$	NA	Asthma	Torgerson, Nat Genet, 2011[25]	Multi-ethnic
17	rs7216389	38.07	*GSDMB*	C/T	0.46	0.86 [0.81-0.91]	**$3.1x10^{-8}$**	*GSDMB*	T	0.52	$9.0x10^{-11}$	1.45 [1.17-1.81]	Asthma	Moffatt, Nature, 2007[37]	European
17	rs4794820	38.09	.	A/G	0.40	0.86 [0.82-0.89]	**$1.0x10^{-13}$**	*ORMDL3 - GSDMA*	NA	NA	$1.0x10^{-8}$	1.33 [1.20-1.45]	Asthma	Wan, Thorax, 2012[38]	European
17	rs3894194	38.12	*GSDMA*	A/G	0.49	1.16 [1.11-1.2]	**$1.4x10^{-13}$**	*GSDMA*	A	0.45	$5.0x10^{-9}$	1.17 [1.11-1.23]	Asthma	Moffatt, N Engl J Med, 2010[1]	European
				A/G	0.49	1.16 [1.11-1.2]	**$1.4x10^{-13}$**	*GSDMA*	A	NA	$3.0x10^{-21}$	1.59 [1.44-1.76]	Childhood severe Asthma	Bonnelykke, Nat Genet, 2014[35]	European
17	rs7212938	38.12	*GSDMA*	G/T	0.50	1.18 [1.13-1.23]	**$1.1x10^{-15}$**	*GSDMA*	G	0.46	$4.0x10^{-10}$	1.16 [1.11-1.20]	Asthma & hay fever	Ferreira, J Allergy Clin Immunol, 2014[28]	European

				Meta-analysis of time-to-asthma onset GWASs				GW results reported in GWAS for asthma and asthma-related traits NCBI GWAS Catalog, June 2015							
Chr	**SNP or proxy***	**Pos (Mb)**	**Gene Symbol**	**Effect/Ref Alleles†**	**Effect Freq**	**HR [95% CI]**	**P‡**	**Mapped Genes**	**Effect Allele**	**Effect Freq in ctrls**	**P**	**OR [95% CI]**	**Trait**	**References**	**Pop**
22	rs2284033	37.53	*IL2RB*	A/G	0.41	0.94 [0.9-0.98]	$2.1x10^{-3}$	*IL2RB*	G	0.56	$1.0x10^{-8}$	1.12 [1.08-1.16]	Asthma	Moffatt, N Engl J Med, 2010[1]	European

* The SNP with the strongest LD with the reported SNP in the literature was used if the SNP reported in the literature was not available in the imputed data

†For the calculation of the hazard and odds ratios, Effect alleles (Effect) were designated as risk alleles. Effect Freq denotes frequency of the effect allele, CI confidence interval, and Ref reference allele

‡P-values are shown for tests of association under a fixed-effect model when there was no significant evidence of heterogeneity or under a random-effect model otherwise.

Associations with P-values ≤ 0.001 in the time-to-asthma onset meta-analysis are indicated in bold

NA: Not available

FIGURE LEGENDS

Figure E1. Distribution of age-of-asthma onset

Figure E2. Quantile-quantile (QQ) plots of 2,387,926 SNPs of nine GWAS (N = 13,886) after quality control (Rsq≥0.50, MAF≥0.01, ≥6 contributing studies) under a fixed-effect model (inflation factor, λ_{GC}=1.04). The dots represent the distribution of observed Chi-Square values against the theoretical model distribution of expected Chi-Square values. The red line represents the theoretical model distribution of expected Chi-Square values under the null distribution.

Figure E3. Regional association plots of the genome-wide associated regions using Locuszoom[39] software: 2q12, 6p21, 9p24, 17q12-q21. SNPs are plotted with their P-values (-$\log_{10}$ values, left y-axis) as a function of genomic position (x-axis). Estimated recombination rates (right y-axis) taken from 1000 Genomes (March 2012 EUR) are plotted in cyan to reflect the local LD structure. SNPs surrounding the most significant SNP (purple circle) are color-coded according to LD with lead SNP (pairwise r2, according to a blue to red scale from r^2=0 to 1). In 9p24 and 17q12-q21 regions, additional SNP detected by conditional analyses is indicated by an arrow (Part A). Two additional plots show SNPs color-coded according to LD with additional SNP detected in conditional analysis (Part B).

Figure E4. Forest plots of hazard ratios for SNPs associated with time-to-asthma onset at genome-wide significant level ($P\leq5x10^{-8}$) and two additional SNPs detected by conditional analyses in 9p24 and 17q12-q21 regions. The hazard ratios and 95% confidence intervals for seven loci show distinct effect on time-to-asthma onset. In each plot, the diamond indicates the effect size and the 95% CI derived from the meta-analysis of nine studies.

Figure E5. Regional plots of 9p24 (Part-a) and 17q12-q21 (Part-b) regions for distinct association signals using sequential conditional analysis and time-to-asthma onset as an outcome variable: original meta-analysis (A), adjusted for lead SNP (B) and additionally adjusted for the secondary signal (C). Signals above the red line ($P<10^{-5}$) were considered to

exhibit evidence of association in the regions. SNPs are colored according to their pairwise LD r^2 with the lead SNP. r^2 was estimated from 1000 Genomes (March 2012 EUR).

Figure E6. Association plots of the fine-mapping conducted in 16q12 region using 1000G CEU reference sample. SNPs are plotted with their P-values ($-\log_{10}$ values, left y-axis) as a function of genomic position (x-axis). Estimated recombination rates (right y-axis) taken from 1000G are plotted to reflect the local LD structure around the top associated SNP (purple circle) and correlated proxies (according to a blue to red scale from $r^2 = 0$ to 1).

Figure E1.

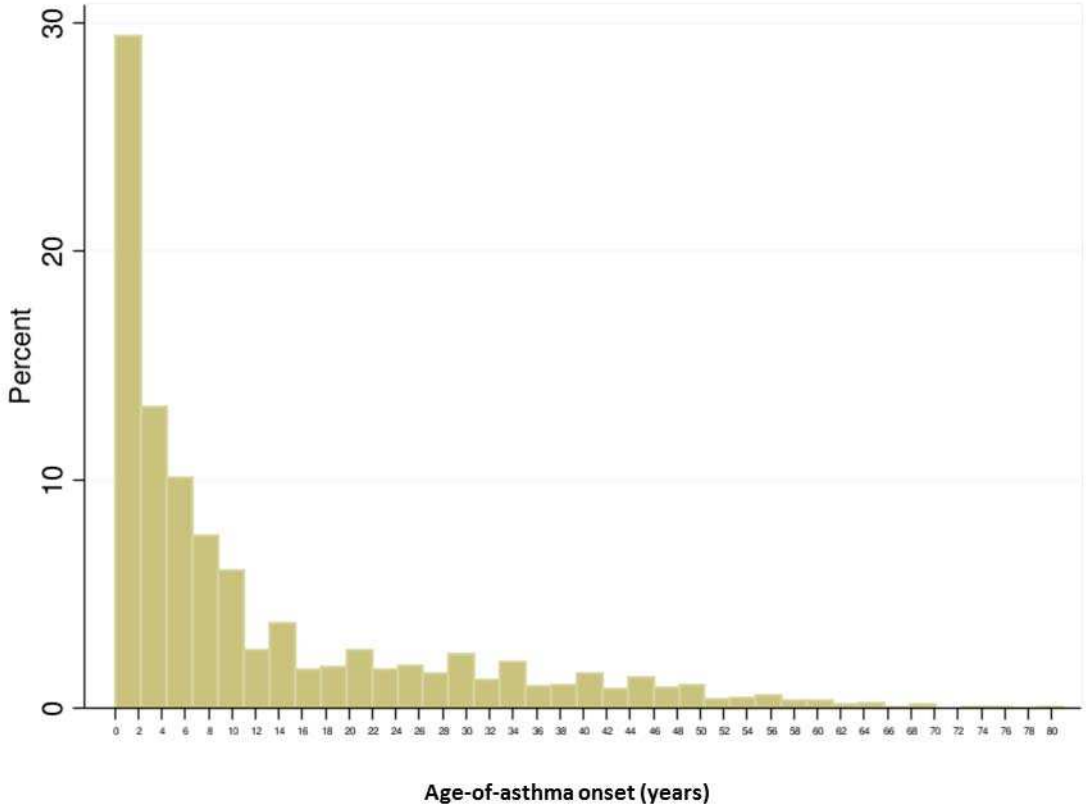

Percent
0
10
20
30
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80
Age-of-asthma onset (years)

Figure E2.

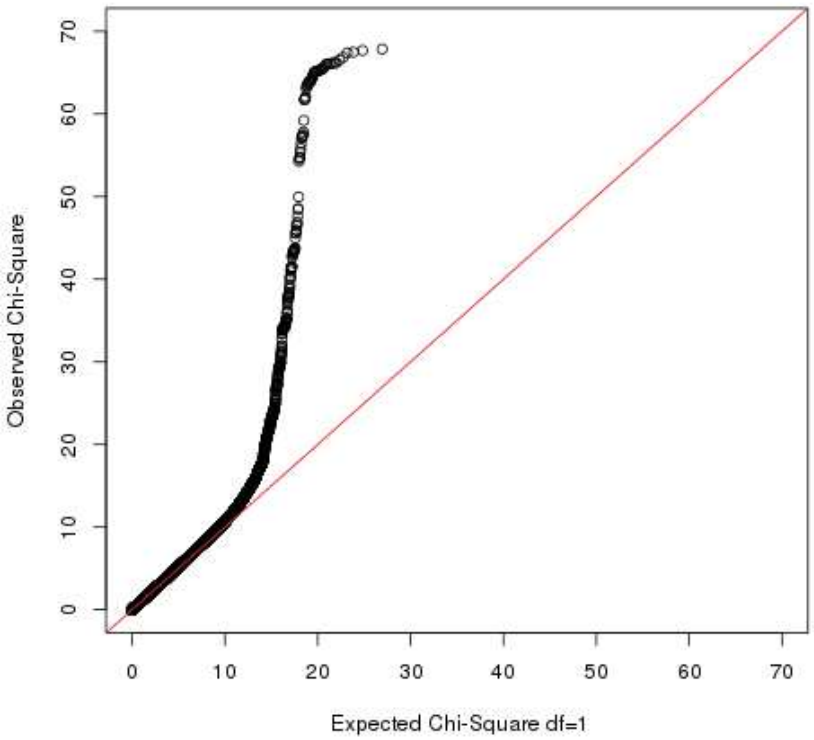

Observed Chi-Square
Expected Chi-Square df=1
0
10
20
30
40
50
60
70

Figure E3.

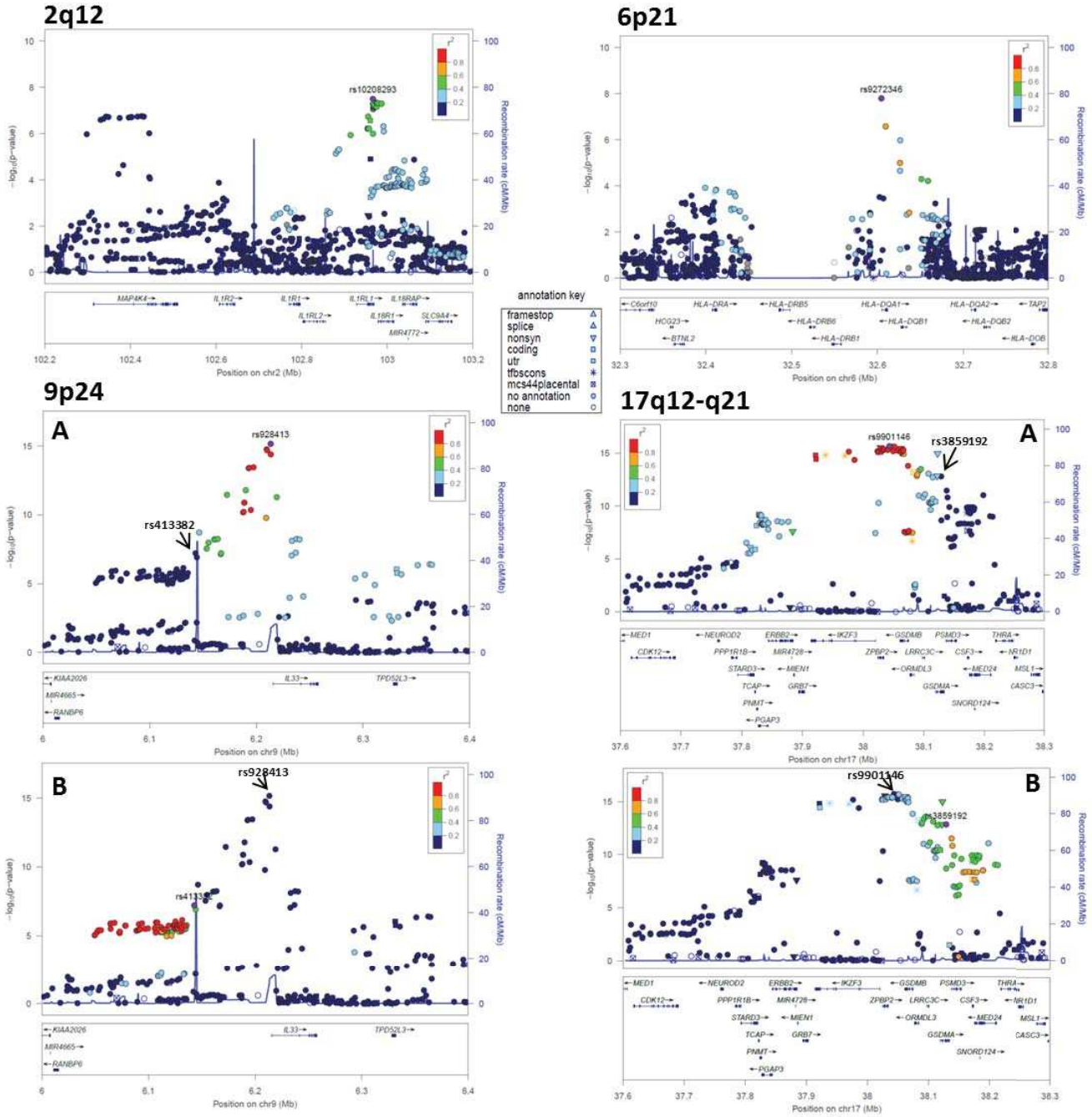
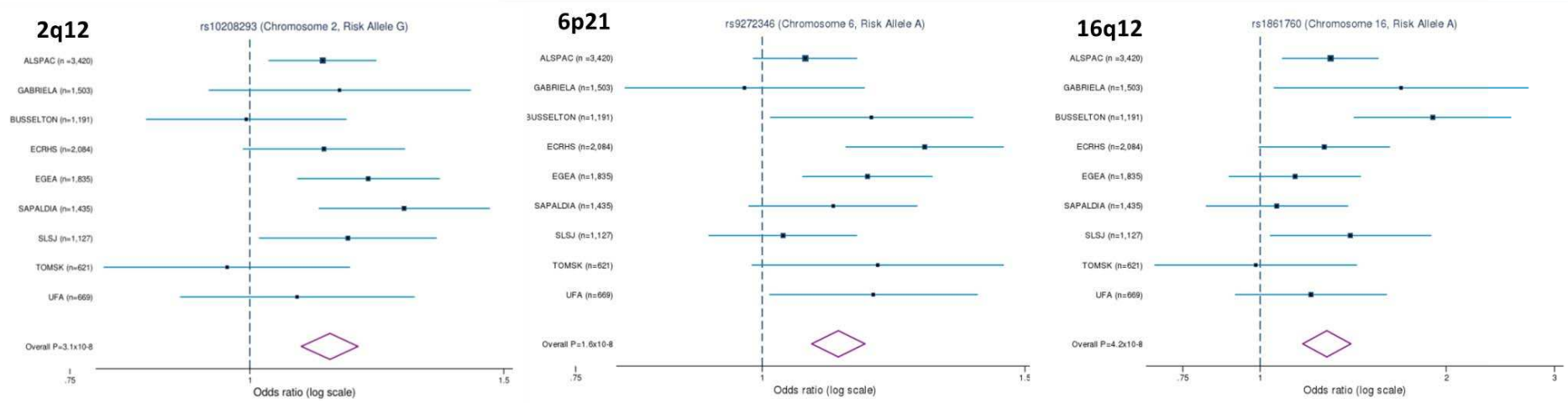
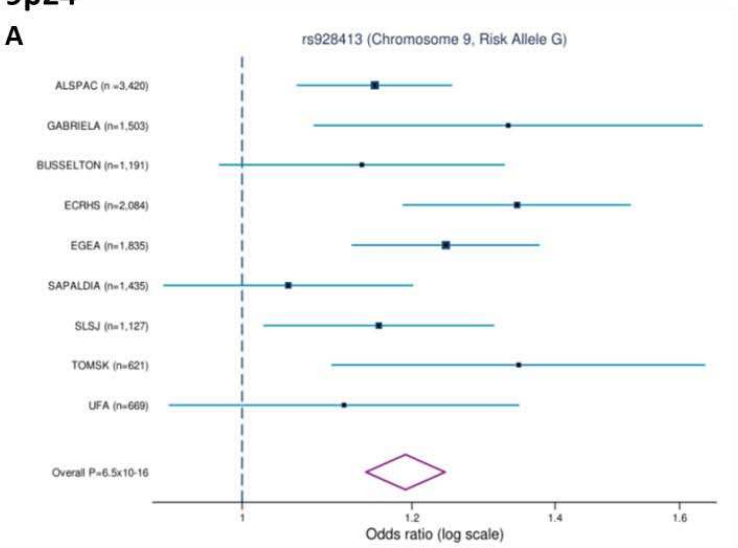

2q12
rs10208293
6p21
rs9272346
annotation key
framestop
splice
nonsyn
coding
utr
tfbscons
mcs44placental
no annotation
none
9p24
A
rs928413
rs413382
B
rs928413
17q12-q21
A
rs9901146
rs3859192
B
rs9901146
rs3859192
Position on chr2 (Mb)
Position on chr6 (Mb)
Position on chr9 (Mb)
Position on chr17 (Mb)
Recombination rate (cM/Mb)

Figure E4.

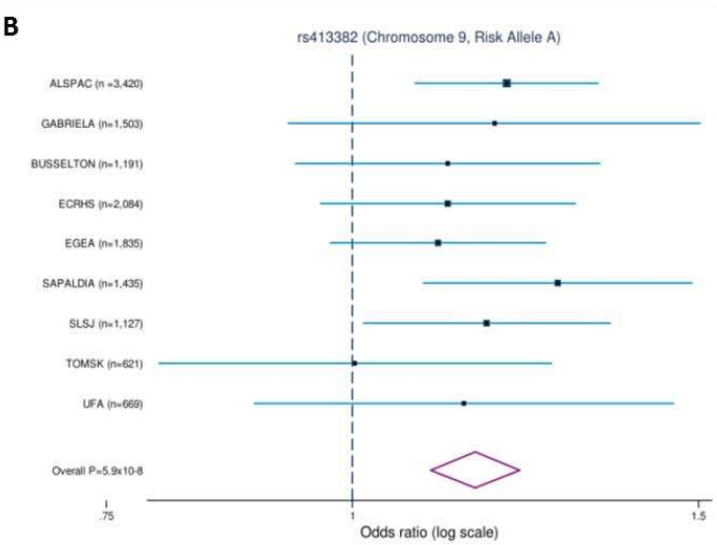

2q12
rs10208293 (Chromosome 2, Risk Allele G)
ALSPAC (n =3,420)
GABRIELA (n=1,503)
BUSSELTON (n=1,191)
ECRHS (n=2,084)
EGEA (n=1,835)
SAPALDIA (n=1,435)
SLSJ (n=1,127)
TOMSK (n=621)
UFA (n=669)
Overall P=3.1x10-8
.75
1
1.5
Odds ratio (log scale)
6p21
rs9272346 (Chromosome 6, Risk Allele A)
ALSPAC (n =3,420)
GABRIELA (n=1,503)
BUSSELTON (n=1,191)
ECRHS (n=2,084)
EGEA (n=1,835)
SAPALDIA (n=1,435)
SLSJ (n=1,127)
TOMSK (n=621)
UFA (n=669)
Overall P=1.6x10-8
.75
1
Odds ratio (log scale)
16q12
rs1861760 (Chromosome 16, Risk Allele A)
ALSPAC (n =3,420)
GABRIELA (n=1,503)
BUSSELTON (n=1,191)
ECRHS (n=2,084)
EGEA (n=1,835)
SAPALDIA (n=1,435)
SLSJ (n=1,127)
TOMSK (n=621)
UFA (n=669)
Overall P=4.2x10-8
.75
1
2
3
Odds ratio (log scale)

9p24

A


rs928413 (Chromosome 9, Risk Allele G)
ALSPAC (n =3,420)
GABRIELA (n=1,503)
BUSSELTON (n=1,191)
ECRHS (n=2,084)
EGEA (n=1,835)
SAPALDIA (n=1,435)
SLSJ (n=1,127)
TOMSK (n=621)
UFA (n=669)
Overall P=6.5x10-16
1
1.2
1.4
1.6
Odds ratio (log scale)

B


rs413382 (Chromosome 9, Risk Allele A)
ALSPAC (n =3,420)
GABRIELA (n=1,503)
BUSSELTON (n=1,191)
ECRHS (n=2,084)
EGEA (n=1,835)
SAPALDIA (n=1,435)
SLSJ (n=1,127)
TOMSK (n=621)
UFA (n=669)
Overall P=5.9x10-8
.75
1
1.5
Odds ratio (log scale)

17q12-q21

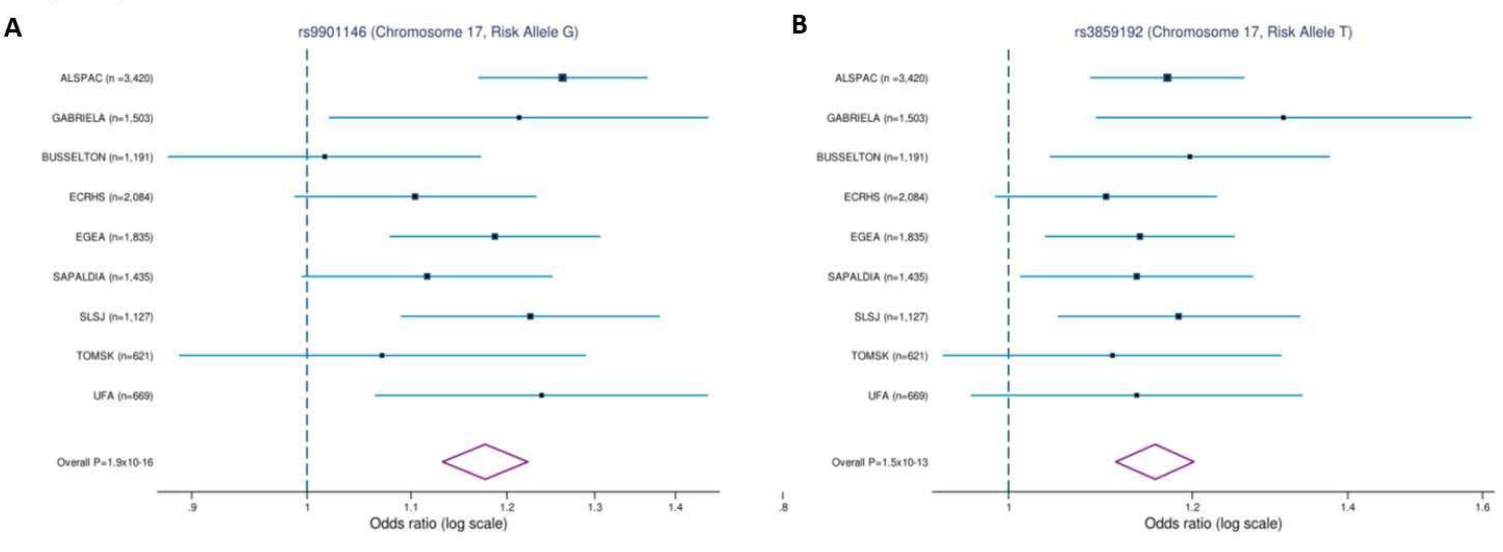

A
rs9901146 (Chromosome 17, Risk Allele G)
ALSPAC (n =3,420)
GABRIELA (n=1,503)
BUSSELTON (n=1,191)
ECRHS (n=2,084)
EGEA (n=1,835)
SAPALDIA (n=1,435)
SLSJ (n=1,127)
TOMSK (n=621)
UFA (n=669)
Overall P=1.9x10-16
.9
1
1.1
1.2
1.3
1.4
Odds ratio (log scale)
B
rs3859192 (Chromosome 17, Risk Allele T)
ALSPAC (n =3,420)
GABRIELA (n=1,503)
BUSSELTON (n=1,191)
ECRHS (n=2,084)
EGEA (n=1,835)
SAPALDIA (n=1,435)
SLSJ (n=1,127)
TOMSK (n=621)
UFA (n=669)
Overall P=1.5x10-13
.8
1
1.2
1.4
1.6
Odds ratio (log scale)

Figure E5.

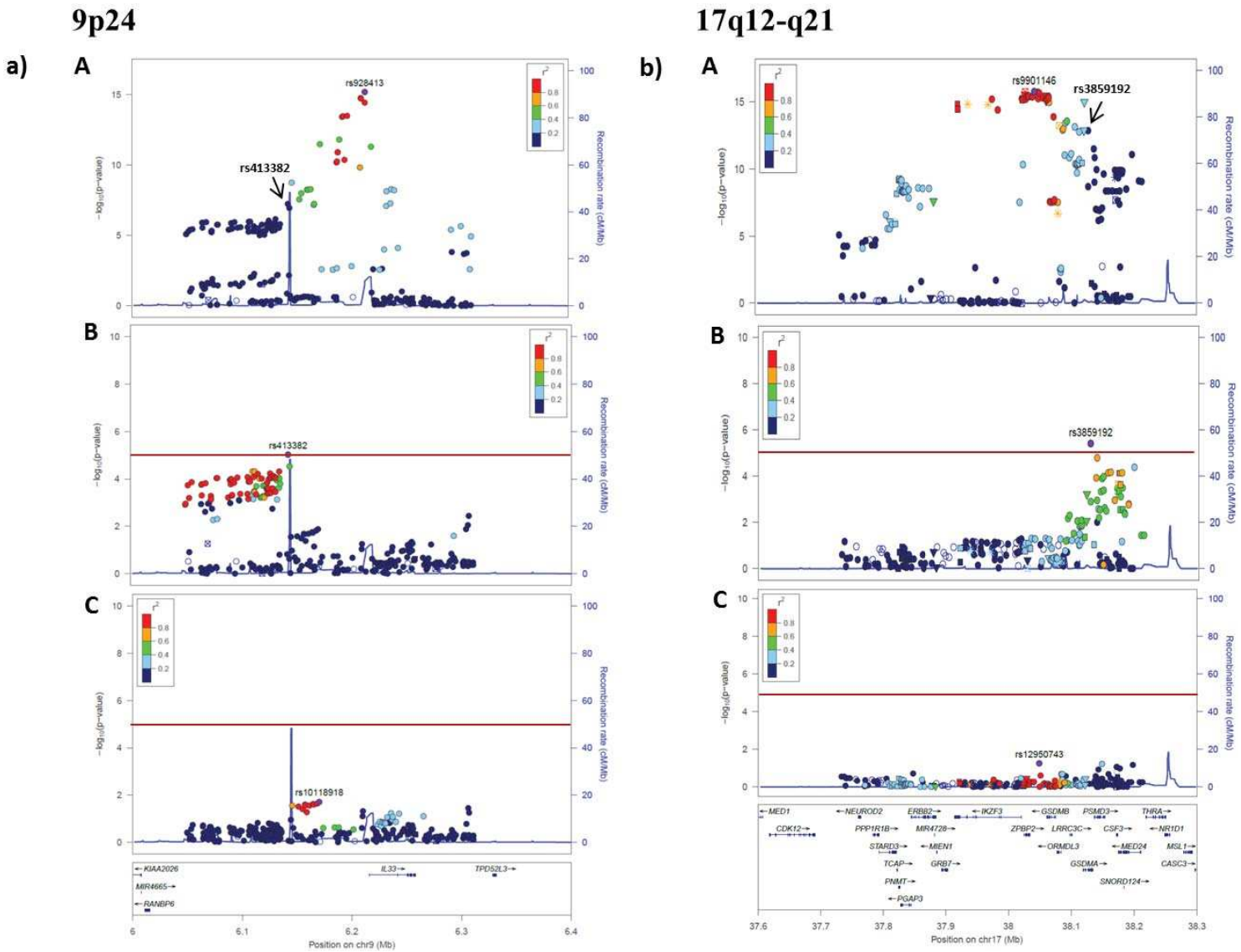
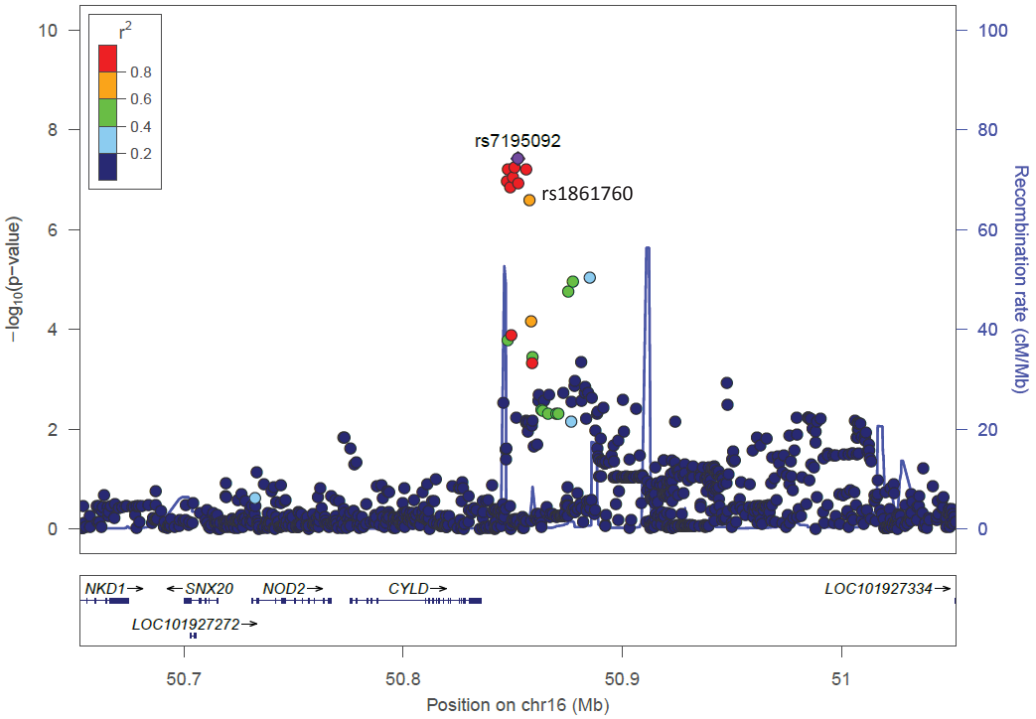

9p24
17q12-q21
a)
b)
A
B
C
rs928413
rs413382
rs10118918
rs9901146
rs3859192
rs12950743
Recombination rate (cM/Mb)
Position on chr9 (Mb)
Position on chr17 (Mb)
KIAA2026
MIR4665
RANBP6
IL33
TPD52L3
MED1
CDK12
NEUROD2
PPP1R1B
STARD3
TCAP
PNMT
PGAP3
ERBB2
MIR4728
MIEN1
GRB7
IKZF3
ZPBP2
GSDMB
LRRC3C
ORMDL3
GSDMA
PSMD3
CSF3
MED24
SNORD124
THRA
NR1D1
MSL1
CASC3

Figure E6.


r²
0.8
0.6
0.4
0.2
rs7195092
rs1861760
−log10(p−value)
Recombination rate (cM/Mb)
NKD1→
←SNX20
NOD2→
CYLD→
LOC101927334→
LOC101927272→
50.7
50.8
50.9
51
Position on chr16 (Mb)