Les interactions gène-environnement dans les études génétiques des maladies complexes

Rémi Kazma

UNIVERSITÉ PARIS-SUD XI
FACULTÉ DE MÉDECINE
ÉCOLE DOCTORALE 420 « SANTÉ PUBLIQUE »

Thèse de Doctorat en Santé Publique
Spécialité Génétique Statistique

Les interactions gène-environnement dans les études génétiques des maladies complexes

présentée par
Rémi KAZMA

soutenue publiquement le
jeudi 17 juin 2010

sous la direction de
Emmanuelle GÉNIN

réalisée au sein de l'unité
Inserm U946
Variabilité Génétique et Maladies Humaines

Composition du jury

Monsieur Denis HÉMON	Président
Monsieur Jacques BÉNICHOU	Rapporteur
Monsieur Bertram MÜLLER-MYHSOK	Rapporteur
Monsieur André GARCIA	Examinateur
Madame Núria MALATS	Examinatrice
Madame Emmanuelle GÉNIN	Directrice

A little knowledge is a dangerous thing,
So is a lot.

Albert Einstein

À Mimat et Lina,
pour tout l'amour qu'elles m'ont donné.

À Kaia et Mick, à l'avenir,
et à tous ceux des générations qui suivent.

À la Famille,
source intarissable de questions
... et parfois de réponses.

De Tampa Bay à Waikiki Beach, des alligators paresseux aux tortues brouteuses, du Yellow Tail australien au Faux Rhum hawaïen, quatre années riches de formation, mais pas que … Merci à tous ceux qui m'ont accompagné et soutenu tout le long du chemin.

L'itinéraire de cette thèse, je l'ai emprunté entre autres car il prévoyait un guide de qualité, Emmanuelle Génin. Son enthousiasme, sa patience et sa perspicacité ont été les ingrédients essentiels à la concrétisation de ce périple. Aujourd'hui s'achève ma formation sous son aile et c'est grandi d'une richesse tant scientifique qu'humaine que je m'envole vers d'autres cieux. Merci, Emmanuelle, pour tout ce que tu m'as appris.

Même si l'on n'y partage pas les proportions requises d'ADN, la génétique statistique est une grande famille où l'accueil est toujours chaleureux. Les génotypes de la convivialité, du respect et du partage y sont homozygotes et chacun peut y trouver sa place et y tisser ses liens.

Catherine m'en a ouvert les portes, en me proposant mon sujet de stage de Master, puis de nombreuses autres collaborations. Sa rigueur scientifique et sa sagacité sont des qualités que j'admire chez elle. Ce fut un plaisir d'hériter du flambeau … EURECA, nous avons trouvé !

En m'accueillant dans son laboratoire, Françoise m'a permis d'y apprendre les fondamentaux du travail de paillasse : nos bactéries sont fromagères et notre éthanol est viticole ! L'un ne va pas sans l'autre et le carré cacaoté est toujours de rigueur. C'est ensuite chez Florence que j'ai pu poursuivre ces TP « vie de famille ». Je vous remercie toutes deux de m'avoir offert l'environnement spatial et humain propice à la réalisation de ces travaux-ci et à celle de ma thèse aussi.

Porteuse de la double mutation « black inside », elle ferme toujours les yeux en riant, manie aussi bien le dico que l'extincteur et a la prostate lycopénifiée. Elle cumule avec brio les casquettes de GI, CoF, Soprano et Mamie d'Unité. Son sport de prédilection est la génuflexion-enfonçage de prise de courant. Avec son humour décoiffant et décapant, elle tourne les situations les plus sérieuses en moments burlesques. Il est parfois des personnes dont la simple présence apaise et tu en fais partie.
Merci, Marie-Claude pour ton encadrement (!) et ta clairvoyance si souvent salvatrice.

Juger du travail d'autrui est une responsabilité de la plus haute importance. Nous pouvons, Emmanuelle et moi, nous enorgueillir d'avoir sélectionné pour cela une équipe de choc ! Avec Denis Hémon aux commandes, Núria Malats, Jacques Bénichou, André Garcia et Bertram Müller-Myhsok pour l'entourer, mes travaux sont ainsi soumis au regard d'experts internationaux et impartiaux.
Je vous suis reconnaissant d'avoir accepté de faire partie de mon jury de thèse.

« *On n'enseigne pas ce que l'on sait ou ce que l'on croit savoir :*
On n'enseigne et on ne peut enseigner que ce que l'on est. »
J'ai toujours ressenti de l'admiration pour mes enseignants, et aujourd'hui je les remercie toutes et tous de ce qu'ils m'ont apporté. C'est à Henri Helou que je pense tout particulièrement car il fut le premier à semer en moi le goût de la génétique. L'adolescent boutonneux que j'étais ne savait encore pas jusqu'où cela l'emmènerait. Merci Monsieur Helou.
Ce fut ensuite mon tour de transmettre un peu de moi aux étudiants du Master de Santé Publique. Et quel plaisir ! Ils ne s'en doutent pas, mais eux m'ont bien plus appris encore. Merci.

Ma Sara, l'amitié requiert du temps et pourtant toi, j'ai su qui tu étais dès le premier instant. Du rire aux larmes … personne ne sait autant que toi ce qu'il y a d'invisible entre ces lignes de thèse, et d'ailleurs celles-ci n'existeraient probablement pas si tu n'avais pas été présente. Mon Pedro, merci pour ton amitié sans faille. Je compte sur toi pour veiller sur cette perle.
Kiss Kiss Bang Bang Luvs.

Merci aussi à …

Simone, avec qui c'est un réel plaisir de travailler. Ce projet m'a permis d'apprendre à utiliser de nouveaux outils et d'appréhender la génétique sous une perspective plus appliquée.

Hervé, parce que derrière un ordinateur, un bouquin ou un fourneau, tu es un dieu !

Pascal, pour m'avoir laissé tes traces de pas sur le chemin. Je t'admire pour ta constante bonne humeur et ton positivisme à toute épreuve. C'est un plaisir de t'avoir pour grand frère.
Tu vois que je peux placer trois phrases sans nécessairement parler de ton hépatite psychosomalcoolique à Naples (Florida) en novembre 2006.

Gaëlle, ma formidable petite sœur qui m'a fait danser jusqu'au bout de la nuit sur des rythmes endiablés ! Tu as rendu ma semaine madrilène inoubliable. Muchas gracias también para Gus por las lecciones de gramática española: ratas, ratones, fresas, fresones, cojas … Oh!

Flora, une nana chouette comme y'en a pas des masses dans la vraie vie.

Céline, ma fermière virtuelle préférée. Merci pour toutes les briques (- ;

Emmanuelle, pour toutes les fois où nous nous sommes complétés les phrases l'un de l'autre.

Hélène, pour m'avoir enfin éveillé à l'art suprême de la découpe du melon.

Anne-Louise, pour m'avoir poussé vers l'avant lorsque je reculais.

Marie-Hélène, pour les thés, les chocolats, et ton implication dans la relecture d'une lettre.

Michel, parce que des crèmes comme toi, il n'y en a plus en stock. Et que ça c'est le drame !

Ève, pour tous les scripts R et Stata qui tombaient à pic et pour toutes les fois que j'ai trouvé réponse à ma question sur le chemin vers ton bureau.

Hamida, pour ta générosité tant en espace disque qu'en friandises.

Christine et Martine, pour vos sourires … et aussi parce que la recherche française serait foutue sans votre capacité à parler couramment le SAFIR.

Un peu de chaque Piou Piou do Brazil se trouve dans ces pages. Vous ne les voyez pas car pas plus le Tsipouro que la Cachaça ne tache. Merci à Inês, Dimitris, Valentine, Marc, Sophie, Loïc, Hélène, Lolo, Lika, Jako, Réjane, François, Sylvie et les Yolandes pour les petits bonheurs du mardi de 22h à minuit.

Une pensée aussi à …

Celle et ceux qui m'ont tant appris sur moi-même.

Catherine qui garde quoi qu'il advienne sa place au chaud dans mon cœur.

Roula et Samar qui lisent en moi comme dans un livre ouvert.

Antoine, parce que tu comptes comme un frère pour moi.

Charbel et Patrice parce que des amis comme ça, c'est précieux.

Claire, en souvenir de nos échanges qui ont tant fait avancer le schmilblick.

Et aussi aux agriculteurs corses qui m'ont fourni mes doses de vitamine C pendant l'hiver rigoureux.

L'argent ne fait pas le bonheur, mais il paraît qu'il y contribue. C'est pourquoi, je remercie également la Présidence de l'Université Paris-Sud XI, la Fondation pour la Recherche Médicale et l'Inserm pour leur part de contribution à mon bonheur durant ces années de thèse.

À ceux qui pensent que Paul Brousse à Villejuif, c'est le trou du cul du monde, je réponds qu'où que je sois je penserai à mon vieux cèdre ridé trônant derrière ma fenêtre et agitant ses branches.
À ceux qui pensent que le CEPH c'est un vieux bâtiment déglingué et mal agencé, je réponds qu'où que je sois, je penserai à mon petit couple d'amoureux déjeunant sur leur balcon au soleil.
Je n'oublierai jamais ces lieux et les gens que j'y ai côtoyé, tout cela est gravé dans mon cœur et ça, c'est bien plus important que toutes les bases nucléiques de nos génomes.

Finalement, merci aussi à Juanito, Juanita, Blanche Neige, les sept nains et les sept cent millions de chinois qui m'ont permis de surmonter ma timidité !

Rawi

Résumé

Les maladies humaines les plus fréquentes sont complexes avec plusieurs facteurs génétiques et environnementaux qui interagissent. Ce travail propose deux nouvelles méthodes statistiques pour étudier les interactions gène-environnement. La première méthode utilise la récurrence familiale de la maladie pour identifier une interaction entre un facteur environnemental et la composante génétique impliquée dans la maladie. La seconde méthode permet de prendre en compte ces interactions dans les études d'associations pangénomiques lorsque l'information sur le facteur d'exposition n'est pas disponible chez les témoins. Cette situation est devenue fréquente avec l'utilisation de panels de témoins de référence. Ces deux méthodes apportent de nouveaux outils pour étudier simultanément les facteurs génétiques et environnementaux dans les maladies complexes. Elles ont été appliquées sur deux jeux de données concernant le diabète de type 2 et les réactions cutanées sévères aux médicaments.

Mots-clefs

interaction gène-environnement, génétique statistique, développement méthodologique, puissance statistique, biais de confusion, agrégation familiale, régression logistique multinomiale, étude d'association pangénomique, témoins de référence, diabète de type 2, Réactions cutanées sévères secondaire à un médicament.

Coordonnées du laboratoire d'accueil

Inserm UMR_S946 – Variabilité Génétique et Maladies Humaines
Fondation Jean Dausset – CEPH
27 rue Juliette Dodu
75010 Paris
FRANCE

E-mail

remi.kazma@gmail.com

Title

Gene-environment interactions in genetic studies of complex diseases

Abstract

Most common human diseases are complex with multiple genetic and environmental factors that can interact. This doctoral work proposes two new statistical methods to study gene-environment interactions. The first method uses information on the familial recurrence of the disease to identify an interaction between an environmental factor and the genetic component involved in the disease. The second method allows accounting for gene-environment interactions in genome-wide association studies when the exposure information is not available for controls. This situation has become more frequent with the use of reference control panels in those studies. These two methods bring new tools to study simultaneously the genetic and the environmental factors involved in complex diseases. Two examples of how they can be used are provided that concern type 2 diabetes and severe cutaneous adverse drug reactions.

Keywords

gene-environment interaction, statistical genetics, methodological development, statistical power, confusion bias, familial aggregation, multinomial logistic regression, genome-wide association study, reference controls, type 2 diabetes, severe cutaneous adverse drug reactions.

Contact information of the hosting lab

Inserm UMR_S946 – Genetic Variability and Human Diseases
Fondation Jean Dausset – CEPH
27 rue Juliette Dodu
75010 Paris
FRANCE

E-mail

remi.kazma@gmail.com

PRODUCTION SCIENTIFIQUE

Articles publiés (3)

Kazma R, Bonaïti-Pellié C, Norris JM, Génin E. On the use of sibling recurrence risks to select environmental factors liable to interact with genetic risk factors. *Eur J Hum Genet* 2010;18(1):88-94

Kazma R, Dizier MH, Guilloud-Bataille M, Bonaïti-Pellié C, Génin E. Power comparison of different methods to detect genetic effects and gene-environment interactions. *BMC Proc* 2007;1 Suppl 1:S74

Culverhouse RC, Suarez BK, Beckmann L, Chen P, Chen YS, Chiu YF, Chang-Claude J, Dempfle A, Hein R, Kazma R, Lebrec JJ, Lee S, Lim S, Maher BS, Park T, Perdry H, Wang KS, Wolkow PP, Xu W. Gene by environment interactions. *Genet Epidemiol* 2007;31 Suppl 1:S68-74

Article en préparation (1)

Kazma R, Babron MC, Génin E. Gene-environment interactions and genetic association: A new method to overcome the lack of environmental information in controls. *Am J Epidemiol*, en révision

Communications orales (10)

Kazma R, Babron M-C, Génin E. Joint testing of genetic association and gene-environment interaction with reference controls has a new flavour: MInT! Young Researchers in Life Sciences Congress, Institut Pasteur, Paris, France. 7-9 juin 2010.

Kazma R, Babron M-C, Génin E. Interaction gène-environnement et étude d'association : Faut-il vraiment tout savoir sur les témoins ? 5èmes Assises de Génétique Humaine et Médicale, Strasbourg, France. 28-30 janvier 2010. Médecine/Sciences 2010;26(S1):17-18
Prix de la meilleure présentation orale, session « maladie multifactorielle »

Kazma R, Babron M-C, Génin E. How to account for gene-environment interaction when testing for association with a reference control panel? 18th annual meeting of the International Genetic Epidemiology Society. Kahuku, Hawaii, USA. 18-20 octobre 2009
Finaliste du prix Williams des doctorants
Bourse de voyage de l'International Genetic Epidemiology Society

Kazma R, Génin E. Joint test of genetic effect and gene-environment interaction in the setting of a case - population control design. European Mathematical Genetic Meeting, Munich, Germany, 14-15 mai 2009. Ann Hum Genet 2009;73:658-669

Kazma R, Bonaïti-Pellié C, Norris JM, Génin E. Interaction gène-environnement et risque familial. Journée des jeunes chercheurs, Société Française de Biométrie, 5 décembre 2008

Kazma R, Bonaïti-Pellié C, Norris JM, Génin E. Gene-environment interactions involved in type 2 diabetes: application of a genotype-free method. European Mathematical Genetic Meeting, Rotterdam, Netherlands, 10-11 avril 2008. Ann Hum Genet 2008;72: 687-695

Kazma R, Bonaïti-Pellié C, Norris JM, Génin E. Interactions gène-environnement : un nouveau test sans génotypage ! 4èmes Assises de Génétique Humaine et Médicale, Lille, France. 17-19 janvier 2008. Médecine/Sciences 2008;24(S1):30

Kazma R, Bonaïti-Pellié C, Génin E. A new test to detect a possible gene-environment interaction without genotyping! European Mathematical Genetic Meeting, Heidelberg, Germany, 12-13 avril 2007. Ann Hum Genet 2007;71:550-559

Kazma R, Bonaïti-Pellié C, Génin E. Interaction gène-environnement et risque de récurrence familiale : Application au cancer du poumon et au tabac. 17e journée scientifique de la cohorte Gazel, Paris, 7 février 2007

Kazma R (sous la direction de : Génin E, Bonaïti-Pellié C). Interaction gène-environnement et risque de récurrence chez un germain. Résumés des mémoires de Master Recherche. Revue d'Épidémiologie et de Santé Publique 2007;55:151-152

Communications affichées (6)

Kazma R, Génin E, Gaborieau V, Babron M-C, Brennan P, Hung RJ, Krokan H, Metspalu A, Field JK, Lathrop M, Sarasin A, Benhamou S et le consortium ILCCO. Cancer du poumon et gènes de réparation de l'ADN : Association significative avec MSH5 et interaction potentielle entre le tabac et RAD54B. Cinquièmes Assises de Génétique Humaine et Médicale, Strasbourg, France. 28-30 janvier 2010. Médecine/Sciences 2010;26(S1):

Kazma R, Génin E, Gaborieau V, Babron M-C, Brennan P, Hung RJ, Krokan H, Metspalu A, Field JK, Lathrop M, Sarasin A, Benhamou S, the ILCCO consortium. Lung cancer and candidate DNA repair genes: Evidence of significant association with MSH5 and potential interaction between smoking and RAD54B. 59th annual meeting of the American Society of Human Genetics. Honolulu, Hawaii, USA. 20-24 octobre 2009

Kazma R, Bonaïti-Pellié C, Norris JM, Génin E. On the choice of an exposure to test for gene-environment interactions in type 2 diabetes: a new genotype-free method! European Society of Human Genetic, Barcelona, Spain, 31 mai – 3 juin 2008. Eur J Hum Genet 2008;16(S2):307

Kazma R, Bonaïti-Pellié C, Norris JM, Génin E. Detection of gene-environment interaction: a new test based on sibling recurrence risks. 16th annual meeting of the International Genetic Epidemiology Society. Genet Epidemiol 2007;31:633

Kazma R, Bonaïti-Pellié C, Norris JM, Génin E. Detection of gene-environment interaction: a new test based on sibling recurrence risks. European Society of Human Genetics 2007, Nice, 16-19 juin 2007. Eur J Hum Genet 2007;15(S1):291

Kazma R, Dizier MH, Guilloud-Bataille M, Bonaïti-Pellié C, Génin E. Power comparison of different methods to detect genetic effects and gene-environment interaction. Genetic Analysis Workshop 15, St. Pete's Beach, USA, 12-15 novembre 2006

TABLE DES MATIÈRES

Introduction

Si les premières interrogations sur le rôle des facteurs du milieu dans la survenue des maladies humaines datent de la Grèce Antique (Hippocrate 1996), ce n'est qu'en 1801 qu'est apparu le terme d'« épidémiologie » en titre d'un ouvrage sur les épidémies de peste publié à Madrid : *Epidemiología española* (1831). Alors que les premières études épidémiologiques se sont intéressées principalement aux maladies « transmissibles » comme le choléra (Snow 2008), l'attention des épidémiologistes s'est tournée depuis la seconde moitié du *XX*e siècle vers l'étude des maladies chroniques « non transmissibles » par un agent infectieux, telles que le cancer, les maladies cardio-vasculaires, le diabète ou l'asthme, dont les prévalences ont explosé dans les pays développés. Ainsi de nombreux facteurs environnementaux non infectieux ont été étudiés dans ces maladies et l'étude de Doll et Hill (1954) qui met en évidence un lien entre le tabagisme actif et la survenue du cancer du poumon dans une cohorte de médecins britanniques fait figure de pionnière en la matière.

Parallèlement, la redécouverte trente-cinq ans après des travaux de Gregor Mendel (1865) décryptant les lois de l'hérédité de caractères phénotypiques discrets chez des végétaux, puis la découverte et l'étude de la structure de l'ADN (acide désoxyribonucléique), support physique de l'information génétique ont été les éléments fondateurs de la génétique médicale. L'objectif de cette discipline est d'étudier les maladies monogéniques dont le déterminisme est lié à la présence d'une mutation sur un gène causant une altération de la fonction d'une protéine. Ce sont des maladies relativement rares, apparaissant souvent de façon précoce dans la vie et suivant un modèle Mendélien de transmission familiale (Feingold et Martinez 1998).

Cependant la vaste majorité des maladies humaines ne sont ni le résultat de l'effet d'une exposition environnementale isolée, ni celui d'une mutation génétique unique, mais sont le plus souvent dues à de nombreux facteurs de risque génétiques et environnementaux dont les effets peuvent se conjuguer de différentes manières. Dans les années soixante émerge une nouvelle discipline, l'épidémiologie génétique qui va permettre d'intégrer ces deux dimensions dans l'étude des maladies complexes. Le terme « complexe » fait référence ici à tout phénotype dont l'agrégation familiale ne peut être expliquée par un mode de transmission Mendélien d'un seul gène (Ottman 1995; Risch 1990a). L'épidémiologie génétique vise à étudier le rôle des facteurs génétiques et de leurs interactions avec des facteurs de l'environnement dans le déterminisme de la santé et des maladies dans les familles et dans les populations humaines (Khoury *et al.* 1993c). Les progrès technologiques qui ont révolutionné

la biologie moléculaire et la bioinformatique au cours des trente dernières années ont joué un rôle capital dans l'évolution de cette discipline où les nombreuses problématiques méthodologiques nécessitent le développement perpétuel d'outils statistiques adaptés.

Pour de nombreuses maladies le mécanisme d'action des facteurs de risque qui leurs sont associés demeure néanmoins énigmatique. Archibold Garrod notait déjà en 1902 que « les influences de l'alimentation et d'autres maladies » pouvaient « masquer » certaines des « erreurs congénitales du métabolisme » et que « l'idiosyncrasie relative aux médicaments » était probablement reliée à des facteurs héréditaires (Hunter 2005). Reprenant l'exemple de la relation entre le tabac et le cancer du poumon, de nombreuses observations décrivent des personnes ne développant pas de cancer du poumon malgré un tabagisme important et inversement des personnes développant la maladie sans qu'aucune exposition à ce facteur de risque ne soit retrouvée. Bien que ces observations puissent être partiellement expliquées par la présence d'autres facteurs d'exposition non identifiés, l'idée que des facteurs génétiques, puissent interagir de façon synergique ou antagoniste avec le tabagisme est particulièrement séduisante. Par exemple, on peut se demander s'il n'existerait pas un facteur génétique – conférant une certaine forme « d'immunité » – qui protègerait certains fumeurs du risque de développer un cancer. Comprendre et élucider le rôle des facteurs synergétiques ou antagonistes peut être une étape cruciale afin de déterminer les mécanismes biologiques sous-jacent du développement d'une maladie complexe mais également afin de structurer la mise en place de programmes de prévention primaire. C'est pour ces raisons que les questions relatives aux interactions entre des facteurs de risque génétiques et environnementaux sont d'un intérêt à la fois biologique et épidémiologique.

Comme en témoigne la figure 0.1 ci-dessous, il s'agit d'une thématique très récente et en plein essor. Près de la moitié des publications citant ce terme dans leurs titres ou leurs résumés ont été publiées après 2006. Cependant, la majorité des études d'épidémiologie génétique réalisées aujourd'hui continuent de faire abstraction de la présence potentielle d'interactions entre les facteurs génétiques et les facteurs environnementaux étudiés. Le manque d'outils statistiques efficaces pour prendre en compte les interactions peut être l'une des explications. En effet, les quelques méthodes proposées sont soit peu puissantes nécessitant donc des tailles d'échantillons irréalistes, soit difficiles à mettre en œuvre car s'appuyant sur des schémas d'étude différents de ceux généralement utilisés en génétique.

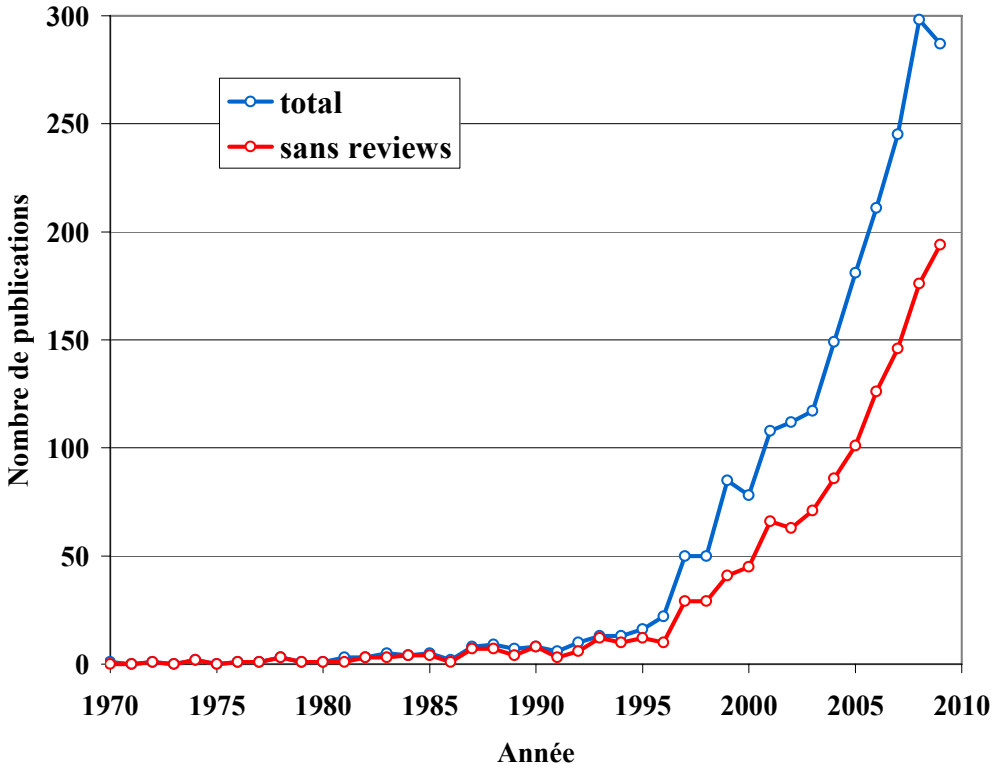


Figure 0.1 Nombre annuel de publications sur les interactions gène-environnement de 1970 à 2009

Requête PubMed (17/01/10) : (gene-environment[Title/Abstract] OR GxE[Title/Abstract] OR G x E[Title/Abstract] OR genotype-environment[Title/Abstract]) AND (interaction[Title/Abstract] OR interactions[Title/Abstract]) AND "humans"[MeSH Terms] AND English[lang]

Le développement de nouvelles méthodes pour prendre en compte les facteurs environnementaux dans les études génétiques apparaît donc comme une étape importante vers une meilleure compréhension des maladies. C'est l'objectif de ce travail qui propose deux nouvelles méthodes statistiques répondant à deux problématiques couramment rencontrées lors de l'étude des interactions gène-environnement dans les maladies complexes.

La première problématique est celle de la sélection des facteurs d'exposition qu'il faut prendre en compte dans les études génétiques. Les facteurs environnementaux qui peuvent jouer un rôle dans une maladie sont souvent multiples et il est impossible de les considérer tous ensemble dans une étude génétique. Nous avons donc recherché des critères permettant de sélectionner ceux qui sont les plus susceptibles d'interagir avec des facteurs génétiques.

Pour ce faire, nous avons développé un test statistique, qui permet, avant même de disposer d'informations génétiques, de détecter si un facteur environnemental est potentiellement impliqué dans une interaction avec un facteur génétique de susceptibilité.

La seconde problématique que nous aborderons est liée à l'utilisation récente de bases de données génotypiques de témoins de référence dans les études d'association pangénomiques. Pour ces témoins, nous ne disposons généralement pas d'information sur les variables d'exposition, ce qui constitue un obstacle important à la prise en compte des interactions gène-environnement. En effet, les méthodes statistiques disponibles pour tester ou prendre en compte les interactions gène-environnement nécessitent de disposer de l'information génétique et environnementale chez tous les individus, malades et non malades. Nous proposons une nouvelle approche fondée sur une régression logistique multinomiale qui permet de surmonter ce problème et ouvre donc de nouvelles perspectives pour la prise en compte des facteurs environnementaux dans les études d'association pangénomiques.

Cette thèse est organisée en trois parties. La première partie sera consacrée à la présentation des principes de l'épidémiologie génétique et des différentes méthodes disponibles pour prendre en compte les facteurs environnementaux dans les études génétiques. Dans la seconde partie, nous exposerons la première méthode de sélection des facteurs environnementaux que nous avons développée. Nous étudierons ses propriétés statistiques et l'appliquerons à des données concernant le diabète de type 2. La troisième partie sera consacrée à la deuxième méthode permettant de prendre en compte des interactions gène-environnement dans les études d'association génétique réalisées avec des panels de témoins de références pour lesquels les variables environnementales ne sont généralement pas documentées. Après avoir décrit la méthode et étudié ses propriétés statistiques, nous présenterons une application aux réactions cutanées sévères suite à la prise de médicaments.

Partie 1

Interaction gène-environnement : définitions, modèles et tests

Chapitre 1
Introduction à la génétique épidémiologique

Ce chapitre introduit les définitions et les concepts de base indispensables à la compréhension des principes de l'épidémiologie génétique. Nous y abordons la notion de variabilité génétique en présentant les différents types de polymorphismes génétiques que l'on peut étudier puis nous décrivons les caractéristiques de ce que l'on désigne par « facteur génétique » : les modèles de risque génétique, le modèle de Hardy-Weinberg qui régit les relations entre les fréquences des allèles et des génotypes à un locus donné, et la notion d'équilibre et de déséquilibre gamétique qui décrit les relations entre les fréquences alléliques à deux locus.

1. Variabilité du génome humain

1.1. Le matériel génétique

L'homme est un organisme eucaryote pluricellulaire diploïde. En cela, le noyau de la majorité de ses cellules contient de l'ADN ou acide désoxyribonucléique en double dose. L'ADN est une macromolécule constituée de l'union de deux brins dits anti-sens ou antiparallèles ayant une structure spatiale en « double hélice ». Chaque brin est constitué de l'assemblage de constituants élémentaires appelés les nucléotides dont on dénombre quatre types : l'adénine (A), la guanine (G), la cytosine (C) et la thymine (T). Les deux brins s'associent entre eux au niveau de ces bases nucléotidiques en établissant des liaisons « faibles » hydrogènes qui sont spécifiques. Ainsi, le nucléotide A d'un brin s'associe toujours avec T du brin complémentaire et C s'associe toujours avec G. L'ADN nucléaire de l'homme est composé de 3×10^9 paires de bases et constitue avec l'ADN mitochondrial ce que l'on appelle le génome humain. On y distingue les séquences codantes et les séquences non codantes. La séquence codante d'un gène est tout d'abord transcrite sous la forme d'une molécule d'ARN (acide ribonucléique) simple brin, appelée ARN messager, qui est transférée du noyau au cytoplasme où se déroule la traduction de l'ARN messager en protéines. Les séquences non codantes correspondent à des séquences d'ADN qui ne sont pas transcrites en ARN ou qui le sont sans être traduites en protéines. La définition du gène fondée sur l'idée

qu'à un gène correspond une protéine à laquelle correspond une fonction précise a été bouleversée par les nombreuses découvertes liées à l'étude du génome au cours des dix dernières années (Gerstein *et al.* 2007). Nous nous contenterons cependant ici de cette définition tout en gardant à l'esprit que nos modèles simplistes sont une approximation plus ou moins juste de la complexité des mécanismes que l'on cherche à décrire.

1.2. Les marqueurs génétiques

Les séquences nucléotidiques du génome de deux individus issus d'une même espèce sont identiques à 99,99 %. Cependant, le développement des techniques de biologie moléculaire a montré qu'à certains locus génétiques (localisation spécifique sur le génome), différentes formes de la séquence d'ADN pouvaient être observées. Ces formes possibles sont appelées des allèles et la paire d'allèles observée à un locus chez un individu constitue son génotype à ce locus. Ces variations se révèlent extrêmement utiles dans les analyses génétiques des maladies complexes. Elles représentent en quelque sorte des balises qui vont permettre de déterminer la localisation des locus responsables d'une susceptibilité génétique à une maladie. Nous allons brièvement décrire les différents types de polymorphismes qui peuvent être utilisés comme marqueurs génétiques. Classiquement, on distingue les polymorphismes dialléliques qui peuvent présenter uniquement deux allèles possibles, des polymorphismes multialléliques qui peuvent en présenter plus de deux. Nous présenterons également les variants structuraux qui forment une catégorie à part de polymorphismes pouvant être dialléliques ou multialléliques mais qui sont caractérisés par la grande taille du segment d'ADN concerné.

1.2.1. Polymorphismes dialléliques

Historiquement, les polymorphismes de longueur des fragments de restriction ou RFLP (*Restriction Fragment Length Polymorphism*) furent les premiers à être mis en évidence par des analyses biomoléculaires (*southern blot*). Ce sont des variations au niveau d'un seul nucléotide qui entraînent une modification des cartes de découpage des enzymes de restriction (enzymes qui coupent l'ADN au niveau d'une séquence spécifique). Elles sont définies par un couple sonde / enzyme de restriction qui permettent de différencier les deux formes alléliques possibles et donc de connaître le génotype à un locus donné.

Le développement des techniques de séquençage de l'ADN a par la suite permis d'identifier un nombre beaucoup plus important de polymorphismes dus à la substitution ponctuelle d'un nucléotide que l'on appelle les polymorphismes nucléotidiques simples ou SNP (*Single Nucleotide Polymorphism*). Bien qu'en théorie, ils puissent être trialléliques voire quadrialléliques, la rareté des évènements mutationnels font que la grande majorité des SNPs décrits sont dialléliques. On estime le nombre total de SNPs sur le génome à 10^6, dont plus d'un tiers ont été identifiés, localisés et décrits dans différentes populations grâce aux différentes phases du projet HapMap (The International HapMap Consortium 2003, 2005, 2007). Les SNPs sont distribués tout le long du génome à la fois dans des régions codantes et non codantes. Leur grande abondance, la simplicité de leur génotypage et leur caractère binaire en font le polymorphisme de choix pour « baliser » le génome à la recherche de facteurs de risque génétiques. La base de données dbSNP gérée par le *National Center for Biotechnology Information* (NCBI) répertorie l'ensemble des nouveaux SNPs découverts (Sherry *et al.* 2001). Notons que les RFLPs sont en réalité un sous-ensemble de SNPs qui ont la particularité de modifier le site d'action d'une enzyme de restriction.

1.2.2. Polymorphismes multialléliques

Les minisatellites ou VNTR (*Variable Number Tandem Repeat*) sont des RFLPs multialléliques. Il s'agit d'une répétition en nombre variable d'une séquence de nucléotides de 10 à 50 pb (paires de bases). Le nombre de répétitions qui caractérise un allèle peut être très élevé allant jusqu'à 1000 et la taille totale d'un minisatellite est de l'ordre de 1 à 10 kb (1 kb = 10^3 pb). Les minisatellites ont l'inconvénient d'être multilocus, c'est-à-dire qu'ils peuvent se trouver de façon identique à plusieurs endroits sur le génome avec à chaque fois différents allèles possibles.

Les microsatellites ou STR (*Short Tandem Repeat*) sont constitués de très courtes séquences d'ADN (1 à 6 pb) répétées en tandem. Détectés uniquement par amplification de l'ADN par PCR (*Polymerase Chain Reaction*), ces polymorphismes sont également caractérisés par une grande variabilité du nombre de répétitions. Comparés aux minisatellites, ils ont un degré de polymorphisme plus élevé, une répartition plus uniforme le long du génome, une taille plus petite (80 à 400 pb) et un nombre de répétitions moins important (5 à 40). Par ailleurs, chaque microsatellite correspond à un locus unique sur le génome, parfaitement défini par les séquences uniques qui l'encadrent. Il existe différents types de

microsatellites en fonction de la longueur du motif qui est répété, mais les plus couramment utilisés dans les analyses génétiques sont les motifs répétés en tandem de 2 nucléotides $(CA)_n$ sur un brin et $(GT)_n$ sur le brin complémentaire car ceux-ci sont facilement détectables, très polymorphes et sont les motifs les plus abondants sur le génome.

1.2.3. Variants structuraux

Un variant structural est une altération génomique d'un segment d'ADN de taille supérieure ou égale à 1 kb observable à l'échelle microscopique ou submicroscopique. En particulier, les études en épidémiologie génétique s'intéressent de plus en plus aux variants du nombre de copies. Un variant du nombre de copies ou CNV (*Copy Number Variant*) est un segment d'ADN d'une taille supérieure ou égale à 1 kb et qui est présent en un nombre variable de copies comparativement à un génome de référence. Le mécanisme sous-jacent d'apparition des CNVs peut être une insertion ou délétion du segment mais ce qui les différencie des indels (contraction utilisée pour désigner ces deux autres types de polymorphismes) est la taille du segment impliqué qui est inférieure à 1kb pour les indels (Feuk *et al.* 2006; Freeman *et al.* 2006).

2. Caractéristiques du facteur génétique

Nous allons maintenant aborder trois notions fondamentales relatives au « facteur génétique » dont on cherche à étudier l'effet sur une maladie complexe : la modélisation du risque génétique de la maladie, la modélisation de la relation entre les fréquences des allèles et des génotypes et la modélisation de la relation entre les fréquences des allèles entre deux locus. Si ce facteur génétique peut être multiallélique ou être une variation du nombre de copie, on considère en général pour la modélisation de ses effets qu'on peut le représenter comme un locus avec deux allèles *S* et *s* dont les fréquences respectives en populations sont q et p ($p = 1 - q$). Nous traiterons donc ici uniquement le cas de locus autosomiques dialléliques.

2.1. Modèle génétique de risque

Chaque individu peut porter l'un des trois génotypes $S||S$ $S||s$ et $s||s$. La pénétrance d'un génotype est la probabilité qu'un individu porteur de ce génotype soit malade. Ainsi, dans le modèle général de risque, trois pénétrances peuvent être calculées, P_2, P_1 et P_0 telles que :

$$P_2 = P \text{ (Malade sachant } S||S) \quad (1.1)$$
$$P_1 = P \text{ (Malade sachant } S||s) \quad (1.2)$$
$$P_0 = P \text{ (Malade sachant } s||s) \quad (1.3)$$

En prenant pour référence le génotype $s||s$, on définit deux risques relatifs génotypiques :

$$RR_{G2} = P_2 / P_0 \quad (1.4)$$
$$RR_{G1} = P_1 / P_0 \quad (1.5)$$

Il est cependant possible de restreindre le nombre de paramètres en ne modélisant les relations entre les pénétrances qu'avec un seul paramètre, le risque relatif du facteur génétique RR_G, en imposant des relations entre ces pénétrances.

Dans le cas d'un modèle dominant, la présence d'au moins une copie de l'allèle S augmente le risque de la maladie avec :

$$P_2 = P_1 = P_0 \times RR_G \quad (1.6)$$

Dans un modèle récessif, deux copies de l'allèle S sont nécessaires pour que le risque de la maladie soit augmenté avec :

$$P_2 = P_1 \times RR_G = P_0 \times RR_G \quad (1.7)$$

Finalement, dans un modèle additif les relations entre les pénétrances sont :

$$P_2 = P_1 \times RR_G = P_0 \times RR_G^2 \quad (1.8)$$

Afin de simplifier et d'uniformiser l'ensemble du travail exposé dans ce manuscrit, nous ne considèrerons par la suite qu'un modèle génétique dominant.

2.2. Modèle de Hardy-Weinberg

Le devenir de la variabilité génétique à un locus est difficilement prévisible d'une génération à l'autre et de nombreuses forces évolutives peuvent influer sur ces fréquences alléliques et génotypiques. En 1908, le mathématicien britannique G.H. Hardy et le physiologiste allemand R. Weinberg font simultanément le constat que sous certaines conditions, à une génération donnée, les fréquences des allèles à un locus sont reliées aux fréquences des génotypes. Ils démontrent également qu'en ajoutant d'autres hypothèses, ces fréquences demeurent stables d'une génération à l'autre (Hardy 1908). Le modèle proposé par Hardy et Weinberg est en quelque sorte la clef de voûte de la génétique des populations dont l'objectif est d'étudier l'effet des forces évolutives qui influent sur cet équilibre des polymorphismes du génome au sein des populations.

Reprenons l'exemple de notre locus de susceptibilité autosomique diallélique : *S et s* de fréquences respectives q et p dans la population étudiée. La relation de Hardy-Weinberg établit que les fréquences respectives des génotypes *S*||*S*, *S*||*s* et *s*||*s* sont égales à q^2, $2pq$ et p^2. Ces proportions également appelées proportions de Hardy-Weinberg nécessitent deux conditions pour être valides. La panmixie (pangamie) est l'union aléatoire des individus (et de leurs gamètes) qui permet un assortiment indépendant des deux allèles au locus. L'effectif infini de la population permet d'appliquer la loi des grands nombres afin de considérer que les fréquences des génotypes sont égales à leurs probabilités respectives d'être observées.

L'équilibre de Hardy-Weinberg stipule lui que ces fréquences alléliques et génotypiques restent inchangées aux générations suivantes. On dit alors que la population est à l'équilibre. Quatre conditions supplémentaires sont nécessaires. L'absence de mutation (ou de distorsion de ségrégation méiotique) implique qu'un individu *S*||*s* par exemple produira toujours pour moitié des gamètes *S* et pour moitié des gamètes *s*. L'absence de sélection permet à tout individu, quel que soit son génotype, d'avoir la même probabilité de contribuer par ses gamètes à la génération suivante. L'absence de migration évite une modification de la constitution génotypique de la population par un apport extérieur différent. Finalement, on suppose des générations non chevauchantes, c'est-à-dire qu'aucun croisement n'a lieu entre individus de générations différentes (Serre 1997).

2.3. Équilibre gamétique

Deux allèles situés sur des locus différents sont dits associés ou en déséquilibre gamétique lorsque ces allèles sont présents chez un même individu plus souvent que ne le prédit un assortiment gamétique au hasard. Si l'assortiment des deux allèles se fait au hasard, on dit qu'il y a équilibre gamétique. À l'équilibre gamétique, la fréquence d'un gamète porteur des deux allèles est égale au produit des fréquences de ces deux allèles dans la population (Lewontin 1988).

Considérons donc maintenant deux locus dialléliques situés sur des autosomes, par exemple notre locus de susceptibilité avec ses allèles *s* et *S* de fréquences *p* et *q* ($p > q$) et un locus marqueur pouvant porter les allèles *m* ou *M* avec les fréquences *u* et *v* ($u > v$). Les gamètes produits dans la population à une génération donnée sont porteurs des allèles *sm*, *Sm*, *sM* et *SM* avec les fréquences respectives g_{00}, g_{01}, g_{10} et g_{11}.

A l'équilibre gamétique (indépendance de la distribution des allèles aux deux locus), les fréquences attendues des gamètes *sm*, *Sm*, *sM* et *SM* sont égales à *pu*, *qu*, *pv* et *qv* respectivement. Le déséquilibre gamétique est mesuré par l'écart des fréquences gamétiques observées à celles attendues à l'équilibre et est égal à :

$$D = g_{00}\, g_{11} - g_{01}\, g_{10} \qquad (1.9)$$

Une valeur de *D* positive indique une plus grande fréquence des gamètes *sm* et *SM* par rapport aux gamètes *Sm* et *sM* alors qu'une valeur négative indique une plus grande fréquence des gamètes *Sm* et *sM* par rapport aux gamètes *sm* et *SM*. Le χ^2 d'association entre les deux locus d'un échantillon observé de *N* gamètes s'exprime en fonction du déséquilibre gamétique par :

$$\chi^2 = \frac{D^2 N}{pquv} \qquad (1.10)$$

L'inconvénient de cette mesure est que la valeur maximale qu'elle peut prendre dépend des fréquences alléliques. La valeur du déséquilibre *D* ne permet donc pas de mesurer l'importance quantitative du déséquilibre et de comparer différentes valeurs de déséquilibre gamétique mesurées entres des locus différents.

Afin de s'affranchir de cette dépendance, Lewontin (Lewontin 1964) introduit le *D*' qui consiste à réduire la valeur du déséquilibre gamétique *D* par la valeur maximale qu'il peut prendre :

$$D' = D / D_{max} \quad (1.11)$$

où $D_{max} = \min (pv, qu)$ quand $D > 0$

$D_{max} = \min (pu, qv)$ quand $D < 0$

Les valeurs de *D*' peuvent varier entre –1 et 1. La valeur 0 correspond à l'équilibre gamétique et les valeurs –1 et 1 indiquent un déséquilibre maximal lorsque l'un des quatre gamètes possibles n'est pas observé.

Une troisième mesure du déséquilibre gamétique consiste à calculer une corrélation entre les deux locus par la formule :

$$r^2 = \frac{D^2}{pquv} \quad (1.12)$$

La valeur du r^2 varie entre 0 (équilibre gamétique) et 1 (association complète). Dans la situation d'une association complète, un allèle donné au premier locus est toujours associé au même allèle au second locus. En effet, un r^2 égal à 1 nécessite que *p* soit égal à *u* et *q* à *v* et par conséquent seuls deux des quatre gamètes possibles sont observés (Hill et Robertson 1968).

Cependant, comme le note Lewontin lui-même, que ce soit pour le *D*' ou le r^2, ces deux mesures ont des bornes standardisées de telle façon à ce qu'elles représentent les situations de déséquilibre maximum ou d'association complète, mais leurs variations entre les bornes demeurent dépendantes des fréquences alléliques (Lewontin 1988).

Jusqu'à présent nous n'avons fait aucune hypothèse concernant la localisation chromosomique des deux locus concernés. Un déséquilibre gamétique peut exister entre deux locus situés sur des chromosomes différents et résulte souvent d'un mélange de deux populations génétiquement distinctes avec des compositions différentes. Dans la situation particulière où les deux locus sont situés sur le même chromosome, on parle de déséquilibre de liaison. Le déséquilibre de liaison suppose donc à la fois un déséquilibre gamétique et une liaison génétique. Cette association allélique est maintenue par la proximité physique des locus et le caractère récent de la mutation ayant produit l'un des deux allèles.

Chapitre 2
Qu'est-ce qu'une interaction gène-environnement ?

Dans ce chapitre, nous définissons tout d'abord la notion d'interaction gène-environnement (G×E) d'un point de vue exclusivement statistique. Puis ensuite, nous présentons les mesures de risque utilisées en épidémiologie génétique et leur expression dans la modélisation d'une interaction G×E. Nous terminons ce chapitre en présentant une classification des différents types d'interaction G×E.

1. Définition statistique

Afin de définir l'interaction G×E, il est nécessaire de rappeler le concept d'indépendance conditionnelle qui stipule que la relation entre deux facteurs demeure inchangée en stratifiant sur un troisième facteur. Dans le contexte de deux facteurs de risque génétique *G* et environnemental *E* et d'une maladie *M*, l'indépendance conditionnelle implique que l'association mesurée entre le facteur *G* et la maladie *M* demeure identique, dans les différentes strates possibles de *E* et que symétriquement, l'association mesurée entre le facteur *E* et la maladie *M* reste constante dans toutes les strates de *G*. Si cette condition n'est pas remplie, on dit alors qu'il y a interaction entre les deux facteurs *G* et *E* (Ottman 1995).

Ainsi, une interaction G×E est présente si le risque de développer la maladie associé à l'exposition au facteur *E* n'est pas le même selon les génotypes des individus au locus porteur du facteur *G* ou inversement les risques associés aux différents génotypes du facteur *G* varient selon le statut d'exposition à *E*. Il s'agit donc d'un écart à l'hypothèse d'indépendance des effets respectifs des facteurs *G* et *E* sur *M* qui peut se produire dans les deux sens. Soit l'effet observé en présence des deux facteurs de risque est supérieur à celui attendu sous l'hypothèse d'indépendance des effets. Cette situation correspond à une interaction G×E synergétique. Soit inversement, l'interaction tend à diminuer l'effet global observé et alors il s'agit d'une interaction G×E antagoniste. Cette déviation peut être statistiquement quantifiée par un terme d'interaction mesurant l'écart par rapport à un modèle de risque multiplicatif ou additif que nous décrirons plus en détail dans les paragraphes qui suivent.

2. Modèle général d'interaction gène-environnement

2.1. Description des variables du modèle

Considérons une maladie M, un facteur de risque génétique G et un facteur de risque environnemental E. Ces trois variables sont supposées dichotomiques : 0 indiquant l'absence et 1 indiquant la présence de la maladie ou du facteur. Nous nous plaçons ici délibérément dans une situation simple où les variables sont dichotomiques afin de restreindre le nombre de paramètres du modèle. La plupart des méthodes employées en épidémiologie génétique sont cependant adaptables à des variables polytomiques ou quantitatives.

La maladie M a une prévalence f_M et une probabilité de base f_B, en population. La probabilité de base f_B correspond au risque de maladie pour un individu non exposé au facteur E et non porteur du génotype de susceptibilité G, soit $P(M = 1 | G = 0, E = 0)$.

Le facteur de risque génétique correspond à un locus génétique diallélique avec pour allèle de référence s et pour allèle de susceptibilité à la maladie S dont les fréquences respectives en population sont p et q, avec $q = 1 - p$. Ainsi chaque individu est porteur de l'un des trois génotypes possibles $S||S$ $S||s$ et $s||s$ dont les fréquences dans la population sont respectivement q^2, $2pq$ et p^2 en supposant que les fréquences génotypiques suivent les proportions de Hardy-Weinberg. Afin de limiter le nombre de paramètres dans nos analyses, nous considérerons uniquement un modèle génétique dominant. Par conséquent, la variable G indique la présence d'au moins une copie de l'allèle S ($G = 1$) et la fréquence f_G du facteur G est égale à la somme des fréquences des deux génotypes de susceptibilité, $q^2 + 2pq$.

Le facteur de risque environnemental peut être une exposition physique, chimique, ou biologique ; un facteur comportemental, social ou un évènement de vie ; une mesure anthropométrique ou un marqueur biologique d'exposition. Ce facteur de risque environnemental a une fréquence dans la population générale f_E. Dans la majorité des situations, l'hypothèse est faite que les deux facteurs G et E sont indépendants dans la population générale, c'est-à-dire que la probabilité qu'un individu soit à la fois porteur d'un génotype de susceptibilité ($G = 1$) et exposé au facteur de risque environnemental ($E = 1$) est le produit des probabilités respectives de chacun des deux facteurs, $P(G = 1, E = 1) = f_G \times f_E$.

Un moyen de modéliser une corrélation gène-environnement consiste à conditionner la prévalence du facteur environnemental en fonction du génotype. Pour cela, on utilise un coefficient qui représente le degré d'association entre G et E, θ_{GE} tel que proposé par Lindström *et al.* (2009) :

$$\theta_{GE} = \frac{P(E=1|\,G=1)/P(E=0|\,G=1)}{P(E=1|\,G=0)/P(E=0|\,G=0)} \qquad (1.13)$$

Une valeur de θ_{GE} égale à 1 indique l'indépendance entre G et E et alors $P(E = 1|\ G = 1) = P(E = 1|\ G = 0) = f_E$. Le coefficient θ_{GE} peut ensuite varier entre 0 et $+\infty$ mesurant le degré d'association entre G et E dans la population.

2.2. Modélisation des risques

On appelle effet indépendant (ou parfois effet propre) du facteur de risque G (ou E) sur la maladie, la mesure de risque liée au facteur G (ou E) dans le sous-groupe d'individus non exposés au facteur E (ou non porteurs du facteur G). On appelle effet marginal du facteur de risque G (ou E) sur la maladie, la mesure de risque liée au facteur G (ou E) sur l'ensemble des individus de la population étudiée en faisant abstraction de la présence ou de l'absence de l'autre facteur.

Dans le cadre d'une étude en population, deux cas de figure sont possibles. D'une part, l'étude de cohorte consiste à mesurer les facteurs de risque étudiés G et E d'un échantillon aléatoire de la population à un temps donné, puis à observer de façon prospective la survenue de la maladie M. Ce type d'étude permet de mesurer pour chacun des facteurs de risque étudiés un risque relatif (*RR*) de maladie. Il est d'autre part possible de recruter deux échantillons, l'un composé de sujets malades (les cas, $M = 1$) et l'autre de sujets non malades (les témoins, $M = 0$), puis de rechercher rétrospectivement la présence des facteurs de risques G et E. Dans ce type d'étude où les malades sont surreprésentés dans l'échantillon total, la mesure de l'association entre un facteur de risque donné et la maladie se fait par le calcul d'un odds ratio (*OR*). L'*OR* est une bonne approximation numérique du *RR* lorsque la maladie est rare et que l'effet du facteur n'est pas trop grand (Bouyer *et al.* 1995). Il est donc possible de modéliser les risques de chacun des facteurs G et E et leur interaction selon ces deux types de mesure de risque. Le tableau 1.1 représente la distribution des catégories (M, G et E) des sujets et les mesures de risque associées dans une étude de cohorte ou une étude cas-témoin.

Par ailleurs, nous avons choisi de modéliser l'interaction G×E sur une échelle multiplicative. Nous reviendrons sur la différence entre l'échelle multiplicative et l'échelle additive au paragraphe 2.3.

2.2.1. Risques relatifs

Dans ce type de modélisation, l'effet indépendant de *G* est mesuré par le risque relatif génétique RR_G égal au rapport de probabilités de maladie ($M = 1$) chez des porteurs et des non porteurs de *G*, tous non exposés :

$$RR_G = \frac{P(M = 1 | G = 1, E = 0)}{P(M = 1 | G = 0, E = 0)} \quad (1.14)$$

L'effet indépendant de *E* est mesuré par le risque relatif environnemental RR_E égal au rapport de probabilités de maladie ($M = 1$) chez des exposés et non exposés à *E*, tous non porteurs d'un génotype de susceptibilité :

$$RR_E = \frac{P(M = 1 | E = 1, G = 0)}{P(M = 1 | E = 0, G = 0)} \quad (1.15)$$

En analysant la distribution jointe de *G* et de *E*, l'effet lié à la présence simultanée des deux facteurs *G* et *E* est mesuré par le risque relatif RR_{GE} :

$$RR_{GE} = \frac{P(M = 1 | G = 1, E = 1)}{P(M = 1 | G = 0, E = 0)} \quad (1.16)$$

En l'absence d'interaction G×E et en supposant un modèle multiplicatif à l'échelle des risques (ou additif à l'échelle des logarithmes de risques), RR_{GE} est égal au produit des effets indépendants des deux facteurs RR_G et RR_E.

En présence d'une interaction G×E, les équations 1.14, 1.15 et 1.16 permettent de mesurer la déviation par rapport à l'indépendance des effets des deux facteurs par un risque relatif d'interaction RR_I :

$$RR_I = \frac{RR_{GE}}{RR_G \times RR_E} \quad (1.17)$$

Lorsque $RR_I > 1$ (< 1) l'exposition à E augmente (diminue) l'effet mesuré de G, ou symétriquement, la présence de G entraîne une susceptibilité accrue (une protection par rapport) à l'effet de E.

2.2.2. Odds ratios

L'effet indépendant de G est mesuré par l'odds ratio génétique OR_G égal au rapport des cotes de G chez des malades et chez des témoins, tous non exposés ($E = 0$) :

$$OR_G = \frac{P(G=1 \mid M=1, E=0) / P(G=0 \mid M=1, E=0)}{P(G=1 \mid M=0, E=0) / P(G=0 \mid M=0, E=0)} \quad (1.18)$$

L'effet indépendant de E est mesuré par l'odds ratio environnemental OR_E égal au rapport des cotes de E chez des malades et chez des témoins, tous non porteurs d'un génotype de susceptibilité ($G = 0$) :

$$OR_E = \frac{P(E=1 \mid M=1, G=0) / P(E=0 \mid M=1, G=0)}{P(E=1 \mid M=0, G=0) / P(E=0 \mid M=0, G=0)} \quad (1.19)$$

En analysant la distribution jointe de G et de E, l'effet lié à la présence simultanée des deux facteurs G et E est mesuré par l'odds ratio OR_{GE} :

$$OR_{GE} = \frac{P(E=1, G=1 \mid M=1) / P(E=0, G=0 \mid M=1)}{P(E=1, G=1 \mid M=0) / P(E=0, G=0 \mid M=0)} \quad (1.20)$$

En l'absence d'interaction G×E et en supposant un modèle multiplicatif à l'échelle des cotes (ou additif à l'échelle des logarithmes de cotes), OR_{GE} est égal au produit des effets marginaux des deux facteurs OR_G et OR_E.

En présence d'une interaction G×E, les équations 1.18, 1.19 et 1.20 permettent de mesurer la déviation par rapport à l'indépendance des effets des deux facteurs par un odds ratio d'interaction OR_I :

$$OR_I = \frac{OR_{GE}}{OR_G \times OR_E} \quad (1.21)$$

Lorsque $OR_I > 1$ (< 1) l'exposition à E augmente (diminue) l'effet mesuré de G, ou symétriquement, la présence de G entraîne une susceptibilité accrue (une protection par rapport) à l'effet de E.

	$G = 1$		$G = 0$		
	$E = 1$	$E = 0$	$E = 1$	$E = 0$	
			Étude de cohorte		
Malades ($M = 1$)	a	b	e	f	n_0
Non malades ($M = 0$)	c	d	g	h	n_1
Risque de M $P(M = 1\|G, E)$	$r_{GE} = a / (a + c)$	$r_G = b / (b + d)$	$r_E = e / (e + g)$	$f_B = f / (f + h)$	$f_M = n_0 / n_1$
Risques relatifs (RR)	$RR_{GE} = r_{GE} / f_B$	$RR_G = r_G / f_B$	$RR_E = r_E / f_B$	$RR_0 = 1$ (réf.)	
			Étude cas-témoin		
Cas ($M = 1$)	a	b	e	f	n_0
Témoins ($M = 0$)	c	d	g	h	n_1
Odds ratio (OR)	$OR_{GE} = ah / cf$	$OR_G = bh / df$	$OR_E = eh / gf$	$OR_0 = 1$ (réf.)	

Tableau 1.1 Mesures de risque des facteurs génétique et environnemental dans une étude de cohorte et dans une étude cas-témoin (Ottman 1996)

f_B est le risque de base de maladie chez les non exposés et non porteurs d'un génotype à risque

r_E est le risque de maladie chez les exposés et non porteurs d'un génotype à risque

r_G est le risque de maladie chez les non exposés et porteurs d'un génotype à risque

r_{GE} est le risque de maladie chez les exposés et porteurs d'un génotype à risque

2.3. Interaction multiplicative versus interaction additive

Lorsqu'il s'agit d'étudier l'effet simultané de deux facteurs sur le risque de survenue d'une maladie, deux échelles de mesure existent en épidémiologie quantitative : l'échelle additive et l'échelle multiplicative. Ce qui différencie ces deux types d'échelle est la relation mathématique existant entre le risque relatif (ou l'odds ratio) de l'effet conjoint de *G* et *E* et les risques relatifs (ou les odds ratios) des effets indépendants de ces deux facteurs. Le tableau 1.2 présente ces relations pour les deux échelles de mesure, en l'absence d'interaction G×E, en présence d'une interaction synergétique ou en présence d'une interaction antagoniste. Si l'effet conjoint des deux variables respecte la condition d'absence d'interaction G×E sur l'échelle additive (ou multiplicative), on dira des données qu'elles suivent « un modèle additif (ou multiplicatif) de risques ». L'interaction G×E, telle que nous l'avons définie au paragraphe 1 de ce chapitre, correspond alors à une déviation par rapport à l'un de ces deux modèles. On parlera alors d'une interaction à « l'échelle additive (ou multiplicative) ».

	Échelle de mesure	
	additive	**multiplicative**
Pas d'interaction	$RR_{GE} = RR_G + RR_E - 1$	$RR_{GE} = RR_G \times RR_E$
	$OR_{GE} = OR_G + OR_E - 1$	$OR_{GE} = OR_G \times OR_E$
Interaction synergétique	$RR_{GE} > RR_G + RR_E - 1$	$RR_{GE} > RR_G \times RR_E$
	$OR_{GE} > OR_G + OR_E - 1$	$OR_{GE} > OR_G \times OR_E$
Interaction antagoniste	$RR_{GE} < RR_G + RR_E - 1$	$RR_{GE} < RR_G \times RR_E$
	$OR_{GE} < OR_G + OR_E - 1$	$OR_{GE} < OR_G \times OR_E$

Tableau 1.2 Définition de l'interaction gène-environnement sur une échelle additive ou multiplicative (Ottman 1996)

La question du choix de l'échelle de risque à adopter pour déterminer s'il y a interaction entre deux facteurs a été largement débattue dans la littérature en épidémiologie (Kupper et Hogan 1978; Ottman 1996; Rothman *et al.* 1980; Walter et Holford 1978). Cependant, Rothman *et al.* (1980) indiquent que le choix de l'échelle de mesure dépend

principalement de l'objectif de l'étude. Si le but principal est de mettre en évidence le rôle étiologique de ces facteurs dans la maladie, une échelle multiplicative sera plus appropriée ; s'il s'agit plutôt de prédire le nombre de cas dans la population, c'est l'échelle additive qu'il faudra choisir. La tendance actuelle dans les études en épidémiologie génétique est d'adopter un modèle multiplicatif de risque. Du fait de l'utilisation des modèles linéaires généralisés, l'interaction est souvent mesurée à partir de la déviation par rapport à un modèle multiplicatif (additif à l'échelle des logarithmes de risques). C'est pourquoi, nous avons également choisi d'adopter le modèle multiplicatif au cours des travaux présentés ici.

3. Types d'interaction gène-environnement

Khoury *et al.* (1988a) proposent de classer les interactions G×E en six catégories décrivant six situations biologiques plausibles (tableau 1.3). Ces types d'interaction G×E dépendent des risques relatifs indépendants des facteurs environnemental (RR_E) et génétique (RR_G) et du risque relatif lié à la présence des deux facteurs (RR_{GE}) qui est le produit de RR_G, de RR_E et du coefficient d'interaction G×E (RR_I). Le même raisonnement est valable en utilisant les odds ratios. La figure 1.1 illustre par des représentations schématiques les relations entre les deux facteurs de risque *G* et *E* et la maladie *M* pour une situation sans interaction G×E (modèle 0), pour les six types d'interaction G×E (modèles 1 à 6) et pour une situation sans interaction G×E mais avec une corrélation entre *G* et *E* (modèle C). Dans la figure 1.2 sont représentés les risques de maladie sachant *G* et *E* de ces différents cas de figure en prenant les valeurs particulières de risques relatifs indiquées dans le tableau 1.3. Les pentes des droites rouge, jaune et orange représentent les augmentations de risque de maladie liée au facteur génétique mesurées respectivement chez les exposés, chez les non exposés, et marginalement. En l'absence d'interaction G×E, ces trois droites sont parallèles (sur une échelle logarithmique du risque) ce qui est cohérent avec la définition statistique d'indépendance conditionnelle que nous avons précédemment décrite. En présence d'une interaction G×E, ces droites ne sont plus parallèles ce qui indique que l'effet du facteur *G* n'est plus indépendant de *E*. Dans certaines situations, l'effet marginal du facteur *G* est faible (modèles 1, 2 et 6), alors que les effets de *G* stratifiés en fonction de *E* sont beaucoup plus contrastés. À l'extrême, le modèle 5 représente une situation où aucun effet de *G* n'est observé marginalement alors que son effet est tantôt protecteur (en l'absence de *E*), tantôt délétère (en présence de *E*). Ce type d'interaction où le génotype étudié a un effet qui s'inverse en fonction de la présence ou de l'absence du facteur *E* correspond à une interaction

antagoniste, souvent désignée par le terme anglais « *flip-flop interaction* ». Finalement, notons le biais de confusion induit par la corrélation entre le facteur *G* et *E* (coefficient d'association de l'équation 1.13, $\theta_{GE} = 2$) en présence d'un effet du facteur *E* sur la maladie. L'effet de *G* qui est observé marginalement n'est en réalité que le reflet de cette corrélation et il disparaît lorsque l'on étudie son effet dans les deux strates de *E*.

Modèle	Description	RR_G	RR_E	RR_I	RR_{GE}
0	Effets indépendants de *G* et de *E* Pas d'interaction G×E	2	2	1	4
1	Pas d'effets indépendants de *G* ou de *E* Présence d'une interaction G×E pure	1	1	2	2
2	Effet indépendant de *E* seulement Interaction G×E : *G* renforce l'effet de *E*	1	2	2	4
3	Effet indépendant de *G* seulement Interaction G×E : *E* renforce l'effet de *G*	2	1	2	4
4	Effets indépendants de *G* et de *E* Interaction G×E : synergie des deux facteurs	2	2	2	8
5	Effet indépendant protecteur de *G* Interaction G×E antagoniste par rapport à *G*	0,5	1	5	2,5
6	Effet indépendant protecteur de *G* Effet indépendant de *E* inverse à *G* Interaction G×E antagoniste par rapport à *G*	0,5	2	2	2
C	Effet indépendant de E seulement Pas d'interaction G×E Corrélation entre les facteurs *G* et *E* ($\theta_{GE} = 2$)	1	2	1	2

Tableau 1.3 Types d'interaction gène-environnement (Khoury *et al.* 1988a)

RR_G, RR_E et RR_I représentent les risques relatifs du facteur génétique, du facteur environnemental et de l'interaction G×E. RR_{GE} représente le risque relatif associé à la présence simultanée des deux facteurs génétique et environnemental, soit le produit de RR_G, RR_E et RR_I. θ_{GE} correspond à la mesure de l'association entre le facteur *G* et le facteur *E* indépendamment de la maladie (équation 1.13).

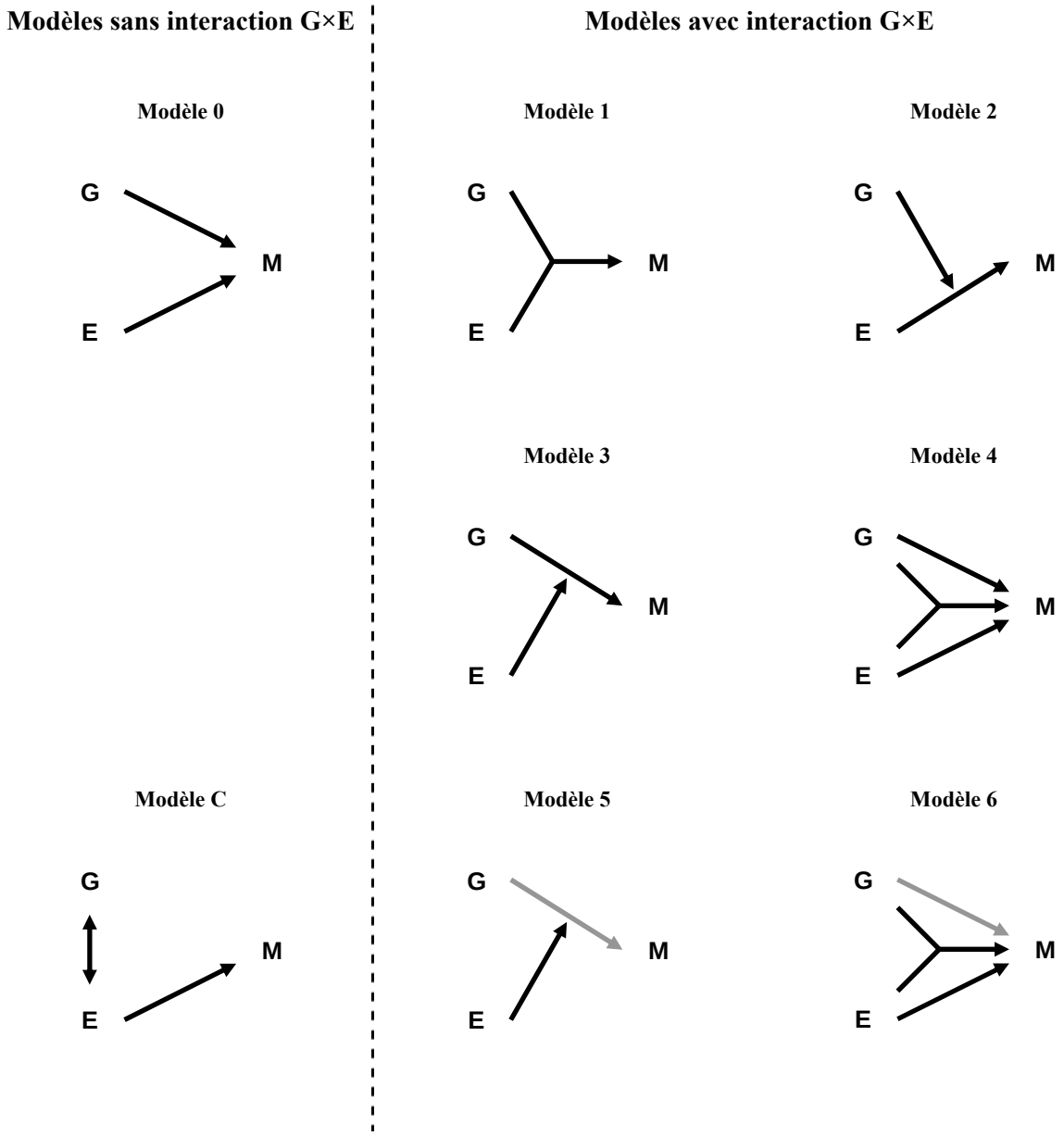


Figure 1.1 Représentation schématique des types d'interaction gène-environnement

Une flèche unidirectionnelle entre un facteur et la maladie indique une relation de causalité.

Une flèche unidirectionnelle entre un facteur et une autre flèche unidirectionnelle indique une interaction modulant un effet indépendant préexistant.

Une flèche unidirectionnelle démarrant au niveau des deux facteurs indique un effet d'interaction nécessitant la présence simultanée des deux facteurs.

La flèche bidirectionnelle entre les deux facteurs indique une corrélation entre eux (modèle C).

En noir sont indiquées les associations positives (risque relatifs ou odds ratios supérieurs à 1).

En gris sont indiquées les associations négatives (effets protecteurs).

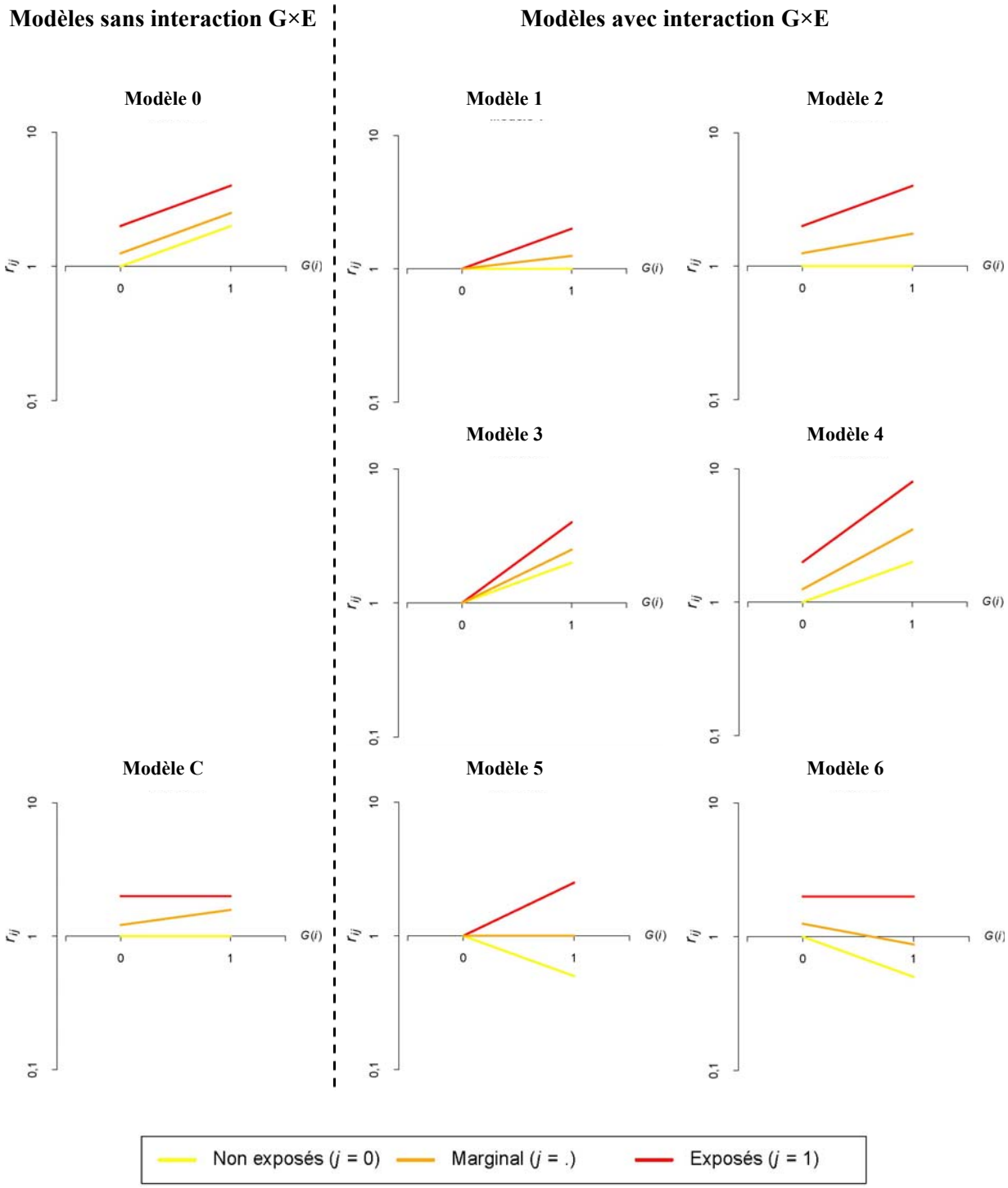


Figure 1.2 Représentation graphique du risque de maladie pour les différents types d'interaction gène-environnement

r_{ij} correspond au risque de maladie, $P(M = 1 | G = i, E = j)$

4. Quid de l'interaction biologique

Le terme « interaction gène-environnement » peut avoir un sens différent selon les personnes. Tandis qu'un biologiste aura tendance à y voir la description d'un mécanisme biologique particulier dans lequel interviennent ces deux facteurs, le statisticien lui y voit une déviation par rapport à un modèle mathématique particulier de leur effet conjoint. Une interprétation du mécanisme de l'interaction statistique n'a de sens que si le modèle mathématique a une interprétation biologique plausible. Ceci est rarement le cas, puisque « la majorité des modèles mathématiques sont des fictions commodes qui seraient très certainement rejetées avec des tailles d'échantillon suffisamment grandes » (Clayton 2009). Wahrendorf (1981) relève que le terme « interaction » est très largement utilisé en statistiques et que bien qu'il ait une « signification arithmétique concrète dans tous les modèles statistiques », il est souvent utilisé lorsque « quelque chose d'inhabituel, quelque chose de non spécifique est décrit » sans que l'on ne tente de comprendre « le phénomène sous-jacent ». Thompson (1991) constate que dix ans après que « le concept d'interaction ait été placé au centre des débats sur la causalité des maladies, la controverse sur la nature de l'interaction s'est atténuée sans qu'aucune réponse adéquate n'ait été apportée aux problèmes conceptuels et pragmatiques qui avaient été soulevés ».

Nous ne nous étendrons pas plus amplement sur ce point qui n'est pas l'objet du travail présenté dans ce manuscrit, mais l'élément essentiel à conserver à l'esprit est que l'interaction G×E telle que nous venons de la définir dans ce chapitre est une interaction au sens statistique. Tout comme la mise en évidence d'une association entre un facteur et une maladie ne signifie pas qu'il y ait nécessairement une relation de causalité, la présence d'une interaction G×E statistiquement significative n'implique pas qu'il y ait une interaction biologique sous-jacente entre ces deux facteurs dans la survenue ou l'évolution de la maladie, bien que nous le souhaitions. Cette mise en garde ne signifie pas que les analyses statistiques des interactions G×E dans les maladies complexes soient dénuées d'un quelconque intérêt scientifique mais que ces résultats n'ont une portée étiologique que lorsqu'ils sont accompagnés d'arguments et d'investigations biologiques.

Chapitre 3
Comment tester une interaction gène-environnement ?

Les outils statistiques pour identifier les facteurs de risque génétiques dans les maladies complexes ont évolué de façon importante en réponse à la fois aux progrès de la biologie moléculaire permettant de mieux caractériser les polymorphismes génétiques mais aussi grâce aux révolutions informatiques permettant aujourd'hui de réaliser des calculs encore inconcevables il y a trente ans. Nous décrivons tout d'abord dans ce chapitre la stratégie générale pour étudier les facteurs génétiques impliqués dans les maladies complexes, puis nous présentons les méthodes les plus couramment utilisées pour tester ou tenir compte des interactions G×E. Classiquement, ces méthodes sont divisées selon qu'elles utilisent, comme mesure du facteur génétique, la liaison et/ou l'association. Nous détaillons ici pour chacune de ces méthodes les échantillons, les types d'informations génétique et environnementale qu'elles nécessitent, ainsi que leurs avantages et inconvénients.

1. Stratégie d'étude du facteur génétique dans une maladie complexe

Comme nous l'avons défini en introduction, l'épidémiologie génétique s'intéresse principalement au lien entre les facteurs génétiques et les maladies complexes. Le facteur génétique que l'on recherche est un locus sur le génome ayant de part sa variation de fréquence en population et sa fonction biologique une influence sur la survenue ou l'évolution de la maladie. Cependant, avant de pouvoir identifier un tel locus spécifiquement, d'en déterminer l'effet et d'en confirmer le lien de causalité par des études fonctionnelles *in vitro* et *in vivo*, la stratégie générale utilisée en épidémiologie génétique est fondée sur différentes approches que nous allons décrire successivement (figure 1.3).

1.1. Étude de l'agrégation familiale

La première étape consiste à rechercher l'existence de la composante génétique dans la maladie et d'en mesurer l'importance. Pour cela, on cherche à mettre en évidence une certaine agrégation familiale de la maladie par exemple en mesurant le risque de récurrence relatif pour un type d'apparentement donné R (λ_R).

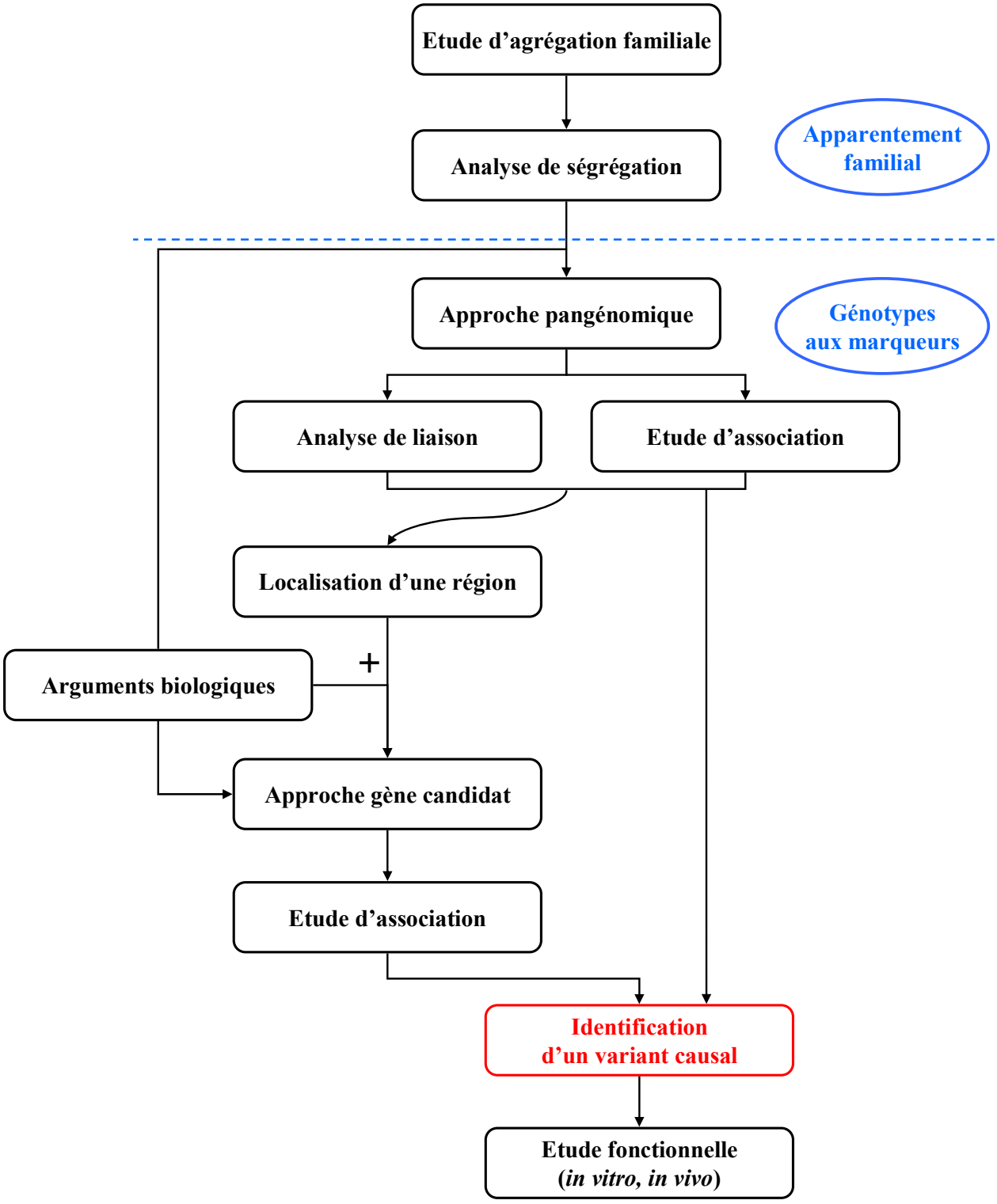


Figure 1.3 Stratégie d'étude de la composante génétique d'une maladie complexe

Par exemple, pour l'apparentement entre germains *S* (pour *sib* : frères et/ou sœurs), le risque de récurrence relatif λ_S est égal au ratio entre le risque de récurrence de la maladie chez les germains d'individus atteints (K_S) avec la prévalence de la maladie dans la population (f_M) (Risch 1990a) :

$$\lambda_S = \frac{K_S}{f_M} \tag{1.22}$$

Cette mesure descriptive rend compte d'une certaine « transmission » familiale et donne donc une première indication sur la présence d'une composante génétique dans la maladie étudiée.

Une autre façon de déterminer la présence d'une composante génétique en utilisant aussi l'agrégation familiale consiste à recruter des échantillons de malades et de non-malades et à comparer les risques de récurrence chez les apparentés de ces deux groupes.

La présence d'une association entre une maladie et le lien de parenté entre individus peut être le reflet d'un facteur *G* transmis mais peut également être due à un facteur *E* partagé au sein de la famille. Cette non-spécificité de l'information de récurrence en limite l'usage à celui d'un simple indicateur qu'il ne faut toutefois pas sous-estimer. Il peut avoir une utilité importante notamment dans la construction de tests de « débrouillage » où l'on cherche à obtenir une bonne sensibilité vis-à-vis du facteur génétique au détriment d'une moindre spécificité. Un autre problème dont souffre la mesure de la récurrence familiale d'une maladie est que son estimation peut être entachée d'un biais de sélection (Cordell et Olson 2000; Guo 1998, 2002).

Dans le même esprit, les études de jumeaux ont traditionnellement été utilisées pour estimer la contribution génétique à une maladie en comparant des paires de jumeaux monozygotes (MZ) qui partagent tous leurs allèles, aux jumeaux dizygotes (DZ) qui n'en partagent que la moitié. L'intérêt de cette approche est qu'elle permet de dissocier les facteurs partagés non génétiques, qui auront des effets similaires chez les jumeaux MZ et DZ, des facteurs génétiques, qui entraînent une plus grande similarité des jumeaux MZ par rapport aux jumeaux DZ. Cependant, le recrutement spécifique de jumeaux limite de façon importante les tailles d'échantillons qu'il est possible de recruter (Khoury *et al.* 1993a).

1.2. Analyse de ségrégation familiale

Une fois qu'une composante génétique est fortement suspectée dans une maladie complexe, la seconde étape cherche à déterminer si un mode de transmission familiale peut être clairement identifié. Pour cela, on utilise des généalogies de familles contenant au moins deux atteints et on étudie la ségrégation de la maladie au sein des familles, c'est-à-dire que l'on cherche à déterminer le mode de transmission qui explique le mieux les généalogies observées (Elston et Yelverton 1975). Par des techniques de maximisation de la vraisemblance, il est possible de tester différentes hypothèses et de déterminer par exemple si un seul gène majeur est responsable de la maladie où si un modèle polygénique (nombreux gènes à effets individuels faibles) est plus vraisemblable au vue des données. Il est évident que l'analyse de ségrégation n'a de réel intérêt que pour des modèles simples (maladies monogéniques dont le gène responsable est diallélique) et que pour la majorité des maladies complexes pour lesquelles de nombreux facteurs génétique et environnementaux interagissent, ces méthodes ont une faible puissance et sont donc d'un usage très limité (Khoury *et al.* 1993b).

Que ce soit pour les études de l'agrégation ou de la ségrégation familiale, la composante génétique a été jusque là mesurée par l'information d'apparentement au sein des familles. Dans les années quatre-vingts, l'accès direct a l'information sur le génome des individus et l'élaboration de méthodes d'analyse appropriées a permis des gains de puissance très importants et une spécificité vis-à-vis du facteur génétique dont on arrive maintenant à localiser la position sur le génome et à mesurer l'effet sur le risque de la maladie.

1.3. Analyse de liaison génétique

L'objectif des analyses de liaison est de localiser la ou les régions chromosomiques où se trouvent des facteurs génétiques impliqués. Elles reposent sur l'observation au sein des familles d'une cotransmission des allèles au niveau d'un marqueur et de la maladie plus fréquente qu'attendue par la transmission Mendélienne. En effet, plus le locus marqueur est proche d'un locus de susceptibilité à la maladie situé sur le même chromosome et plus rares sont les phénomènes de recombinaison génétique entre ces deux locus. On dit alors que ces locus sont liés car la transmission de leurs allèles n'est pas indépendante. C'est l'élément essentiel qui différencie le déséquilibre gamétique de la liaison génétique.

En pratique, l'analyse de liaison d'une maladie complexe est réalisée sur quelques centaines de marqueurs localisés sur une carte génétique, afin d'identifier ceux qui sont liés à la maladie et donc de localiser une ou plusieurs régions contenant des facteurs génétiques de susceptibilité à la maladie. C'est en 1980 que la première proposition de création d'une carte de marqueurs génétiques pour effectuer des analyses de liaison sur tout le génome est faite (Botstein *et al.* 1980). La première carte pangénomique de liaison recensant près de 400 RFLPs est mise au point en 1987 (Donis-Keller *et al.* 1987) et est rapidement supplantée par des cartes d'environ 1000 microsatellites. On commence maintenant à utiliser des cartes de SNPs disponibles à l'aide d'algorithmes spécifiques permettant de sélectionner les SNPs en minimisant le déséquilibre gamétique (Bellenguez *et al.* 2009).

Le taux de recombinaison θ entre deux locus est la proportion de gamètes recombinés parmi l'ensemble des gamètes transmis par les parents (gamètes parentaux et gamètes recombinés). En l'absence de liaison génétique, la proportion de gamètes recombinés est égale à celle des gamètes parentaux et θ est égal à 1/2. En présence d'une liaison génétique, la proportion de gamètes recombinés est inférieure à celle des gamètes parentaux et θ est inférieur à 1/2. Plus les locus sont proches, plus θ tend vers 0. Ce taux de recombinaison est donc un bon indicateur de la distance entre le locus marqueur et le locus de susceptibilité. Le test du *LOD Score* développé par Morton (1955) permet de tester un écart à l'hypothèse d'indépendance de transmission (absence de liaison génétique) en comparant la vraisemblance maximisant le paramètre θ à celle imposant une valeur de $\theta = 1/2$. Dans le cadre d'une maladie monogénique, l'utilisation de cartes de microsatellites sur le génome a permis de réaliser les premiers criblages par analyse de liaison et ainsi de découvrir de nombreux gènes responsables de ces maladies. Cependant, les estimations de vraisemblance par cette méthode nécessitent de spécifier le modèle de pénétrance et les fréquences alléliques, ce qui constitue une contrainte importante dans les maladies complexes où les paramètres du facteur génétique sont le plus souvent inconnus (Clerget-Darpoux et Bonaïti-Pellié 1993).

Des méthodes dites « non paramétriques » ont alors été développées permettant de tester la liaison génétique sans spécifier le modèle génétique. Par exemple, le *Maximum Likelihood Score* (MLS) proposé par Risch (Risch 1990b) repose sur la mesure du nombre d'allèles reçus identiques par descendance (IBD, *Identical By Descent*) par des paires de germains (frères et/ou sœurs) atteints à un locus donné. On dit que deux allèles sont IBD

lorsqu'ils sont la copie d'un même allèle présent chez un ancêtre commun. Ce nombre d'allèles IBD peut donc être égal à 2 si les deux germains ont reçus les mêmes allèles de leurs deux parents, 1 s'ils ont reçu le même allèle d'un des parents et un allèle différent de l'autre parent ou 0 s'ils ont reçu des allèles différents des deux parents. En l'absence de liaison génétique entre le locus marqueur étudié et la maladie (indépendance de transmission), les proportions des paires attendues dans chacun des trois états IBD (2, 1, 0) sont respectivement égales à 1/4, 1/2 et 1/4. À chaque marqueur testé, le MLS estime la distribution IBD (π_2, π_1, π_0) qui maximise la vraisemblance de l'échantillon et la compare à la distribution attendue en l'absence de liaison génétique (1/4, 1/2, 1/4) par un rapport de vraisemblances. Un écart significatif à ces proportions suggère l'existence d'une liaison génétique entre la maladie et le locus étudié. Cette méthode permet ainsi de tester la liaison génétique sans spécifier les paramètres du modèle génétique sous-jacent et est donc appropriée aux maladies complexes.

Il existe d'autres méthodes qui utilisent l'IBD comme information de liaison génétique et nous présenterons notamment par la suite le *Mean Sharing Test* dont l'extension permet de prendre en compte des interactions G×E.

1.4. Étude d'association génétique

1.4.1. Échantillon d'individus non apparentés provenant de la population

Le principe des études d'association génétique consiste à comparer les distributions alléliques (ou génotypiques) au niveau d'un marqueur génétique de deux groupes de personnes non apparentées, des malades (les cas) et des non malades (les témoins). Si ces distributions sont significativement différentes, alors on conclut qu'il existe une association entre l'allèle (ou le génotype) le plus fréquent chez les cas et la maladie étudiée (Balding 2006). Contrairement à la liaison où l'allèle cotransmis peut être différent au sein des familles ou des paires de germains, l'association génétique mesure la corrélation d'un allèle (ou d'un génotype) en particulier avec la maladie. Les études d'association sont réalisées le plus souvent avec des SNPs car ces marqueurs sont dialléliques. Cela permet de limiter le nombre de catégories du facteur génétique étudié et par conséquent le degré de liberté des tests effectués, mais aussi d'éviter d'avoir de petits effectifs dans certaines catégories. Mais rien n'empêche d'étudier l'association avec des polymorphismes multialléliques ou des variants structuraux.

Que l'on veuille prendre en compte ou pas un facteur E ou une interaction G×E dans une étude cas-témoin, les méthodes de modélisation les plus utilisées en épidémiologie sont fondées sur la forme binomiale de la régression logistique, où la variable dépendante (la maladie M) est binaire et où les covariables indépendantes (X_i) sont observées chez tous les individus. La modélisation logistique de la probabilité de la maladie M sachant les covariables X_i est donnée par :

$$P\left(M=1\middle|X_i\right)=\frac{e^{\beta_0+\beta_i X_i}}{1+e^{\beta_0+\beta_i X_i}} \qquad (1.23)$$

où l'ordonnée à l'origine $\beta_0 = \ln (f_B / 1 - f_B)$, f_B est le risque de base de la maladie $P(M| X_i = 0)$ et $\beta_i = \ln (OR_i)$, OR_i est l'odds ratio de l'association du facteur X_i avec la maladie M.

En utilisant la fonction logit :

$$\text{logit } P=\ln\frac{P(M=1|X_i)}{1-P(M=1|X_i)} \qquad (1.24)$$

Le modèle s'écrit alors :

$$\text{logit } P = \beta_0 + \beta_i X_i \qquad (1.25)$$

A partir de ce modèle, il est possible d'estimer les paramètres du modèle par maximum de vraisemblance et de tester différentes hypothèses en utilisant le test du rapport de vraisemblances, le test de Wald ou le test du Score. Ces trois tests sont asymptotiquement équivalents.

Cependant, les études d'association en population peuvent souffrir de nombreux biais plus ou moins importants selon la rigueur avec laquelle les témoins ont été sélectionnés. Nous reviendrons plus amplement sur ce sujet dans la partie 3 de ce manuscrit, mais il est important de préciser que l'existence dans la population source d'une stratification ou d'un mélange de plusieurs populations hétérogènes d'un point de vue génétique et phénotypique peut entraîner des résultats d'association faussement positifs (Cardon et Palmer 2003).

Supposons par exemple que la population d'étude soit composée de sous-groupes avec des prévalences différentes de maladie. Ceux ayant une forte prévalence seront surreprésentés dans l'échantillon de cas tandis que ceux ayant une faible prévalence de la maladie seront sous-représentés. Si par ailleurs, tout à fait indépendamment de la maladie, un marqueur

génétique présente des variations de fréquences alléliques (ou génotypiques) entre ces groupes, on risque de conclure à tort qu'il y a une association entre la maladie et ce marqueur. Plusieurs méthodes de correction de ce biais ont été proposées, mais une alternative en amont de la conception de l'étude consiste à réaliser une étude d'association sur des données familiales.

1.4.2. Échantillon de trios cas-parents

Le *Transmission Disequilibrium Test* (TDT) est une méthode d'analyse génétique qui utilise un échantillon composé de trios où tous les individus sont génotypés (Spielman et Ewens 1996; Spielman *et al.* 1993). Un trio est composé d'un cas atteints de la maladie étudiée et de ses deux parents. Pour un marqueur génétique donné, le principe du TDT consiste à comparer la proportion de parents hétérozygotes ($M||m$) ayant transmis l'un des allèles (M) à la proportion de parents hétérozygotes ayant transmis l'autre allèle (m). En l'absence de liaison ou d'association entre le marqueur étudié et la maladie, la probabilité de transmission de chacun des allèles M et m par un parent hétérozygote est de 0,5. Si l'un des deux allèles est préférentiellement transmis, alors on conclut à la fois à la présence d'une association de l'allèle plus fréquemment transmis avec la maladie et à la liaison génétique du locus étudié avec la maladie. Ainsi le TDT teste simultanément l'association et la liaison génétique. La statistique de test consiste à réaliser un test de Mac Nemar sur le tableau de contingence croisant les allèles transmis et non transmis des parents hétérozygotes de l'échantillon. Si l'échantillon est composé de n_M et n_m parents hétérozygotes qui ont transmis l'allèle M et m respectivement, le TDT est égal à :

$$TDT = \frac{(n_M - n_m)^2}{n_M + n_m} \tag{1.26}$$

Ce test suit sous l'hypothèse nulle d'absence de liaison ou d'association un χ^2 à 1 degré de liberté (ddl). Contrairement aux études d'association en population, l'appariement de chaque cas au parent hétérozygote permet au TDT d'être robuste à la stratification de population.

A partir de ce test simple, plusieurs extensions ont été proposées afin notamment de mesurer l'association de l'allèle avec la maladie. Parmi ces extensions, nous présenterons par la suite la modélisation log-linéaire de données trios qui permet d'incorporer le facteur E et l'interaction G×E.

1.4.3. Stratégie gène candidat versus stratégie pangénomique

Jusqu'à la fin des années 1990, la stratégie d'étude d'une maladie complexe consistait à rechercher dans un premier temps des régions avec un « pic » de liaison génétique, puis à sélectionner dans ces régions des gènes candidats potentiels qui sont étudiés dans un second temps par une étude d'association.

L'approche gène candidat consiste à rechercher la présence d'une association génétique avec un nombre limité de marqueurs (environ 5 à 50 SNPs par gène) situés dans des régions codantes ou avoisinantes (au niveau de sites d'épissage ou de sites régulateurs par exemple) d'un ou de plusieurs gènes candidats. Les caractéristiques qui font qu'un gène est considéré être un bon candidat à une étude d'association avec une maladie donnée sont les suivantes : protéine codée ayant une fonction susceptible d'être impliquée dans la maladie, localisation physique dans une région précédemment suggérée par une étude pangénomique de liaison ou d'association, profil d'expression cellulaire en relation avec la maladie étudiée, homologie avec un gène identifié dans un modèle animal de la maladie, etc.

Avec la disponibilité récente de puces à ADN permettant de génotyper rapidement plus de 300 000 SNPs pour de grands échantillons d'individus, il est aujourd'hui possible de réaliser des études d'association pangénomiques (*Genome Wide Association Studies*). Les études du déséquilibre gamétique entre les SNPs du génome ont montré que l'on peut individualiser des « blocs » de fort déséquilibre gamétique, au sein desquels le génotypage d'un nombre restreints de SNPs (*tag-SNPs*) permet d'inférer avec peu d'ambigüité le génotype de tous les autres SNPs du bloc (Anderson et Novembre 2003; Daly *et al.* 2001; Gabriel *et al.* 2002). Il est ainsi possible de n'étudier que l'association de ces tag-SNPs avec la maladie pour « couvrir » l'ensemble des SNPs du génome. L'idée est d'utiliser cette propriété du déséquilibre gamétique pour tester l'association d'un variant causal avec la maladie de façon indirecte. En effet, l'association du variant causal sera indirectement détectée par l'association d'un (ou de plusieurs) SNP(s) du panel de tag-SNPs étudié qui est (sont) en fort déséquilibre gamétique avec ce variant causal.

2. Traitement des interactions gène-environnement

Indépendamment du type d'analyse choisi, deux stratégies complémentaires de l'étude des interactions G×E dans les maladies complexes peuvent être individualisées.

La première stratégie consiste à tester simultanément l'effet du facteur *G* et l'interaction G×E. Cette stratégie se fonde sur l'argument que de négliger une interaction existante avec un facteur *E* peut empêcher la détection du facteur *G*, comme nous l'avons observé dans la présentation des différents types d'interaction G×E. Que ce soit en liaison ou en association génétique, les quelques études comparatives de puissance réalisées ont montré que pour une grande variété de modèles d'interaction G×E, un test combiné du facteur *G* et de l'interaction G×E peut être plus puissant qu'un test de l'effet marginal de *G* (Andrieu *et al.* 2001; Dizier *et al.* 2003; Gauderman et Siegmund 2001; Kraft *et al.* 2007; Selinger-Leneman *et al.* 2003).

La seconde stratégie consiste à ne tester que l'interaction G×E seule. L'argument sur lequel se basent les partisans de cette approche est qu'après une étude d'association pangénomique qui teste l'effet du facteur *G* marginalement, l'étude isolée de l'interaction G×E est indépendante et permet d'identifier de nouveaux variants génétiques impliqués dans des interactions G×E pures (Murcray *et al.* 2009).

Quoi qu'il en soit, le traitement des interactions G×E consiste le plus souvent à étendre les différentes méthodes utilisées pour étudier le facteur *G* en stratifiant l'analyse en fonction du facteur *E* des sujets étudiés. Nous présenterons successivement dans ce paragraphe, les méthodes de prise en compte des interactions G×E fondées sur l'analyse de liaison génétique, sur les études d'association génétique en population et sur les études utilisant des données trios fondées à la fois sur la liaison et sur l'association génétique. Pour chacune de ces trois approches, nous présenterons les différents tests du facteur génétique seul (G), du facteur génétique et de l'interaction G×E conjointement (GI) et de l'interaction G×E seule (I).

2.1. Tests fondés sur la liaison génétique

Afin de prendre en compte une hétérogénéité de la liaison génétique en fonction d'un facteur *E*, l'échantillon de paires de germains atteints est stratifié en fonction de l'exposition

des deux germains, ce qui conduit à distinguer trois catégories : les paires de germains exposés (*EE*), les paires discordantes pour l'exposition avec un germain exposé et l'autre non (*EU*) et les paires de germains non exposés (*UU*) (Dizier *et al.* 2003; Gauderman et Siegmund 2001; Khoury *et al.* 1991). Ainsi l'extension du MLS afin de prendre en compte une interaction G×E consiste à sommer les statistiques MLS estimées dans chacune des trois catégories d'exposition des paires de germains. Ce sMLS suit sous l'hypothèse nulle un χ^2 à 6 ddl. Cependant, nous avons choisi d'utiliser comme méthode de référence dans l'étude de comparative de puissance (dans la Partie 3) le *Mean Interaction Test* qui découle du *Mean Sharing Test*.

Une alternative à l'estimation de la distribution IBD utilisée dans le MLS consiste à estimer la moyenne π des proportions de paires de germains aux trois statuts IBD possibles :

$$\pi = \sum_{i=0}^{2} i\,\pi_i \qquad (1.27)$$

En l'absence de liaison génétique, on s'attend à ce que π soit égal à 0,5. En présence d'une liaison génétique, π est supérieure à 0,5.

Le *Mean Sharing Test* (MST) tel que proposé initialement consiste à effectuer un test unilatéral *z* de la forme :

$$z = \frac{(\pi - 0{,}5)\sqrt{N}}{\sigma} \qquad (1.28)$$

où N est le nombre de paires de germains et σ est l'écart-type de π. Ce test z suit sous l'hypothèse nulle d'absence de liaison une loi normale centrée réduite (Blackwelder et Elston 1985).

Gauderman et Siegmund (2001) ont proposé une extension du MST pour prendre en compte une interaction avec un facteur *E*. Cette extension consiste à estimer la moyenne des proportions IBD (π) dans chacune des strates de paires de germains *EE*, *EU*, et *UU*.

La méthode est fondée sur le modèle de régression linéaire suivant :

$$\pi_i = \pi + \beta\,(X_i - E(X)) + \varepsilon_i \qquad (1.29)$$

où π est l'ordonnée à l'origine (moyenne de l'IBD combiné des trois catégories), β est le coefficient de régression du facteur d'exposition et X_i est la covariable d'exposition centrée sur sa moyenne $E(X)$.

Différents codages sont proposés pour X_i. Le premier consiste en deux covariables X_{EE} et X_{EU} contrastant les paires *EE* et *EU* par rapport aux paires *UU*. Le second modélise une tendance linéaire avec X_{lin} qui est respectivement égal à 1, 0,5 et 0 pour les paires *EE*, *EU* et *UU*. Le troisième codage contraste les paires *EE* par rapport aux paires non-*EE* et finalement, le dernier contraste les paires *UU* par rapport aux paires non-*UU*. À partir de ce modèle, trois tests peuvent être réalisés.

2.1.1. Test de liaison génétique (G)

L'hypothèse nulle d'absence de liaison génétique est testée par le rapport de vraisemblances :

$$T_G = -2 \ln [L(\pi = 0{,}5 ; \beta = 0) / L(\pi ; \beta = 0)] \quad (1.30)$$

où L indique la log-vraisemblance.

L'hypothèse alternative étant unilatérale ($\pi > 0{,}5$), la distribution sous l'hypothèse nulle est un mélange d'un χ^2 à 0 ddl et d'un χ^2 à 1 ddl. Ce test est asymptotiquement équivalent à celui proposé initialement dans le MST (équation 1.28).

2.1.2. Test conjoint de la liaison génétique et de l'interaction G×E (GI)

Il s'agit du MIT proprement dit tel que proposé par Gauderman et Siegmund (2001). L'hypothèse nulle d'absence de liaison et d'interaction G×E est testée par le rapport de vraisemblances :

$$T_{GI} = -2 \ln [L(\pi = 0{,}5 ; \beta = 0) / L(\pi ; \beta)] \quad (1.31)$$

L'hypothèse alternative étant unilatérale pour $\pi > 0{,}5$ et bilatérale pour $\beta \neq 0$, la distribution sous l'hypothèse nulle est un mélange d'un χ^2 à p ddl et d'un χ^2 à $p + 1$ ddl, où p est le nombre de coefficients de régression du facteur d'exposition.

2.1.3. Test de l'interaction G×E en présence d'une liaison génétique (I)

L'hypothèse nulle d'absence d'interaction G×E est testée par le rapport de vraisemblances :

$$T_I = -2 \ln [L(\pi ; \beta = 0) / L(\pi ; \beta)] \quad (1.32)$$

L'hypothèse alternative étant bilatérale pour $\beta \neq 0$, la distribution sous l'hypothèse nulle est un χ^2 à p ddl, où p est le nombre de coefficients de régression du facteur d'exposition. Notons qu'ici pour mettre en évidence une interaction G×E, la présence d'une liaison génétique est indispensable car ce test mesure l'hétérogénéité de la liaison génétique entre les trois groupes.

Outre le fait qu'elles nécessitent de recruter des paires de germains atteints qu'il faut génotyper tous deux, ces méthodes ont l'inconvénient d'être assez peu puissantes pour détecter des effets modérés ou faibles. Par exemple, Gauderman et Siegmund (2001) ont comparé la puissance du MIT à celle du MST qui ne prend pas en compte l'interaction G×E. Ils ont montré que la prise en compte de l'interaction G×E augmentait la puissance de détection de la liaison génétique pour des coefficients d'interaction G×E supérieurs à 3 ou inférieur à 1/3. Autrement, la stratification de l'échantillon en 2 ou 3 catégories entraîne une perte de puissance du MIT par rapport au MST. Par ailleurs, ces méthodes ne permettent de tester ni l'effet d'un allèle spécifique, ni l'effet d'un facteur *E* seul sur la susceptibilité à la maladie, ni l'interaction G×E de façon isolée (sans liaison génétique), mais il est en revanche possible de conclure à la liaison génétique entre le locus étudié et la maladie en prenant en compte une hétérogénéité de liaison en fonction du facteur *E*.

2.2. Tests fondés sur l'association génétique

Si nous reprenons le modèle que nous avons introduit au chapitre 2. La maladie *M* est la variable dépendante que nous souhaitons expliquer par les deux facteurs de risque *G* et *E* et par leur interaction G×E. À partir de ces facteurs, différents modèles de régression logistique sont possibles (équation 1.25, page 59) selon que l'on utilise (ou que l'on dispose) ou pas (de) l'information sur l'exposition de tous les individus de l'étude.

2.2.1. Tests d'association génétique (G)

Test de l'effet marginal du facteur génétique

Dans la situation où l'on ne s'intéresse qu'à l'effet du facteur *G* sans prendre en compte l'exposition à *E*, le modèle logistique s'écrit :

$$\text{logit } P = \beta_{0M} + \beta_{GM} G \qquad (1.33)$$

où β_{GM} = ln (OR_{GM}), OR_{GM} est une mesure marginale de l'association du facteur G avec la maladie. Ce modèle peut être utilisé lorsque l'on ne dispose pas d'information sur le facteur E ou bien si l'on est sûr que ce facteur E n'a aucun effet confondant avec la variable G.

L'hypothèse nulle d'absence d'association entre G et M ($\beta_{GM} = 0$) est testée par un test de rapport de vraisemblances qui suit sous cette hypothèse un χ^2 à 1 ddl :

$$T_{GM} = -2 \ln [L(\beta_{0M} ; \beta_{GM} = 0) / L(\beta_{0M} ; \beta_{GM})] \quad (1.34)$$

Test de l'effet du facteur génétique ajusté sur l'environnement

Il est possible lorsque l'on dispose de l'information sur l'exposition des cas et des témoins de tenir compte de ce facteur en le modélisant de la façon suivante :

$$\text{logit } P = \beta_{0A} + \beta_{EA} E + \beta_{GA} G \quad (1.35)$$

où β_{GA} = ln (OR_{GA}). OR_{GA} mesure l'association du facteur G avec la maladie en ajustant sur le facteur E. Ce genre de modélisation est utile lorsque le facteur E est à la fois associé à la maladie et au facteur G. Dans ce cas, ne pas le prendre en compte risque d'entraîner un biais de confusion dans l'estimation marginale du facteur G.

L'hypothèse nulle d'absence d'association entre G et M ($\beta_{GA} = 0$) est testée par un test de rapport de vraisemblances qui suit sous cette hypothèse un χ^2 à 1 ddl :

$$T_{GA} = -2 \ln [L(\beta_{0A} ; \beta_{EA} ; \beta_{GA} = 0) / L(\beta_{0A} ; \beta_{EA} ; \beta_{GA})] \quad (1.36)$$

2.2.2. Test conjoint d'association génétique et de l'interaction G×E (GI)

Le modèle complet (ou modèle saturé) incluant les deux facteurs G et E et l'interaction G×E permet de réaliser différents tests. Ce modèle qui nécessite de connaître les génotypes et les statuts d'exposition de tous les individus s'écrit :

$$\text{logit } P(M = 1 | G, E) = \beta_{0CT} + \beta_{ECT} E + \beta_{GCT} G + \beta_{ICT} GE \quad (1.37)$$

où β_{ECT} = ln (OR_{ECT}), β_{GCT} = ln (OR_{GCT}) et β_{ICT} = ln (OR_{ICT}). OR_{ECT} et OR_{GCT} mesurent respectivement l'association des facteurs E et G avec la maladie chez des individus non porteur de G et non exposés à E. OR_{ICT} mesure l'odds ratio multiplicatif d'interaction G×E. Ce modèle est celui qui s'adapte le mieux au modèle général que nous avons exposé au chapitre 2.

Kraft *et al.* (2007) proposent de tester conjointement l'association génétique et l'interaction G×E en réalisant le test de rapport de vraisemblances qui suit sous l'hypothèse nulle ($OR_{GCT} = OR_{ICT} = 1$) un χ^2 à 2 ddl :

$$T_{GI} = -2 \ln [L(\beta_{0CT} ; \beta_{ECT} ; \beta_{GCT} = \beta_{ICT} = 0) / L(\beta_{0CT} ; \beta_{ECT} ; \beta_{GCT} ; \beta_{ICT})] \quad (1.38)$$

2.2.3. Tests d'interaction gène-environnement

Test d'interaction à partir d'un échantillonnage cas-témoins

Une seconde possibilité à partir du modèle logistique complet (équation 1.37) est de ne tester que l'interaction G×E par un test du rapport de vraisemblances qui suit sous l'hypothèse nulle ($\beta_{ICT} = 0$) un χ^2 à 1 ddl :

$$T_{ICT} = -2 \ln [L(\beta_{0CT} ; \beta_{ECT} ; \beta_{GCT} ; \beta_{ICT} = 0) / L(\beta_{0CT} ; \beta_{ECT} ; \beta_{GCT} ; \beta_{ICT})] \quad (1.39)$$

Test d'interaction à partir d'un échantillonnage cas-seuls

Piegorsch *et al.* (1994) proposent le schéma de recrutement cas-seuls (traduction littérale de l'anglais *case-only design*) comme alternative plus puissante pour tester l'interaction G×E en n'utilisant qu'un échantillon de cas (Yang et Khoury 1997). Les malades non exposés au facteur *E* sont considérés comme des pseudo-témoins et ceux exposés au facteur *E* comme des pseudo-malades. Les distributions génotypiques (ou alléliques) des deux groupes sont ensuite comparées. Le modèle logistique consiste à considérer que l'exposition est la variable dépendante et le génotype la variable explicative. En supposant une maladie rare et l'absence de corrélation gène-environnement, le coefficient de régression associé à la variable *G* correspond à l'estimateur de l'interaction G×E. Si l'odds ratio obtenu est significativement supérieur à 1, on conclut à l'existence d'une interaction G×E. Le modèle logistique correspondant s'écrit :

$$\text{logit } P(E = 1 | G) = \beta_{0CS} + \beta_{ICS} G \quad (1.40)$$

Un test du rapport de vraisemblances qui suit sous l'hypothèse nulle ($\beta_{ICS} = 0$) un χ^2 à 1 ddl permet de tester l'interaction G×E :

$$T_{ICS} = -2 \ln [L(\beta_{0CS} ; \beta_{ICS} = 0) / L(\beta_{0CS} ; \beta_{ICS})] \quad (1.41)$$

Même si le test avec un échantillonnage de cas seuls est réputé pour sa puissance supérieure à tous les autres tests d'interaction G×E, l'inconvénient de ce type d'étude est l'impossibilité d'estimer les effets indépendants des deux facteurs *G* et *E*. Par ailleurs il repose sur l'hypothèse d'indépendance entre *G* et *E* qui ne peut être vérifiée qu'en disposant de témoins pour lesquels l'exposition au facteur *E* et les génotypes sont disponibles.

2.3. Tests fondés sur l'association et la liaison génétique

Il est possible d'étudier les effets conjoints de deux facteurs *G* et *E* avec un échantillonnage de trios cas-parents. Pour cela, plusieurs auteurs ont proposé d'étudier les interactions G×E en groupant les trios en fonction du statut d'exposition du cas et en testant la différence de transmission des allèles au marqueur entre les cas exposés et non exposés. Ces propositions peuvent être regroupées en deux catégories. Certaines approches consistent à étendre le TDT en utilisant comme critère de comparaison le taux de transmission d'un allèle donné des parents hétérozygotes aux cas (Harley *et al.* 1995; Maestri *et al.* 1997; Schaid 1999; Thomas 2000). En effet, si les fréquences de transmissions des deux allèles au marqueur génétique diffèrent entre les cas exposés et les cas non exposés, alors cela suggère la présence d'une interaction G×E. Ces méthodes ne permettent pas de tenir compte de la non-indépendance de transmission des couples de parents hétérozygotes et elles sont par ailleurs sensibles à la présence d'une corrélation gène-environnement. D'autres méthodes utilisent la distribution des génotypes des cas conditionnée sur le génotype des parents permettant ainsi de résoudre ces deux problèmes (Khoury 1994; Umbach et Weinberg 2000; Witte *et al.* 1999).

Nous allons présenter ici l'approche proposée par Umbach et Weinberg (2000) qui consiste à comparer les distributions génotypiques des cas exposés et non exposés conditionnées sur les génotypes des parents, en utilisant un modèle log-linéaire. Les génotypes de la mère, du père et de l'enfant sont notés *M*, *P* et *C* et prennent les valeurs 0, 1 ou 2 correspondant au nombre d'allèle *M* au marqueur génétique. En faisant abstraction de l'ordre des parents, six couples parentaux (*M*,*P*) sont possibles : (2,2), (2,1), (2,0), (1,1), (1,0) et (0,0) que nous indiquerons par l'indice *i* variant de 1 à 6. Les trios peuvent alors être subdivisés en 20 catégories en fonction des six catégories de couples parentaux, du génotype

de l'enfant atteint et de son statut d'exposition. Les probabilités attendues dans chacune des catégories de trios peuvent être exprimées à partir d'un modèle log-linéaire :

$$\ln\left(P\left(M=1\middle|i,G,E\right)\right)=\beta_{0i}+\beta_{Ei}E+\beta_G G+\beta_I GE+\ln(2)I_{\{(M,P,C)=(1,1,1)\}} \qquad (1.42)$$

Les six paramètres β_{0i} correspondent aux paramètres de strate des six couples parentaux possibles dont l'enfant n'est pas exposé. Les six valeurs (β_{0i} + β_{Ei}) correspondent à ces paramètres pour les couples parentaux dont l'enfant est exposé. Le paramètre β_G est égal au logarithme du risque relatif associé au facteur *G* chez les enfants non exposés (équation 1.14, page 44) et le paramètre β_I est égal au logarithme du coefficient d'interaction G×E (équation 1.17, page 44). $I_{\{(M,P,C)=(1,1,1)\}}$ est une fonction indicatrice égale à 1 quand (*M*,*P*,C) = (1,1,1) et 0 sinon. La constante ln(2) qui est ajoutée à la probabilité de la catégorie (*M*,*P*,*C*) = (1,1,1) est un *offset* qui contraint la probabilité que des parents hétérozygotes *M*||*m* aient un enfant hétérozygote *M*||*m* à être le double de leur probabilité d'avoir un enfant homozygote *M*||*M* ou d'avoir un enfant homozygote *m*||*m*. À partir de ce modèle, différents tests de rapport de vraisemblances peuvent être réalisés.

2.3.1. Test de l'effet du facteur génétique (G)

L'hypothèse nulle d'absence d'effet du facteur *G* ($RR_G = 0$) est testée par le rapport de la vraisemblance du modèle complet de l'équation 1.42 et de la vraisemblance du modèle contraignant β_G à être égal à 0, en imposant l'absence d'interaction dans les deux modèles :

$$T_G = -2 \ln [L (\beta_{0i} ; \beta_{Ei} ; \beta_G = 0 ; \beta_I = 0) / L (\beta_{0i} ; \beta_{Ei} ; \beta_G ; \beta_I = 0)] \qquad (1.43)$$

T_G suit sous l'hypothèse nulle un χ^2 à 1 ddl.

2.3.2. Test conjoint de l'effet du facteur génétique et de l'interaction G×E (GI)

L'hypothèse nulle d'absence d'effet du facteur *G* et d'absence d'interaction G×E ($RR_G = RR_I = 0$) est testée par le rapport de la vraisemblance du modèle complet de l'équation 1.42 et de la vraisemblance d'un modèle contraignant β_G et β_I à être égaux à 0 :

$$T_{GI} = -2 \ln [L (\beta_{0i} ; \beta_{Ei} ; \beta_G = 0 ; \beta_I = 0) / L (\beta_{0i} ; \beta_{Ei} ; \beta_G ; \beta_I)] \qquad (1.44)$$

T_{GI} suit sous l'hypothèse nulle un χ^2 à 2 ddl.

2.3.3. Test de l'interaction G×E (I)

Finalement, l'hypothèse nulle d'absence d'interaction G×E ($RR_I = 0$) est testée par le rapport de la vraisemblance du modèle complet de l'équation 1.42 et de la vraisemblance d'un modèle contraignant β_I à être égal à 0 :

$$T_I = -2 \ln [L(\beta_{0i} ; \beta_{Ei} ; \beta_G ; \beta_I = 0) / L(\beta_{0i} ; \beta_{Ei} ; \beta_G ; \beta_I)] \quad (1.45)$$

T_I suit sous l'hypothèse nulle un χ^2 à 1 ddl.

Comparativement à une étude d'association sur des données cas-témoins, les approches utilisant des trios permettent d'obtenir des puissances équivalentes en génotypant autant de trios cas-parents que de paires cas-témoin. Ce coût supplémentaire est contrebalancé par la robustesse vis-à-vis de la stratification de population. Cependant, ce type de recrutement n'est pas adapté aux maladies à début tardif pour lesquelles généralement les parents des cas sont décédés. Par ailleurs, ce type d'approche ne permet pas d'estimer, ni de tester l'effet du facteur environnemental.

Par rapport à un TDT stratifié sur l'exposition du cas où l'unité d'étude est l'allèle transmis par un parent hétérozygote, la modélisation log-linéaire a pour unité d'étude le trio au complet. En conditionnant sur les génotypes parentaux, on s'affranchit du problème de la non-indépendance de transmission que l'on a dans la situation de deux parents hétérozygotes. En effet, dans ce cas, Umbach et Weinberg (2000) montrent que la probabilité que la mère transmette l'allèle à risque n'est pas la même sachant que le père l'a transmis ou pas. Par rapport à la modélisation log-linéaire, un autre problème du TDT stratifié est qu'il n'est pas robuste à la présence d'une corrélation gène-environnement dans la population.

Partie 2

EURECA,
Sélection des facteurs environnementaux

Chapitre 1
EURECA, une méthode de contraste de la récurrence familiale

L'une des problématiques inhérente à l'étude des interactions G×E dans les maladies complexes est le choix de la variable environnementale *E* avec laquelle l'interaction est testée ou prise en compte dans l'analyse. L'approche naïve, consistant à ne s'intéresser qu'aux facteurs de risque *E* pour lesquels un effet est documenté, n'est pas nécessairement la meilleure. En effet, dans la partie précédente nous avons montré que pour certains modèles d'interaction G×E, une mesure marginale de l'effet d'un facteur de risque ne permet pas de préjuger *a priori* de la présence ou non d'une interaction G×E. Nous proposons dans ce chapitre un nouvel outil statistique, EURECA (pour *Exposed versus Unexposed Recurrence Analysis*), permettant de sélectionner les facteurs *E* les plus intéressants à prendre en compte. Nous présentons son principe et ses propriétés statistiques pour différents modèles d'interaction G×E.

L'ensemble des travaux de cette partie ont fait l'objet d'une publication qui se trouve en annexe 1 (Kazma *et al.* 2010).

1. Problématique

Même si l'importance de prendre en compte les facteurs *E* et les interactions G×E dans les études génétiques a été de nombreuses fois signalée, la majorité des études épidémiologiques récentes sur les maladies complexes évaluent ces deux types de facteurs de façon indépendante plutôt que de façon concomitante. L'analyse conjointe n'est généralement abordée que dans une seconde étape, lorsque les facteurs de risque associés de façon indépendante ont été clairement identifiés. L'utilisation de cette stratégie a montré ses limites notamment lorsque des associations génétiques qui ne sont pas observables marginalement le deviennent lorsque l'on prend en compte une interaction avec un facteur *E*. En particulier, la puissance de détection d'un facteur de risque *G* interagissant avec un facteur de risque *E* peut être considérablement réduite lorsque l'exposition à *E* des individus n'est pas prise en compte (Selinger-Leneman *et al.* 2003). Cependant, ce phénomène est très dépendant de la fréquence du facteur *E*, de la mesure de son effet indépendant sur la maladie et de la mesure de

l'interaction avec le facteur *G*. Dans certaines situations, prendre en compte un facteur *E* peut même avoir un effet néfaste sur la puissance de détection du facteur *G*. En effet, pour tenir compte des interactions, on stratifie généralement l'échantillon en fonction de l'exposition des malades, ce qui dans certaines situations peut conduire à une diminution de la puissance du test, en particulier si la fréquence du facteur environnemental est forte (Leal et Ott 2000).

Ces résultats soulignent donc l'intérêt, avant toute recherche de facteurs génétiques de susceptibilité, de l'identification des facteurs *E* qui interagissent avec ces facteurs *G* et dont la prise en compte peut être nécessaire à leur mise en évidence. Une sélection fondée uniquement sur la mesure marginale des effets des facteurs *E* n'est pas une stratégie efficace d'où l'utilité qu'il y aurait à développer un outil statistique pour identifier ceux qui sont susceptibles d'interagir avec des facteurs de risque *G*.

Discriminer ces facteurs *E* devient encore plus capital lorsque des études d'association pangénomiques sont réalisées avec des centaines de milliers de marqueurs génétiques. La prise en compte de multiples facteurs *E* oblige à effectuer des corrections drastiques pour contrôler le risque de faux positifs lié à la multiplicité des tests effectués. Les tailles d'échantillon nécessaires deviennent alors rapidement irréalistes. Le développement de méthodes pour sélectionner les facteurs *E* pertinents s'avère donc aussi être une étape préalable indispensable avant de tester des interactions G×E à l'échelle pangénomique. Par ailleurs, l'utilisation de l'information familiale en substitut des données génotypiques présente l'avantage de permettre une sélection de ces facteurs au préalable de la réalisation de l'étude pangénomique.

Une méthode fondée sur la modélisation par analyse de la variance d'un échantillon de jumeaux a été proposée par Purcell (Purcell 2002) pour étudier les interactions G×E sans utiliser l'information génotypique de l'échantillon. Une première contrainte de cette méthode est la difficulté d'obtenir des tailles d'échantillon suffisamment grandes en recrutant des paires de jumeaux. Par ailleurs, elle nécessite également de disposer d'une mesure de l'exposition des deux jumeaux de la paire. Or, dans la majorité des études épidémiologiques utilisant des informations familiales, l'histoire de la maladie est documentée chez les sujets apparentés alors que le facteur d'exposition *E* n'est généralement recueilli que chez l'index atteint et assez rarement chez les autres membres de sa famille.

La méthode que nous proposons permet dans le cadre plus large d'un échantillon de paires de germains (frères et/ou sœurs) recrutées sur la base d'un index atteint d'analyser les interactions G×E lorsque le facteur *E* n'est documenté que chez cet index.

2. Principe de la méthode EURECA

EURECA emploie d'une part l'exposition *E* mesurée chez un individu atteint, que nous appellerons l'index (noté S1) et d'autre part l'information de récurrence familiale de la maladie *M* chez son germain (noté S2) comme substitut de l'information génotypique. La méthode est fondée sur l'observation suivante que si le facteur environnemental étudié est impliqué dans une interaction G×E, alors on s'attend à trouver des différences de récurrence familiale de la maladie entre des index exposés et non exposés (Stücker *et al.* 1993). L'explication rationnelle de cette propriété est qu'en présence d'une interaction G×E, les index exposés et non exposés ont des distributions génotypiques différentes. Par exemple si l'interaction va dans le sens d'une augmentation de risque de la maladie (interaction synergétique), un génotype interagissant avec ce facteur d'exposition aura une fréquence plus élevée dans le groupe des index exposés que dans celui des non exposés. Les germains auront par conséquent des distributions génotypiques similaires et donc des probabilités différentes de présenter la maladie. Le principe d'EURECA consiste à utiliser cette différence observée entre les risques de récurrence des index exposés et non exposés pour sélectionner les facteurs *E* qui sont les plus probablement impliqués dans une interaction G×E dans la maladie dont on observe la récurrence.

Afin de mettre en évidence une différence entre les risques de récurrence de la maladie chez les germains d'index exposés et non exposés, il est nécessaire d'échantillonner des paires de germains (S1 − S2). Les données peuvent être présentées dans un tableau de contingence (tableau 2.1). La variable d'intérêt à expliquer est le statut maladie du germain S2 et la variable indépendante explicative, le statut d'exposition de l'index S1.

3. L'odds ratio de récurrence

Soit K_S le risque de récurrence défini comme étant la probabilité que S2 soit atteint sachant que S1 est atteint (Risch 1990a). K_{SE} et $K_{S\bar{E}}$ correspondent à ces risques lorsque S1 est respectivement exposé et non exposé au facteur *E*.

		Index S1 (atteint)		
		exposé E (S1) = 1	non exposé E (S1) = 0	
Germain S2	atteint M (S2) = 1	a	b	$a + b$
	non atteint M (S2) = 0	c	d	$c + d$
		$a + c$	$b + d$	N

Tableau 2.1 Distribution de l'échantillon de paires de germains

Les risques de récurrence de l'ensemble de l'échantillon (K_S) et dans chacune des strates d'index exposés (K_{SE}) et non exposés ($K_{S\bar{E}}$) peuvent être calculés à partir des effectifs du tableau de contingence :

$$K_S = (a + b) / \mathrm{N} \tag{2.1}$$

$$K_{SE} = a / (a + c) \tag{2.2}$$

$$K_{S\bar{E}} = b / (b + d) \tag{2.3}$$

Afin de mesurer la différence entre ces deux risques de récurrence, un odds ratio de récurrence (*ORR*) peut être calculé de façon analogue au calcul de l'odds ratio d'association entre un facteur et une maladie à partir des risques stratifiés :

$$ORR = \frac{K_{SE} \times \left(1 - K_{S\bar{E}}\right)}{K_{S\bar{E}} \times \left(1 - K_{SE}\right)} \tag{2.4}$$

En utilisant les risques calculés à partir des effectifs du tableau de contingence (équations 2.2 et 2.3), l'*ORR* peut être exprimé par :

$$ORR = \frac{ad}{bc} \tag{2.5}$$

Habituellement, pour l'odds ratio mesurant l'association entre un facteur de risque et une maladie dans une étude cas-témoin, les deux variables dont on évalue l'association proviennent du même individu. La particularité de l'*ORR* est que l'exposition est ici mesurée chez l'index S1 et le statut maladie chez son germain S2.

4. Corrélation de l'environnement chez les germains

Le principe de la méthode repose sur l'interprétation faite que la différence observée entre les risques de récurrence stratifiés sur le statut d'exposition de S1 est due à une interaction G×E. Cependant, il est possible d'observer une différence entre les deux strates de S1 en l'absence d'interaction G×E lorsque la distribution du facteur *E* présente une corrélation entre les germains et que ce facteur a un effet sur la maladie. C'est pourquoi, nous avons également modélisé une agrégation familiale du facteur *E*, en utilisant les formules proposées par Khoury *et al.* (1988b) où la probabilité d'exposition au facteur *E* de S2 est définie conditionnellement au statut d'exposition de S1 :

$$P(E\,(S2) = 1 \mid E\,(S1)) = (1 - C_E) \times f_E + C_E \times E\,(S1) \qquad (2.6)$$

où E (S1) est égal à 1 si S1 est exposé et 0 sinon. La valeur de C_E mesure de la corrélation environnementale de la paire de germains. Lorsque $C_E = 0$, le statut d'exposition de S2 est indépendant de celui de S1 et sa probabilité est égale à la fréquence de *E* dans la population générale, f_E. Lorsque $C_E = 1$, la corrélation est complète et le statut d'exposition de S2 est identique à celui de S1.

5. Modélisation de l'interaction gène-environnement

Même si la méthode n'utilise en pratique que l'information *E* de S1 et ne nécessite aucun génotypage ; afin d'étudier le comportement de l'*ORR*, de construire le test statistique EURECA et d'en étudier les propriétés dans différentes situations, nous avons modélisé l'interaction G×E sous-jacente à partir des paramètres définis dans le tableau 2.2. Ces paramètres sont identiques à ceux d'une modélisation de l'interaction G×E par des risques relatifs présentée dans la partie 1 (page 44). Afin de simplifier la présentation des résultats, nous avons choisi de ne présenter ici qu'un modèle génétique dominant et nous discuterons les différences constatées avec un modèle récessif, dont les résultats se trouvent dans l'article en annexe 1 (Kazma *et al.* 2010).

		Génotype	
		$s\|\|s$ $g = 0$ $(1 - q)^2$	$S\|\|s$ ou $S\|\|S$ $g = 1$ ou 2 $q^2 + 2q\,(1 - q)$
Exposition	$E = 0$ $1–f_E$	f_B	$f_B\,RR_G$
	$E = 1$ f_E	$f_B\,RR_E$	$f_B\,RR_G\,RR_E\,RR_I$

Tableau 2.2 Probabilité de maladie d'un individu sachant son statut d'exposition et son génotype en fonction des paramètres du modèle d'interaction gène-environnement, $P\,(D = 1 \mid g = i, E = j)$

S : allèle conférant une susceptibilité à la maladie ; g : nombre d'allèles S dans le génotype (0, 1 ou 2) ; q : fréquence de l'allèle S ; f_E : fréquence de l'exposition ; f_B : risque de base de la maladie ; RR_E : risque relatif environnemental; RR_G : risque relatif génétique (modèle génétique dominant) ; RR_I : coefficient multiplicatif d'interaction G×E

6. Calcul des effectifs attendus du tableau de contingence

Un individu est porteur de l'un des trois génotypes possibles ($S||S$, $S||s$ ou $s||s$) et est exposé ou non exposé au facteur E. Ainsi, une paire de germains présente 36 combinaisons possibles de génotypes et de statuts d'exposition. La méthode matricielle ITO de Li et Sacks (1954) permet d'obtenir les probabilités conjointes des génotypes d'une paire de germains (tableau 2.3).

En utilisant le tableau 2.2, le tableau 2.3 et l'équation 2.6, le calcul des probabilités conjointes d'observer un index S1 malade et exposé (ou non exposé) et un germain S2 malade (ou non malade) se fait de la façon suivante :

$$\begin{aligned} P_a &= P\,(M\,(\mathrm{S1}) = 1, E\,(\mathrm{S1}) = 1, M\,(\mathrm{S2}) = 1) \\ &= \sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{k=0}^{1}\left[\begin{array}{l} U_{ij} \times P\big(E(S1)=1\big)P\big(E(S2)=k\,\big|E(S1)=1\big) \times \\ \quad P\big(D=1\big|g(S1)=i, E(S1)=1\big) \times P\big(D=1\big|g(S2)=j, E(S2)=k\big) \end{array}\right] \end{aligned} \quad (2.7)$$

$$P_b = P(M(\mathrm{S1})=1, E(\mathrm{S1})=0, M(\mathrm{S2})=1)$$
$$= \sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{k=0}^{1}\left[\begin{array}{l} U_{ij} \times P(E(S1)=0)P(E(S2)=k|E(S1)=0) \times \\ \quad P(D=1|g(S1)=i, E(S1)=0) \times P(D=1|g(S2)=j, E(S2)=k) \end{array}\right] \quad (2.8)$$

$$P_c = P(M(\mathrm{S1})=1, E(\mathrm{S1})=1, M(\mathrm{S2})=0)$$
$$= \sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{k=0}^{1}\left[\begin{array}{l} U_{ij} \times P(E(S1)=1)P(E(S2)=k|E(S1)=1) \times \\ \quad P(D=1|g(S1)=i, E(S1)=1) \times P(D=0|g(S2)=j, E(S2)=k) \end{array}\right] \quad (2.9)$$

$$P_d = P(M(\mathrm{S1})=1, E(\mathrm{S1})=0, M(\mathrm{S2})=0)$$
$$= \sum_{i=0}^{2}\sum_{j=0}^{2}\sum_{k=0}^{1}\left[\begin{array}{l} U_{ij} \times P(E(S1)=0), P(E(S2)=k|E(S1)=0) \times \\ \quad P(D=1|g(S1)=i, E(S1)=0) \times P(D=0|g(S2)=j, E(S2)=k) \end{array}\right] \quad (2.10)$$

"*i*" représente les génotypes possibles de S1, "*j*" représente les génotypes possibles de S2, "*k*" représente les statuts d'exposition possibles de S2 et la matrice U_{ij} fait référence à la matrice ITO du tableau 2.3.

		Germain S1			
		$S\|\|S$	$S\|\|s$	$s\|\|s$	
		$i=2$	$i=1$	$i=0$	
Germain S2	$S\|\|S$ $j=2$	$1/4\,q^2(1+q)^2$	$1/2\,q^2(1-q^2)$	$1/4\,q^2(1-q)^2$	q^2
	$S\|\|s$ $j=1$	$1/2\,q^2(1-q^2)$	$q(1-q)(1+q(1-q))$	$1/2\,q(1-q)^2(2-q)$	$2q(1-q)$
	$s\|\|s$ $j=0$	$1/4\,q^2(1-q)^2$	$1/2\,q(1-q)^2(2-q)$	$1/4(1-q)^2(2-q)^2$	$(1-q)^2$
		q^2	$2q(1-q)$	$(1-q)^2$	1

Tableau 2.3 Matrice U_{ij} des probabilités de la distribution des génotypes de paires de germains d'après la méthode matricielle ITO de Li and Sacks (Li and Sacks 1954)

S : allèle conférant une susceptibilité à la maladie; *q* : fréquence de l'allèle *S* dans la population

La somme P_T de ces quatre probabilités est égale à la probabilité *a priori* de maladie chez S1 :

$$P_T = P\,(M\,(\text{S1})) = P_a + P_b + P_c + P_d \qquad (2.11)$$

Finalement, en utilisant les équations 2.7 à 2.11, les effectifs attendus des différentes cases du tableau de contingence sont :

$$a = N \times P_a / P_T \qquad (2.12)$$

$$b = N \times P_b / P_T \qquad (2.13)$$

$$c = N \times P_c / P_T \qquad (2.14)$$

$$d = N \times P_d / P_T \qquad (2.15)$$

où N est le nombre total de paires de germains.

Le calcul de ces effectifs attendus a permis d'obtenir les différents risques de récurrence (tableau 2.1) et de calculer l'*ORR* attendu sous différents modèles d'interaction G×E.

7. Variations de l'odds ratio de récurrence

Afin d'évaluer la pertinence de l'utilisation de l'*ORR* comme indicateur de la présence d'une interaction G×E, nous avons étudié les variations de l'*ORR* sous différents modèles, tout d'abord en l'absence de corrélation environnementale chez les germains (figure 2.1). Lorsque les valeurs de l'interaction G×E (RR_I) augmentent, une augmentation des valeurs d'*ORR* est constatée, mais cette augmentation dépend également des valeurs des autres paramètres du modèle. L'impact de ces paramètres est représenté dans la figure 2.1 en fonction des fréquences des facteurs génétique et environnemental (f_G et f_E) pour différentes combinaisons de risques relatifs génétique et environnemental (RR_G et RR_E). Pour une valeur donnée de RR_I, on peut individualiser deux situations. En l'absence d'un effet du facteur génétique (RR_G = 1), l'*ORR* est plus grand pour des valeurs élevées de RR_E, de f_E et de f_G (courbes grises et vertes). En présence d'un effet du facteur génétique (RR_G = 5), l'*ORR* est plus grand pour des valeurs élevées de RR_E et de f_E et pour des valeurs faibles de f_G (courbes jaune et rouge). L'*ORR* est globalement plus élevé pour un modèle dominant que pour un modèle récessif pour une valeur de f_G donnée (annexe 1, figures 1 et 2).

Puisque l'existence d'une corrélation de E chez les germains peut entraîner une confusion potentielle avec une interaction G×E, nous avons étudié les variations de l'*ORR* pour différentes valeurs de C_E, en présence d'une interaction G×E ($RR_I = 5$) et en l'absence d'interaction G×E ($RR_I = 1$) (figure 2.2). Sous l'hypothèse nulle ($RR_I = 1$), la valeur de l'*ORR* (que nous appellerons dans cette situation ORR_0) est toujours égale à 1 en l'absence de corrélation ($C_E = 0$) ou en l'absence d'effet du facteur E ($RR_E = 1$). Par contre, en présence simultanée d'un effet ($RR_E \neq 1$) et d'une corrélation ($C_E \neq 0$) du facteur E, les valeurs d'ORR_0 sont augmentées. Ainsi, afin de construire un test de comparaison qui soit valide, les estimations d'*ORR* obtenues avec une interaction G×E et en présence d'une corrélation de E chez les germains doivent être comparées à la valeur de l'ORR_0 pour une même corrélation de E chez les germains, plutôt qu'à une valeur théorique de l'ORR_0 de 1. En prenant sur la figure 2.2 l'exemple d'un effet du facteur E, $RR_E = 2$ et d'une corrélation, $C_E = 0{,}5$, la valeur observée de l'*ORR* en présence d'une interaction G×E ($RR_I = 5$) est de 3,44. La valeur théorique sous l'hypothèse nulle ($RR_I = 1$) de l'ORR_0 est de 1,41. La double flèche bleue indique la comparaison qu'il faut réaliser pour que le test soit robuste à l'inflation de l'*ORR* due à la corrélation de E chez les germains.

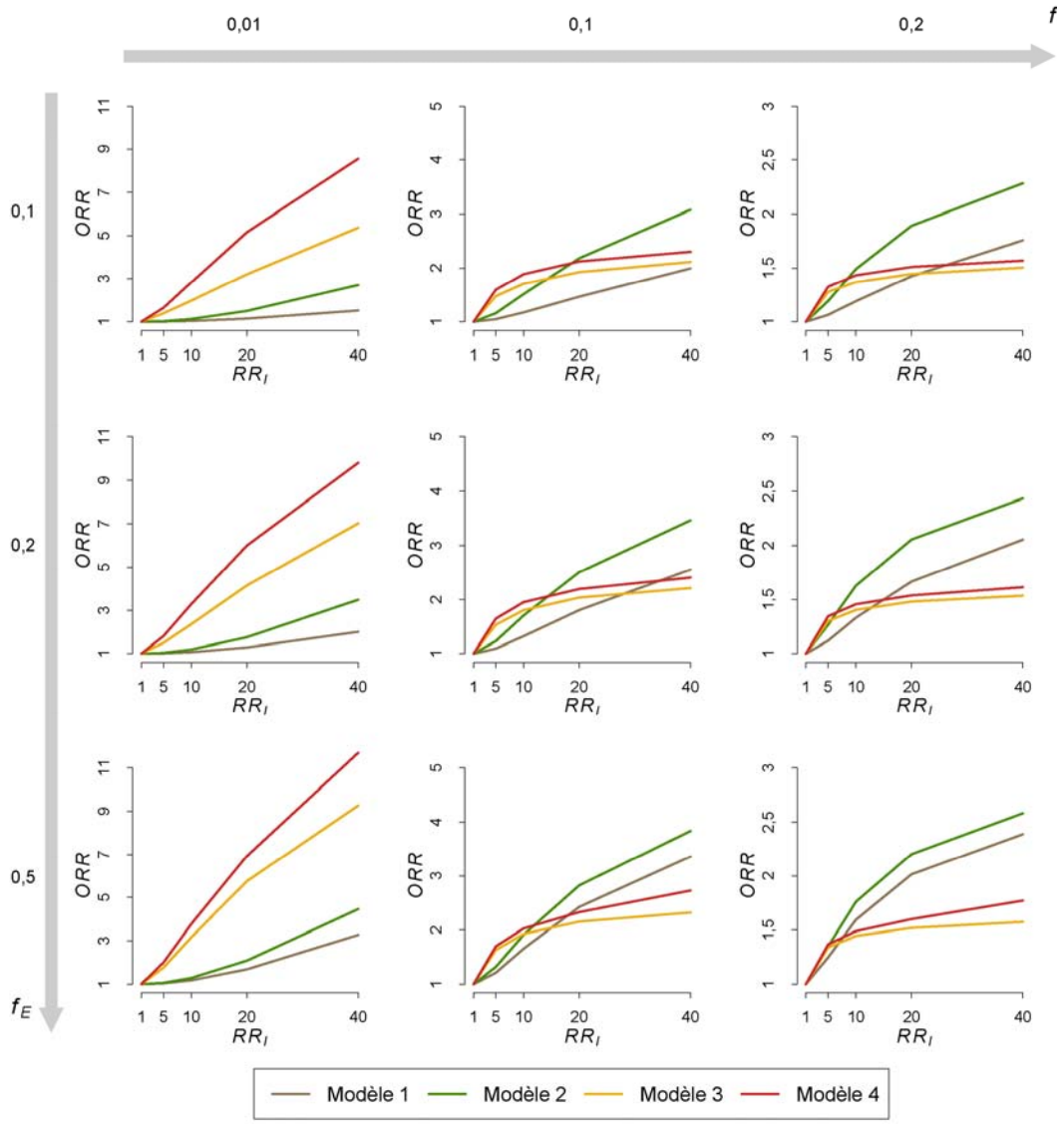


Figure 2.1 Odds ratio de récurrence (*ORR*) en fonction de l'interaction gène-environnement (RR_I), des fréquences des facteurs génétique (f_G) et environnemental (f_E) et pour différents modèles de risques relatifs génétique (RR_G) et environnemental (RR_E)

Modèle 1 – Interaction pure sans effets de *G* et *E* : $RR_G = RR_E = 1$

Modèle 2 – Interaction potentialisant un effet de *E* : $RR_G = 1$ et $RR_E = 5$

Modèle 3 – Interaction potentialisant un effet de *G* : $RR_G = 5$ et $RR_E = 1$

Modèle 4 – Interaction en présence d'un effet de *G* et de *E* : $RR_G = RR_E = 5$

corrélation environnementale chez les germains : $C_E = 0$; prévalence de la maladie $f_M = 0,1$

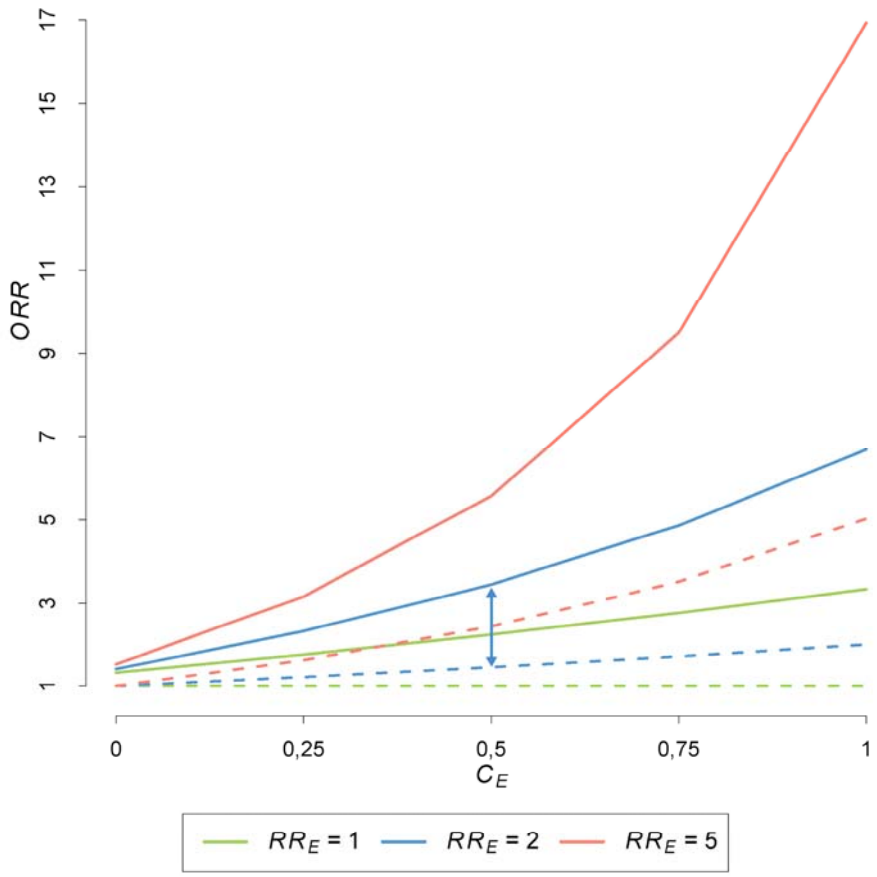


Figure 2.2 Odds ratio de récurrence (*ORR*) en fonction de la corrélation environnementale chez les germains (C_E) et du risque relatif environnemental (RR_E)

Les courbes pleines correspondent à des situations avec une interaction gène-environnement ($RR_I = 5$) et les courbes pointillées à des situations sous l'hypothèse nulle ($RR_I = 1$)

fréquence de l'exposition : $f_E = 0{,}2$; risque relatif génotypique : $RR_G = 2$ (modèle dominant) ; fréquence du facteur génétique : $f_G = 0{,}1$; prévalence de la maladie : $f_M = 0{,}1$

8. Construction du test EURECA

L'hypothèse nulle H0 du test EURECA est donc « $ORR = ORR_0$ ». Pour tester H0, nous proposons de modéliser la relation entre les deux variables maladie du germain M (S2) et exposition de l'index E (S1) par une régression logistique de la forme :

$$\text{logit}\,(P\,(M\,(\text{S2}) \mid E\,(\text{S1})) = \beta_0 + \beta_R\,E\,(\text{S1}) \qquad (2.16)$$

où β_0 est égal à ln ($K_{S\bar{E}}$ / (1 – $K_{S\bar{E}}$)) et β_R est égal à ln (ORR).

Le test de Wald du paramètre β_R de la régression logistique suit sous l'hypothèse nulle ($ORR = ORR_0$) un χ^2 à 1 ddl :

$$T_{Wald} = \sum \frac{\left(\hat{\beta}_R - \beta_0\right)^2}{\hat{\sigma}^2(\beta_R)} \qquad (2.17)$$

où $\beta_0 = \ln(ORR_0)$ et $\sigma^2(\beta_R)$ est la variance de ce paramètre dont l'estimation est donnée par :

$$\hat{\sigma}^2(\beta_R) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \qquad (2.18)$$

où a, b, c et d sont les effectifs observés du tableau de contingence (tableau 2.1).

L'erreur de première espèce et la puissance du test EURECA ont été estimées asymptotiquement pour une taille d'échantillon de 1000 paires de germains en utilisant des distributions de χ^2 non centrés. Nous avons également calculé le nombre de paires de germains nécessaire pour atteindre une puissance de 80 % avec une erreur de première espèce nominale de 5 % sous ces modèles.

9. Propriétés du test statistique EURECA

La puissance du test EURECA en fonction de l'interaction G×E (RR_I), pour différentes valeurs de corrélation (C_E) et pour un modèle génétique dominant est représentée sur la figure 2.3. Le test construit en utilisant la valeur estimée de l'ORR_0 a le niveau d'erreur de première espèce attendu. La puissance augmente avec des valeurs croissante de RR_I mais le plus intéressant est que cette augmentation est plus importante pour des valeurs élevées que pour des valeurs faibles de C_E. Le tableau 2.4 donne les effectifs de paires de germains nécessaires pour obtenir une puissance de 80 % avec une erreur de première espèce de 5 % pour des valeurs de RR_I et de C_E croissantes sous des modèles de maladie plausibles (facteurs de risque fréquents, f_G = 0,1 et f_E = 0,2 avec des effets modérés, RR_G = 2 et RR_E = 2). Pour des situations sans corrélation environnementale (C_E = 0) et des RR_I faibles, les effectifs nécessaires sont trop élevés et en pratique impossible à collecter. Mais pour des situations avec des interactions fortes (RR_I > 3) et en présence d'une corrélation environnementale des germains (C_E > 0), les effectifs sont plus réalistes. En fixant les mêmes fréquences des facteurs de risque et une corrélation à 0,25, la figure 5 (annexe 1) représente les effectifs nécessaires pour différentes valeurs de RR_G et RR_E. Ces effectifs sont plus petits pour des facteurs de risques ayant des effets indépendants forts, particulièrement pour le facteur génétique.

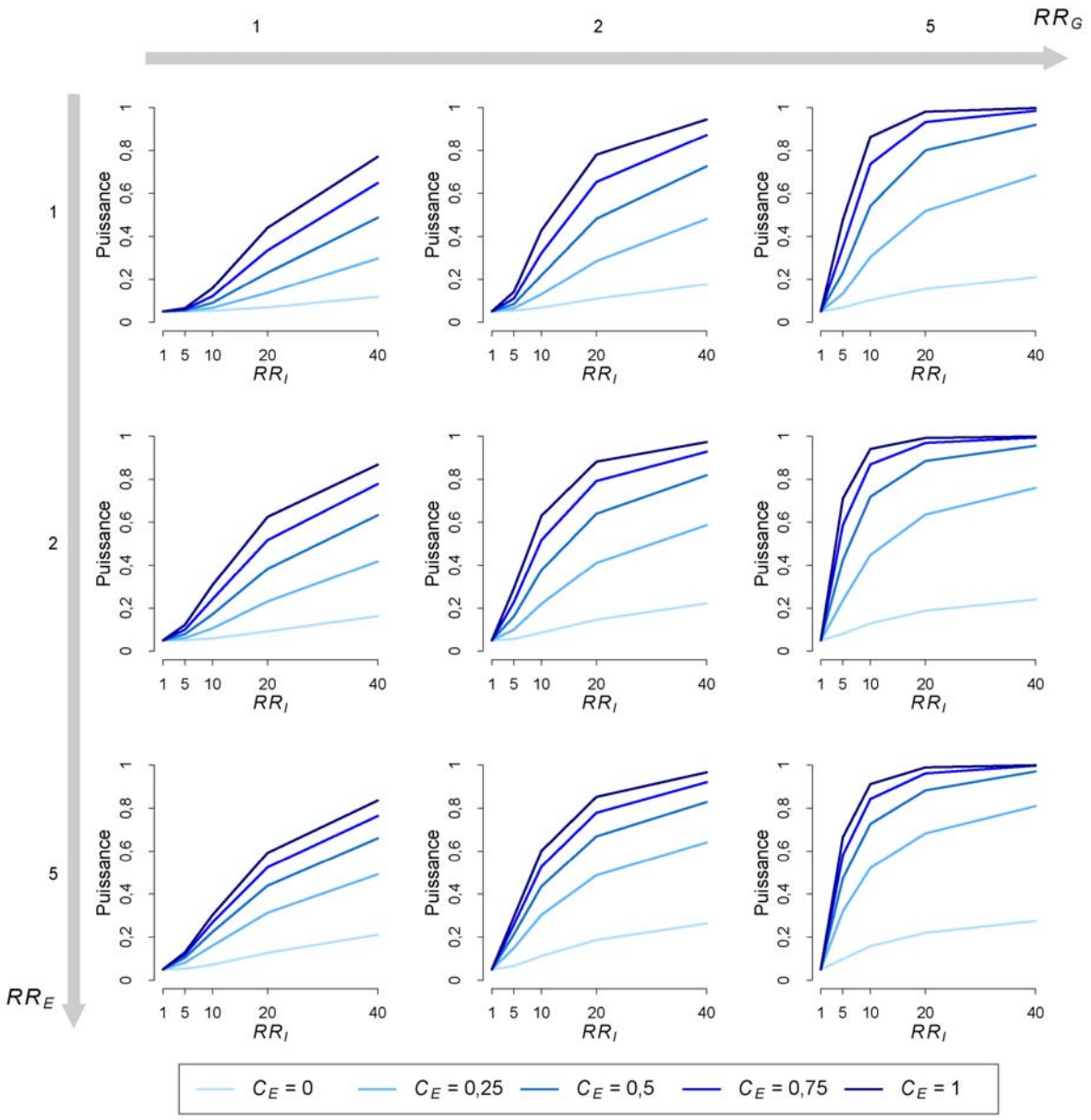


Figure 2.3 Puissance du test EURECA en fonction de l'interaction gène-environnement (RR_I) pour différentes valeurs de corrélation environnementale chez les germains (C_E) et de risques relatifs génétique (RR_G) et environnemental (RR_E)

fréquence du facteur génétique : f_G = 0,1 (modèle dominant) ; fréquence de l'exposition : f_E = 0,2 ; prévalence de la maladie : f_M = 0,1 ; effectif de 1000 paires de germains

		RR_I				
		1	2	3	5	10
C_E	0	∞	41516	8061	1754	448
	0,25	∞	8418	1745	423	126
	0,5	∞	3875	841	219	70
	0,75	∞	2345	529	145	47
	1	∞	1625	379	108	35

Tableau 2.4 Nombre de paires de germains nécessaire pour obtenir une puissance de 80 % avec une erreur de première espèce nominale de 5 % en fonction de l'interaction gène-environnement (RR_I) et de la corrélation environnementale chez les germains (C_E)

fréquence du facteur génétique : f_G = 0,1 (modèle dominant) ; risque relatif génétique : RR_G = 2 ; fréquence de l'exposition : f_E = 0,2 ; risque relatif environnemental : RR_E = 2 ; prévalence de la maladie : f_M = 0,1

10. Calcul de l'ORR_0

L'utilisation du test EURECA nécessite d'avoir une estimation de l'ORR_0. Pour les calculs de puissance, cette estimation a été aisément obtenue à partir des modèles avec une corrélation de *E* chez les germains et sans interaction G×E, tout autre paramètre fixé. Par contre, la question du calcul de l'ORR_0 en présence de données réelles se pose.

Nous avons dérivé la formule de l'ORR_0 à partir de celle de l'*ORR* en considérant un modèle sans interaction G×E (RR_I = 1). Cette formule est très complexe puisqu'elle dépend des six paramètres du modèle : les fréquences des facteurs *G* (f_G) et *E* (f_E), les risques relatifs des facteurs *G* (RR_G) et *E* (RR_E), la corrélation de *E* chez les germains (C_E) et la prévalence de la maladie dans la population (f_M). Cependant, la majeure partie de la variation des valeurs de l'ORR_0 dépend des paramètres du facteur *E* (f_E, RR_E et C_E), à un moindre degré de la prévalence de la maladie (f_M) et quasiment pas des paramètres du facteur génétique (f_G, RR_G).

En supposant l'absence d'effet du facteur génétique (RR_G = 1), la formule de l'ORR_0 se simplifie de la façon suivante :

$$ORR_0 = \left[\frac{C_E(1-RR_E)+(1-RR_E)f_E(1-C_E)-1}{(1-RR_E)f_E(1-C_E)-1} \times \frac{f_E(1-RR_E)\left[f_M(1-C_E)-1\right]-f_M+1}{f_M C_E(1-RR_E)+f_E(1-RR_E)\left[f_M(1-C_E)-1\right]-f_M+1} \right] \quad (2.19)$$

Pour estimer l'ORR_0 en pratique, il est donc nécessaire d'avoir des estimations fiables de ces différents paramètres. La prévalence de la maladie peut être aisément obtenue à partir d'études épidémiologiques descriptives dans la population étudiée. Les paramètres environnementaux (C_E, f_E et RR_E) peuvent être estimés en utilisant l'échantillon d'étude lorsque l'information sur E a été collectée chez tous les sujets (comme dans l'application sur le diabète de type 2 présentée au chapitre 2). Si ce n'est pas le cas, les résultats d'études précédentes sur ce facteur peuvent également être exploités. Seul le modèle génétique demeure inconnu car non spécifié dans la méthode EURECA. Nous proposons de calculer l'ORR_0 pour différentes valeurs des paramètres génétiques (f_G variant de 0,01 à 0,5 et RR_G variant de 0,5 à 10 par exemple) puis d'utiliser la valeur obtenue la plus proche de l'ORR observé sur l'échantillon de germains. Cette valeur la plus proche assure une inférence robuste du test statistique. Afin de calculer l'ORR_0 attendu à partir des paramètres du modèle, un programme écrit sous Maple version 10 (Waterloo Maple Inc. 2005) est disponible sur simple demande.

Chapitre 2
Application au diabète de type 2

Ce second chapitre présente une illustration sur des données familiales concernant le diabète de type 2 où deux facteurs « environnementaux » ont été recueillis dans ces familles (le statut pondéral et l'activité physique). L'objectif ici est de montrer que l'outil que nous avons développé est applicable en pratique sur des données réelles pour lesquelles une estimation de l'ORR_0 devra être réalisée sans information sur le modèle génétique. À la suite de ce chapitre une discussion plus générale mettra en avant les avantages mais également les limites de cette nouvelle approche.

1. Problématique

Le diabète de type 2 (ou diabète non insulinodépendant) est une cause majeure de morbidité et de mortalité dans les pays développés. De 1980 à 2004, la prévalence de diabète de type 2 diagnostiqué a quasiment doublé (*Centers for Disease Control and Prevention* 2004). Les facteurs génétiques jouent un rôle important dans le déterminisme du diabète de type 2 et il est communément admis que des facteurs environnementaux interagissent également avec ces facteurs génétiques pour augmenter le risque de diabète de type 2 (Bartsocas et Leslie 2002; Elbein 1997). L'activité physique est un régulateur important du métabolisme du glucose et des études suggèrent qu'une activité physique importante peut avoir un rôle dans la prévention du diabète de type 2. Par ailleurs, les études cliniques ont également mis en évidence qu'une activité physique régulière, associée à un régime alimentaire et une perte de poids, peut prévenir le diabète dans différentes populations et groupes d'âge (Knowler *et al.* 2002; Pan *et al.* 1997; Tuomilehto *et al.* 2001). Le phénotype « obèse » fait parti du syndrome métabolique caractérisé par une intolérance au glucose et son association au diabète de type 2 est largement documentée (Kriska *et al.* 2003). L'objectif de cette application est d'utiliser le test EURECA comme indicateur pour sélectionner les facteurs environnementaux qu'il faut prendre en compte dans les futures études génétiques.

2. Matériel

L'étude Gene ENvironment Interactions (GENI) (Nelson *et al.* 2007) a recueilli des données phénotypiques et environnementales chez tous les individus de 452 familles recrutées à partir d'un individu atteint de diabète de type 2 (le proposant de la famille) dans la vallée de Saint-Louis et l'agglomération de Denver dans le Colorado (USA). À partir de ces 452 familles, 3090 familles nucléaires ont été identifiées incluant 2699 paires de germains S1-S2, S1 étant atteint et pour lequel au moins l'une des deux variables environnementales étudiées était mesurée : le statut pondéral ou l'activité physique. Parmi ces paires, 1734 étaient d'origine hispanique (H) et 965 étaient caucasiens d'origine non-hispanique (NH). Les sujets dont le diagnostic de diabète de type 2 avait été précédemment posé par un médecin et qui étaient traités par des antidiabétiques oraux ou de l'insuline ont été classés atteints. Le statut diabétique des autres sujets a été déterminé par un test de tolérance au glucose selon les critères de l'Association Américaine du Diabète (*American Diabetes Association*, 1997). Le statut pondéral a été déterminé en fonction de l'Indice de Masse Corporelle (IMC) déterminé au moment du diagnostic, pour les sujets diabétiques, ou calculé au moment du recrutement, pour les autres sujets. Les individus dont l'IMC était supérieur ou égal à 30 kg/m^2 ont été classés obèses, tandis que les autres ont été classés non obèses. L'évaluation de l'activité physique a été effectuée à une seule reprise au cours de l'étude en utilisant un auto-questionnaire précédemment validé (Kriska *et al.* 1990). La dépense énergétique a été évaluée en unités d'équivalence métabolique (MET : *Metabolic Equivalent Task*). Le MET est le rapport de la dépense énergétique métabolique pendant l'exercice sur la dépense énergétique métabolique au repos (Lynch *et al.* 1996). Le MET moyen par semaine (avant le diagnostic de diabète de type 2 pour les sujets précédemment diagnostiqués) a été calculé pour chaque sujet à partir de l'auto-questionnaire. La variable MET a été divisée en tertiles spécifiques à chaque sexe et une variable dichotomique a été créée distinguant les individus du tertile inférieur (« activité physique faible ») de ceux des deux tertiles supérieurs.

3. Méthode

Nous avons effectué l'analyse séparément pour les deux strates de population (H et NH) car les distributions des facteurs d'exposition étaient significativement différentes. Nous avons tout d'abord évalué l'association de chacun des facteurs *E* par une régression logistique conditionnelle en utilisant les paires de germains discordantes pour le statut diabétique (index

+ germain non malade). Les effectifs de paires de germains discordantes étaient de 198 H et 116 NH pour le statut pondéral et 458 H et 309 NH pour l'activité physique. La fréquence d'exposition a été mesurée dans l'échantillon de témoins (germains non atteints) et utilisée pour estimer la prévalence de chaque facteur *E*.

Nous avons ensuite sélectionné un germain aléatoirement pour chaque index dans le but de calculer les effectifs des cases du tableau de contingence et les risques de récurrence global et stratifiés (K_S, K_{SE} et $K_{S\bar{E}}$). Les effectifs de paires de germains utilisés étaient de 267 H et 321 NH pour le statut pondéral, 246 H et 268 NH pour l'activité physique. Puis, nous avons calculé l'*ORR* et appliqué le test EURECA d'interaction en utilisant un modèle de régression logistique non conditionnelle. Afin de prendre en compte la corrélation des paires appartenant à une même famille, nous avons calculé la variance du paramètre de régression en utilisant un estimateur sandwich tel qu'implémenté dans Stata SE version 10.1 (StataCorp. 1984-2008). Lorsque l'exposition du germain était disponible, la paire a également été utilisée pour calculer la corrélation environnementale chez les germains pour chacun des facteurs d'exposition en utilisant l'équation 2.6.

4. Résultats

L'ensemble des résultats de l'application au diabète de type 2 sont présentés dans le tableau 2.5. Pour chaque strate de population (H et NH) et chaque facteur d'exposition étudié, sont présentées d'abord les estimations des paramètres environnementaux : OR_E, f_E et C_E, et ensuite l'analyse d'interaction G×E incluant l'*ORR* et le test *EURECA*. Afin de prendre en compte la corrélation environnementale chez les germains, nous avons calculé l'ORR_0 (*ORR* attendu sous l'hypothèse nulle). La prévalence du diabète de type 2 dans l'état du Colorado en 2001 (année de l'étude) était de 4,5 % selon les données des *Centers for Disease Control and Prevention* (2004). À partir de cette estimation et des estimations des paramètres environnementaux, nous avons calculé les valeurs d'ORR_0 attendues pour différentes valeurs de paramètres génétiques (f_G = 0,01 – 0,5 ; RR_G =0,5 – 10). Afin d'assurer une robustesse maximale du test, nous avons comparé la valeur observée de l'*ORR* à la borne de l'intervalle de variation de l'ORR_0 la plus proche.

	Obésité		**Faible activité physique**	
	H	**NH**	**H**	**NH**
OR_E	2,48	3,87	1,13	0,93
IC à 95 %	1,18 – 5,22	1,54 – 9,65	0,72 – 1,77	0,53 – 1,65
f_E	0,29	0,37	0,31	0,23
C_E	0,22	0,14	– 0,02	0,07
ORR_0	1,25* – 1,27	1,22* – 1,25	0,99 – 1,00*	0,99 – 1,00*
ORR	0,67	1.03	1.14	2,13
IC à 95 %	0,40 – 1,11	0,53 – 1,99	0,62 – 2,08	1,08 – 4,19
EURECA	4,03	0,25	0,15	4,78
p	0,045	0,617	0,70	0,028

Tableau 2.5 Résultats de l'application sur le diabète de type 2

H: hispaniques ; NH: non-hispaniques ; OR_E: estimation de l'odds ratio environnemental ; f_E: estimation de la fréquence du facteur environnemental ; C_E: estimation de la corrélation environnementale chez les germains ; ORR_0: intervalle de variation de la valeur attendue de l'odds ratio de récurrence sous l'hypothèse nulle ; ***** valeur de la borne la plus proche de l'*ORR* observé utilisée pour réaliser le test ; *ORR*: odds ratio de récurrence ; IC : intervalle de confiance ; EURECA : *Exposed versus Unexposed Recurrence Analysis*

Discussion

Contraster les risques de récurrence en fonction du statut d'exposition de l'index est une méthode simple et attrayante pour sélectionner les facteurs environnementaux impliqués dans des interactions G×E. Nous proposons avec la méthode EURECA de mesurer cette différence en calculant l'odds ratio de récurrence (*ORR*) et nous montrons que l'*ORR* est un bon indicateur de la présence d'une interaction G×E. Cet *ORR* n'est pas une mesure directe de l'interaction G×E, mais plutôt une mesure de la différence entre les risques de récurrence des index exposés et non exposés. En utilisant l'exemple de la faible activité physique chez les caucasiens non hispaniques (NH) (tableau 2.5), le risque de diabète de type 2 chez un individu NH est multiplié en moyenne par 2,13 quand son germain atteint présente une activité physique faible comparé à un individu dont le germain atteint n'a pas une activité physique faible.

En substituant l'information familiale à l'information génétique, il est difficile de distinguer la composante génétique sous-jacente qui interagit avec le facteur *E* étudié d'une corrélation familiale de ce facteur *E* qui serait lui-même associé à la maladie, mais nos résultats montrent que cela est possible à condition que l'effet de ce facteur et sa corrélation chez les paires de germains soient bien documentés.

Dans le contexte d'une variable *E* dichotomique, l'intérêt d'utiliser l'*ORR* au lieu du rapport des risques de récurrence réside dans l'application du modèle logistique qui est très couramment utilisé dans les études épidémiologiques traditionnelles, mais la même approche peut être étendue à des variables *E* multinomiales ou continues en utilisant le modèle linéaire généralisé. L'utilisation de variables continues lorsqu'elles sont disponibles peut potentiellement améliorer la puissance mais fait l'hypothèse d'une relation linéaire continue. Afin de tester la différence de l'*ORR* avec sa valeur sous l'hypothèse nulle, nous avons construit un test statistique, EURECA, qui nécessite cependant de pouvoir calculer la valeur de l'ORR_0.

Nous avons dérivé analytiquement une formule pour calculer l'ORR_0 en fonction des paramètres du facteur *E* et de la prévalence de la maladie. Des estimations de ces paramètres peuvent être obtenues à partir de l'échantillon étudié si cette variable a été documentée chez

les deux germains, ou bien à partir de résultats issus de la littérature. Afin de prendre en compte l'impact des paramètres génétiques sur la valeur de l'ORR_0, nous proposons de calculer son intervalle de variation et de choisir la valeur se rapprochant le plus de la valeur observée de l'*ORR*. Notons que la perte de puissance due à l'incertitude sur les paramètres génétiques est minimale puisque l'intervalle de variation de l'ORR_0 est généralement petit. Il était par exemple de l'ordre de 0,01 à 0,03 dans l'application sur le diabète de type 2. Il est intéressant de voir sur notre exemple que nous avons pu mettre en évidence une différence significative entre la valeur observée de l'*ORR* et l'ORR_0 en utilisant cette procédure. Ceci est en accord avec les résultats de Khoury *et al.* (1988b) qui montraient que le degré d'agrégation familiale des maladies les plus fréquentes ne peut être entièrement expliquée par une agrégation familiale des facteurs *E* même si l'on considère des valeurs extrêmes de corrélation familiale pour ce facteur.

L'étude des propriétés statistiques d'EURECA a montré que ce test est plus adapté à l'étude des maladies fréquentes que des maladies rares (figure 6, annexe 1). Il est aussi intéressant de noter que même après correction pour la corrélation du facteur *E* chez les germains, les puissances étaient plus élevées pour des situations présentant des valeurs élevées de C_E plutôt que des valeurs faibles. Une explication possible est que la corrélation a en réalité à la fois un effet confondant qu'il est important de contrôler pour éviter les faux positifs, mais d'autre part qu'elle accentue la différence existante entre les risques de récurrence des index exposés et non exposés due à l'interaction G×E. Nous n'avons pas évalué des situations de corrélation négative qui nous ont semblé peu probables même si en pratique on peut supposer que ce serait une source de perte de puissance.

Les interactions G×E sont difficiles à détecter et nécessitent souvent des tailles d'échantillon très grandes. Afin d'améliorer la puissance de détection des interactions G×E, de nouvelles méthodes fondées sur des échantillonnages particuliers ont été développées. Parmi ces méthodes, nous pouvons citer la modélisation log-linéaire qui utilise des données trios et qui compare les distributions de génotypes d'individus exposés et non exposés en conditionnant sur les génotypes parentaux (Umbach et Weinberg 2000), des méthodes qui utilisent des échantillonnage avec contre-appariement qui permettent d'enrichir l'échantillon avec des facteur d'exposition ou des facteurs génétiques rares (Andrieu *et al.* 2001). Il existe également des schémas cas-témoins-combinés avec des témoins à la fois populationnels et familiaux (Goldstein *et al.* 2006). La caractéristique commune à toutes ces méthodes de

détection des interactions G×E est qu'elles nécessitent d'avoir une information exhaustive des statuts d'exposition et des génotypes des sujets étudiés. Parmi les méthodes utilisant l'agrégation familiale en substitut du facteur génétique latent, Purcell (Purcell 2002) a proposé d'utiliser les modèles de décomposition de la variance dans les études de jumeaux pour mettre en évidence des interactions G×E avec un facteur *E* mesuré chez les deux jumeaux. Ce qui distingue l'approche que nous proposons est justement qu'elle ne nécessite que l'information d'exposition de l'index et la récurrence chez le germain pour appliquer le test. Cette méthode s'adapte donc mieux aux types de recrutements disponibles en épidémiologie génétique. Il n'est pas nécessaire d'avoir une mesure de l'exposition chez le germain et pour des maladies dont le phénotype est simple à déterminer, l'information de récurrence peut être obtenue directement du sujet index. De grands échantillons peuvent donc être recrutés pour un coût minimal. Il est cependant vrai que si le germain peut également être examiné, comme ce fut le cas dans notre application sur le diabète de type 2, la récurrence familiale sera d'autant mieux estimée. Cela permet aussi d'estimer la corrélation du facteur *E* chez les germains directement à partir de l'échantillon plutôt que de données issues de la littérature.

L'utilisation de l'information de récurrence en substitut de l'information génétique a aussi l'avantage de ne pas faire d'hypothèse à priori sur le modèle génétique sous-jacent et sur les gènes impliqués. Contrairement aux autres méthodes qui recherchent une interaction G×E pour de nombreux marqueurs génétiques, cette méthode recherche une interaction du facteur *E* avec l'ensemble de la composante génétique impliquée dans la maladie en un seul test, permettant de s'affranchir du problème des tests multiples. Ceci est un point important car cette problématique dans les études des interactions G×E considérant des milliers de marqueurs génétiques couplés à des dizaines de facteurs d'exposition est un problème encore irrésolu. D'un autre côté, cette approche ne spécifie que le facteur *E* et teste son interaction avec la composante génétique impliquée dans l'augmentation de risque de la maladie sans que cette composante soit clairement identifiée. Comparativement à des méthodes utilisant les deux sources d'information *G* et *E*, notre méthode peut manquer de puissance pour détecter certaines interactions avec des facteurs *G* spécifiques. Mais elle apporte un moyen simple de sélectionner les facteurs *E* qui sont potentiellement impliqués dans des interactions G×E lorsque les génotypes ne sont pas disponibles.

L'association entre le diabète de type 2 et l'obésité était significative dans les deux strates d'hispaniques (H) et de caucasiens non hispaniques (NH), comme cela avait été précédemment mis en évidence dans nombreuses études transversales et longitudinales (Kriska *et al.* 2003). Concernant l'interaction avec l'obésité, le test EURECA était significatif seulement pour la strate des H (p = 0,045) avec un modèle particulier d'interaction qui va en sens inverse de l'effet marginal de l'obésité sur le diabète de type 2. Dans une précédente étude des risques de récurrence dans des familles de sujets atteints diabète de type 2, un résultat analogue avait été trouvé avec des ratios de risque de récurrence plus élevés chez les germains d'index non-obèses comparativement aux index obèses (Weijnen *et al.* 2002). Ce type d'interaction illustre typiquement la situation où l'interaction G×E est un élément de nuisance qu'il faut correctement prendre en compte afin de mieux détecter un effet génétique (Kraft *et al.* 2007; Selinger-Leneman *et al.* 2003).

Concernant l'activité physique faible qui n'avait aucun effet indépendant observé, le test EURECA était significatif chez les NH (p = 0,028) mais pas chez les H. Une précédente étude utilisant des tests d'association fondés sur des trios et des modèles d'équations d'estimation généralisées (*Generalized Estimating Equations*) montrait une interaction G×E entre des polymorphismes du gène du récepteur γ activateur de la prolifération des peroxysomes (PPAR-γ) et l'activité physique faible chez les H également (Nelson *et al.* 2007).

Le recrutement d'index par des familles multiplexes comme dans le cas de l'étude GENI ne permet pas de généraliser ces résultats à la population générale des patients diabétiques. En effet, on s'attend à avoir un enrichissement en allèles de susceptibilité à la maladie dans ces familles et ainsi les estimations de risques de récurrence sont susceptibles d'être augmentées par rapport aux estimations attendues dans la population générale (Weijnen *et al.* 2002). Cependant, ceci ne devrait pas induire une hétérogénéité entre les strates d'index exposés et non exposés à moins que la corrélation chez les germains pour ce facteur E n'ait pas été correctement prise en compte. Dans cet exemple, les résultats orientent vers la sélection de sujets non obèses dans les études génétiques menées chez les H, tandis qu'il serait plus intéressant de rechercher des interactions avec l'activité physique faible chez les NH. Ceci illustre comment il est possible d'utiliser la valeur de l'*ORR*, son intervalle de confiance et la valeur p pour ordonner plusieurs facteurs E et sélectionner en priorité ceux

pour lesquels une interaction G×E devra être testée dans une étude génétique ultérieure ou pour orienter le recrutement de certaines catégories d'individus pour réaliser des études génétiques.

En conclusion, ces résultats montrent qu'une quantité importante d'information familiale peut être exploitée afin de mieux décrire les interactions G×E impliquées dans les maladies complexes. Cet outil permettra aux investigateurs d'identifier les facteurs environnementaux qui sont les plus à même d'interagir avec la composante génétique avant de réaliser des études sur des marqueurs génétiques. Ces facteurs environnementaux doivent donc être pris en compte dans l'analyse génétique ou peuvent aussi être utilisés pour concevoir l'étude en sélectionnant des sous-catégories de la population afin d'améliorer la puissance pour identifier les facteurs génétiques.

Partie 3

Interaction gène-environnement et panel de témoins de référence

Chapitre 1
Étude d'association génétique avec un panel de témoins de référence

La diminution très rapide du coût du génotypage à grande échelle a permis l'émergence des études d'association pangénomiques avec des effectifs suffisamment grands pour détecter des effets génétiques relativement faibles. En apportant de nouvelles pistes fonctionnelles parfois complètement insoupçonnées, cette stratégie fait maintenant parti des méthodes de premier choix pour l'analyse des maladies complexes sur des données de population. La problématique que nous abordons maintenant est liée à l'utilisation de panels de témoins de référence dans ce type d'étude. Nous décrivons dans ce chapitre le principe d'utilisation de ces panels et en présentons les principaux inconvénients

1. Qu'est-ce qu'un panel de témoin de référence ?

Au cours des dernières années, des échantillons d'individus provenant de diverses populations dans le monde ont été collectés et génotypés sur des puces de SNPs à haute densité. Les génotypes de ces individus ont été enregistrés dans des bases de données de génotypes. Ces bases de données ont permis de décrire les variations de fréquence de ces polymorphismes dans les différentes populations humaines, comme en témoigne les avancées importantes apportées par la base de données HapMap (www.hapmap.org) (The International HapMap Consortium 2003, 2005). Elles ont également permis des avancées en génétique des populations et en génétique évolutive, mais ont surtout ouvert de nouvelles perspectives pour l'étude génétique des maladies complexes (Manolio *et al.* 2008).

Ces données génotypiques obtenues sur des centaines voire des milliers d'individus sont habituellement mises à la disposition de la communauté scientifique et peuvent être utilisées comme témoins dans les études d'association génétique (figure 3.1). L'utilisation de tels panels comme témoins de référence est une solution très avantageuse afin de réduire le coût lié au génotypage à haute densité de grands échantillons. D'ores et déjà, des études collaboratives ont fait appel à ce mode d'échantillonnage où les mêmes individus servent de témoins pour différents échantillons de malades atteints de différentes maladies. Les études du *Wellcome Trust Case Control Consortium* (WTCCC) représentent un exemple typique de

cette stratégie d'échantillonnage où 7 échantillons d'environ 2000 cas d'individus atteints de 7 maladies complexes différentes ont été comparés à un échantillon unique de 3000 témoins (The Wellcome Trust Case Control Consortium 2007).

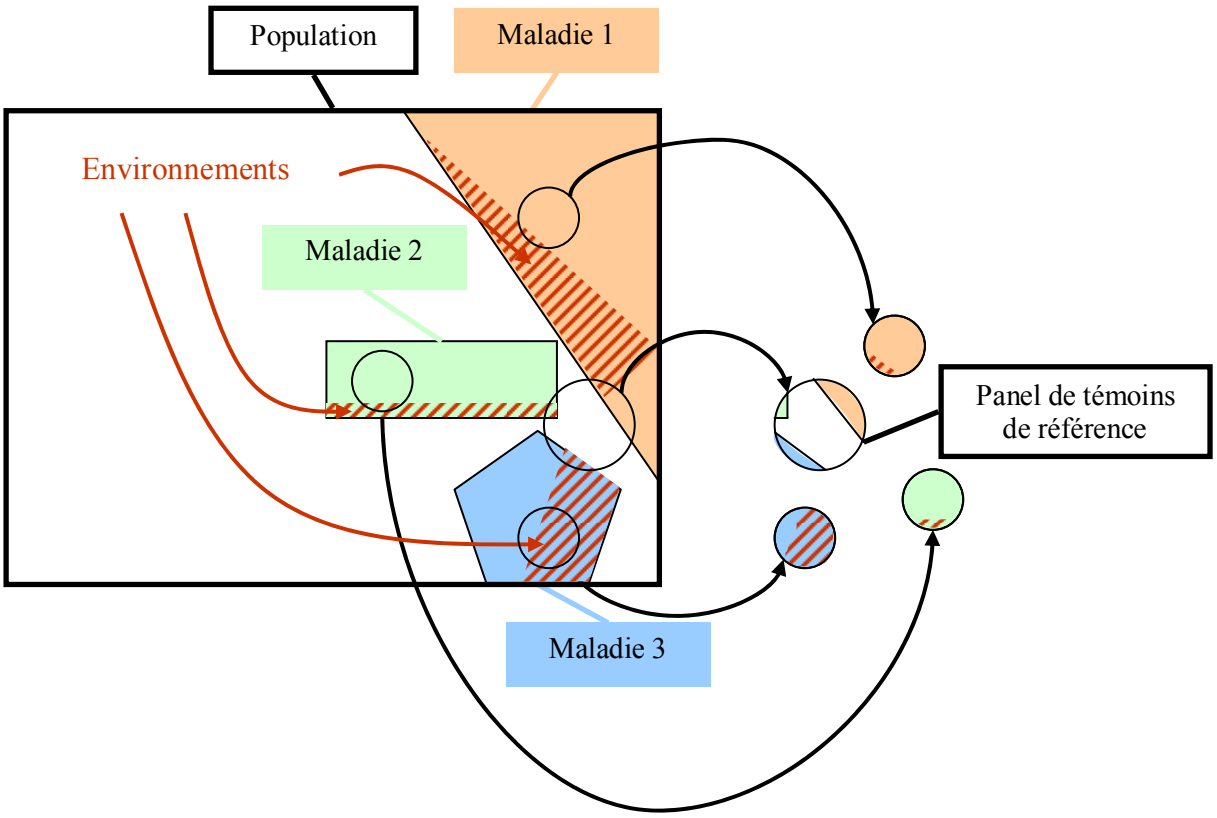


Figure 3.1 Représentation schématique de l'échantillonnage d'un panel de témoins de référence dans l'étude génétique de 3 maladies dans une population donnée

Le tableau 3.1 répertorie quelques exemples de ce type de panels (Cann *et al.* 2002; Holle *et al.* 2005; Krawczak *et al.* 2006; Manolio *et al.* 2007; Nelson *et al.* 2008; The International HapMap Consortium 2003, 2005; The Wellcome Trust Case Control Consortium 2007; Wichmann *et al.* 2005). Par ailleurs dans de nombreux pays, de gigantesques biobanques sont mises en place ayant pour projet la collecte d'échantillons biologiques et d'informations phénotypiques et environnementales sur des centaines de milliers d'individus voire des populations entières (deCODE Genetics Biobank Iceland; Estonian Genome Project; Generation Scotland; Genetic Alliance Biobank (USA); Sweedish National Biobank Program; UK Biobank).

Échelle géographique	Panel	Nombre d'individus	Populations représentées	Puce de génotypage
Mondial	HapMap Phase 1 [a]	270	4 populations	1,3 millions de SNPs
	HapMap Phase 2 [b]	270	4 populations	3,1 millions de SNPs
	HapMap Phase 3	1301	11 populations	1,5 millions de SNPs
	HGDP–CEPH [c]	1 050	51 populations	Illumina 650K et 550K
	POPRES (GSK) [d]	5 886	5 régions du monde	Affymetrix 500K
Régional	POPRES (GSK) [d]	1 387	22 pays européens	Affymetrix 500K
	EuroPa (CNG) [e]	5 811	13 pays européens	Illumina 317K ou 370K
Pays	WTCCC [f]	3000	Royaume-Uni	Affymetrix 500K
	PopGen [g]	10 000	Allemagne (Nord)	Affymetrix 500K
	KORA S4 [h]	2 200	Allemagne (Sud)	Affymetrix 500K
	KORA-gen [i]	18 000	Allemagne (Sud)	(projet de biobanque)

Tableau 3.1 Exemples de panels de témoins de référence dans le monde

[a] (The International HapMap Consortium 2003) ; [b] (The International HapMap Consortium 2007) ; [c] (Cann *et al.* 2002) ; [d] (Nelson *et al.* 2008) ; [e] (Heath *et al.* 2008) ; [f] (The Wellcome Trust Case Control Consortium 2007) ; [g] (Krawczak *et al.* 2006) ; [h] (Holle *et al.* 2005) ; [i] (Wichmann *et al.* 2005)

2. Problèmes liés à l'utilisation d'un panel

L'utilisation d'un panel de témoins de référence en substitut de témoins recrutés de façon spécifique pour l'étude de plusieurs maladies peut présenter de nombreuses difficultés. Tout d'abord, un biais de classement du phénotype des témoins est inhérent au recrutement d'individus aléatoires dans la population. Ensuite, une stratification de population risque d'augmenter le taux de faux positifs notamment lorsque le panel est composé d'individus provenant d'une population différente des cas recrutés. Un autre problème est la corrélation des résultats de l'étude de plusieurs phénotypes avec les mêmes témoins à chaque marqueur génétique. Mais le principal obstacle pour lequel nous avons cherché à apporter une solution est l'absence de mesure du facteur environnemental chez les témoins de référence.

2.1. Biais de classement différentiel

Classiquement lors d'une étude épidémiologique cas-témoins, la sélection des témoins se fait en fonction de critères d'exclusion particuliers qui vont permettre d'éviter le recrutement de sujets susceptibles d'être atteints de la maladie. Une première conséquence directe de l'utilisation d'un panel de sujets recrutés aléatoirement dans la population générale est la possibilité pour certains de ces témoins d'être atteints de la maladie (figure 3.1). Le recrutement n'étant soumis à aucune sélection particulière pour éviter l'enrôlement d'individus malades, on s'attend à ce que ce biais de classement soit généralement de l'ordre de grandeur de la prévalence de la maladie. Ce biais entraîne une perte de puissance des tests de détection du facteur génétique associé à la maladie ou des tests d'interaction G×E. En effet, la présence d'individus atteints de la maladie parmi les témoins va diminuer le contraste observé entre les fréquences génotypiques (ou alléliques) des groupes de malades et de témoins. Cependant, cette perte de puissance est généralement contrebalancée par la possibilité d'avoir des tailles d'échantillons très grandes à moindre coût. En effet, un biais de classement de l'ordre de 5 % chez les témoins entraînerait une perte de puissance équivalente à une diminution de la taille d'échantillon de 10 % (The Wellcome Trust Case Control Consortium 2007).

2.2. Biais de confusion par stratification de population

Une seconde difficulté liée à l'utilisation d'un panel de témoin de référence est l'éventuelle présence d'une stratification de population. Ce biais est potentiellement présent dans toute étude d'association menée sur des échantillons de cas et de témoins non apparentés. En effet, si la population d'étude est divisée en sous-groupes ayant des prévalences de la maladie différentes, la contribution respective des ces sous-groupes aux échantillons de cas et de témoins n'est pas la même. Si entre ces différents sous-groupes il existe de plus des variations des fréquences génotypiques (ou alléliques) à certains marqueurs, il y a alors un risque de conclure à tort à une association entre la maladie et ces marqueurs. Ce problème est accru lorsque l'on utilise des panels de témoins de référence pour lesquels l'information sur l'origine des individus peut être approximative, voire non disponible. Différentes approches pour corriger ce type de biais ont été proposées dans la littérature. Celle qui est actuellement la plus utilisée dans les études d'association pangénomiques se base sur

l'analyse en composantes principales des données génotypiques qui permet d'identifier les axes de variations majeurs (les composantes principales) présents dans les données. Ces axes identifiés, il est possible de réaliser des tests d'association qui les prennent en compte en réalisant une sorte de pseudo-appariement des malades et des témoins sur ces axes (Patterson *et al.* 2006; Price *et al.* 2006).

2.3. Corrélation des résultats d'association pour plusieurs phénotypes

L'utilisation de témoins communs pour étudier l'association d'un marqueur donné avec plusieurs phénotypes entraîne une corrélation entre les tests de comparaison des différentes distributions de génotypes des échantillons de malades à la distribution de génotypes de l'échantillon de témoins. En effet, Zaykin et Kozbur (2009) ont montré que la corrélation entre deux tests d'association ayant un recouvrement partiel des témoins dépend des effectifs des échantillons de cas et de témoins et non pas des fréquences génotypiques au marqueur. La corrélation ρ_{12} entre les tests d'association d'un marqueur génétique donné avec une maladie 1 et une maladie 2 s'exprime par :

$$\rho_{12} = \left[\left(1 + \frac{N_{02}}{N_0} \right) \left(1 + \frac{N_{01}}{N_0} \right) \left(1 + \frac{N_{01}}{N_1} + \frac{N_0}{N_1} \right) \left(1 + \frac{N_{02}}{N_2} + \frac{N_0}{N_2} \right) \right]^{-1} \qquad (3.1)$$

où N_1 et N_2 sont respectivement les effectifs des échantillons de sujets atteints de la maladie 1 et 2, N_0 est l'effectif de l'échantillon de témoins partagé, N_{01} et N_{02} sont respectivement les effectifs de témoins non partagés par les deux tests d'association avec la maladie 1 et la maladie 2.

Par exemple si l'on étudie deux phénotypes avec deux échantillons indépendants de 2000 malades chacun en utilisant un échantillon commun de 3000 témoins, la corrélation pour un marqueur génétique entre les deux tests d'association effectués est de 0,16. Zaykin et Kozbur (2009) proposent d'exploiter cette relation entre la corrélation et les ratios des effectifs de témoins et de cas pour corriger les tests statistiques en conditionnant le calcul de la valeur *p* du test d'association de la maladie 2 sur la valeur observée *p* du test d'association de la maladie 1.

2.4. Information environnementale non mesurée

Enfin, l'inconvénient majeur de ce type d'échantillonnage quand on s'intéresse aux interactions G×E est l'absence d'information environnementale sur ces témoins de référence. En effet, l'information sur les expositions environnementales potentiellement à risque, si elle est disponible, ne le sera que chez les malades puisque cet échantillon aura été collecté spécifiquement pour l'étude de la maladie qui nous intéresse, contrairement à l'échantillon de témoins non spécifiques pour lequel on envisage difficilement de disposer d'une information environnementale cohérente (figure 3.1). Une solution pour résoudre ce problème et tester l'interaction G×E consiste à n'utiliser que les malades dans un schéma d'étude « cas-seuls ». Cependant l'approche cas-seuls ne permet que de mettre en évidence une interaction G×E et ne permet pas d'estimer, ni de tester l'association du facteur génétique. Les méthodes de modélisation les plus utilisées en génétique épidémiologique pour estimer et tester l'association avec un facteur génétique en tenant compte d'un facteur environnemental sont fondées sur la forme binomiale du modèle logistique, où la variable dépendante (ici la maladie) est binaire et où les covariables indépendantes sont observées chez tous les individus. L'absence d'information sur le facteur environnemental chez les témoins rend donc ces méthodes inutilisables en pratique. Le développement d'une approche alternative dans cette situation ouvrirait donc de nouvelles perspectives pour les études d'interactions G×E à l'échelle de tout le génome.

Chapitre 2
MInT, une méthode permettant de s'affranchir de l'environnement des témoins

Nous proposons une nouvelle méthode, la méthode MInT (*Multinomial Interaction Test*), qui s'affranchit du problème de l'absence d'information environnementale chez les témoins. La méthode est fondée sur la régression logistique multinomiale. Nous présentons dans un premier temps le principe de son utilisation pour tester conjointement l'association génétique et l'interaction G×E dans la situation où l'information sur la variable environnementale n'est pas disponible chez les témoins, puis dans un second temps, nous évaluons les propriétés statistiques du test réalisé à partir de ce modèle et nous le comparons aux tests d'association et d'interaction G×E qui sont actuellement utilisés dans les études d'association. Ces travaux ont fait l'objet d'un article en cours de révision (*American Journal of Epidemiology*) qui se trouve en annexe 2.

Comme nous l'avons défini dans la partie 1, nous notons M, G et E les variables dichotomiques représentant respectivement le statut maladie, le facteur génétique et le facteur environnemental des individus (les valeurs 0 et 1 indiquant respectivement l'absence ou la présence de la maladie ou du facteur chez un individu).

1. Modèle logistique multinomiale

Le modèle de régression logistique multinomial (ou polytomique) est un modèle logistique adapté à des variables dépendantes présentant plus de deux catégories (Dobson 2002; Kleinbaum et Klein 2002). Son utilisation pour contraster les effets de facteurs de risque dans différents sous-phénotypes d'une même pathologie est largement répandue en épidémiologie traditionnelle. Par ailleurs, la puissance de ce modèle pour étudier des effets génétiques hétérogènes dans des sous-phénotypes de maladie a récemment été mise en évidence (Morris *et al.* 2009). Le principe de ce modèle logistique repose sur la comparaison simultanée des mesures d'association entre un facteur de risque et différentes sous-catégories d'une variable dépendante multinomiale, en prenant l'une de ces catégories comme référence. La régression logistique multinomiale s'adapte donc parfaitement à notre contexte particulier où l'information environnementale n'est pas recueillie dans le groupe de témoins ($M = 0$). Les

variables dichotomiques M et E peuvent être combinées en une seule variable multinomiale M_P à trois classes prenant les valeurs 0 chez les témoins, 1 chez les cas non exposés et 2 chez les cas exposés. Le codage des trois catégories peut être permuté sans conséquence sur le résultat de l'analyse car le modèle multinomial est non ordinal, mais pour faciliter l'interprétation des estimations de paramètres, il est préférable d'utiliser le groupe de témoins ($M_P = 0$) comme groupe de référence. Le modèle multinomial s'écrit :

$$\text{logit } P\,(M_P = j \mid G) = \log\,[P\,(M_P = j \mid G) / P\,(M_P = 0 \mid G)] = \beta_{0j} + \beta_{Gj}\,G \qquad (3.2)$$

où j prend les valeurs 1 et 2.

Les deux équations logistiques correspondantes sont utilisées simultanément pour estimer les quatre coefficients du modèle β_{01}, β_{02}, β_{G1} et β_{G2} qui maximisent la vraisemblance de l'échantillon sous ce modèle.

Les paramètres β_{01} et β_{02} sont des coefficients difficiles à interpréter cependant l'exponentielle de leur différence est un estimateur du ratio des probabilités d'être exposé et non exposé parmi les cas non porteurs du génotype :

$$e^{\beta_{02}-\beta_{01}} = \frac{P(M_P = 2 \mid G = 0)P(M_P = 0 \mid G = 0)}{P(M_P = 1 \mid G = 0)P(M_P = 0 \mid G = 0)} = \frac{P(E = 1 \mid M = 1, G = 0)}{P(E = 0 \mid M = 1, G = 0)} \qquad (3.3)$$

Les paramètres β_{G1} et β_{G2} représentent respectivement les logarithmes des odds ratios génétiques des cas non exposés et des cas exposés par rapport au groupe de témoins :

$$OR_{G1} = e^{\beta_{G1}} = \frac{P(G = 1 \mid M_P = 1)/P(G = 0 \mid M_P = 1)}{P(G = 1 \mid M_P = 0)/P(G = 0 \mid M_P = 0)} \qquad (3.4)$$

$$OR_{G2} = e^{\beta_{G2}} = \frac{P(G = 1 \mid M_P = 2)/P(G = 0 \mid M_P = 2)}{P(G = 1 \mid M_P = 0)/P(G = 0 \mid M_P = 0)} \qquad (3.5)$$

A priori, l'expression de OR_{G1} laisse présager qu'il s'agit d'un estimateur de l'association du facteur génétique avec la maladie. Tandis que le rapport de ces deux odds ratio (l'exponentielle de la différence des deux paramètres β_{G2} et β_{G1}) correspond à l'estimation de l'odds ratio d'interaction G×E :

$$OR_{IM} = \frac{OR_{G2}}{OR_{G1}} = \frac{P(G = 1 \mid M_P = 2)/P(G = 0 \mid M_P = 2)}{P(G = 1 \mid M_P = 1)/P(G = 0 \mid M_P = 1)} \qquad (3.6)$$

On constate ici que cette estimation ne dépend plus des témoins et en remplaçant la variable M_P par la combinaison des variables correspondantes M et E, on retrouve l'expression de l'estimateur de l'interaction G×E du modèle logistique avec un échantillonnage cas-seuls.

A partir de ce modèle, un test conjoint de l'association génétique et de l'interaction G×E peut être effectué en évaluant l'hypothèse nulle $\beta_{G1} = \beta_{G2} = 0$ par un test de rapport de vraisemblances à 2 ddl comparant le modèle multinomial saturé de l'équation 3.2 au modèle restreint contraignant les deux paramètres β_{Gj} à être égaux à 0 :

$$\text{MInT} = -2 \ln [L(\beta_{01} ; \beta_{02} ; \beta_{G1} = \beta_{G2} = 0) / L(\beta_{01} ; \beta_{02} ; \beta_{G1} ; \beta_{G2})] \qquad (3.7)$$

2. Comparaison de MInT aux autres approches logistiques

L'objectif est d'évaluer les propriétés statistiques (erreur de première espèce et puissance) de MInT et de le comparer aux différents tests d'association génétique et/ou d'interaction G×E disponibles pour différents modèles d'interaction G×E. Le biais, la variance et la probabilité de couverture des coefficients de régression du modèle multinomial ont également été calculés et comparés aux paramètres des autres approches.

2.1. Tests d'association génétique et/ou d'interaction gène-environnement

Nous avons choisi de comparer MInT à ceux couramment utilisés dans la littérature pour tester l'association génétique et/ou l'interaction G×E. Ils sont tous fondés sur une régression logistique binomiale. Ces approches diffèrent principalement par l'information environnementale qu'elles nécessitent d'avoir sur l'échantillon et par l'hypothèse qui est testée. Le tableau 3.2 présente un résumé de l'ensemble de ces méthodes qui ont été décrites au chapitre 3 de la partie 1. Brièvement, les différents tests sont les suivants :

Test conjoint de l'association génétique et de l'interaction G×E (GI–binomial)

Lorsque l'information sur l'exposition est disponible chez les cas et chez les témoins, le modèle logistique complet s'écrit :

$$\text{logit } P(M = 1 | G, E) = \beta_{0CT} + \beta_{ECT} E + \beta_{GCT} G + \beta_{ICT} GE \qquad (3.8)$$

Le test conjoint de l'association génétique et de l'interaction G×E se fait par un test du rapport de vraisemblances à 2 ddl dont l'hypothèse nulle est $\beta_{GCT} = \beta_{ICT} = 0$ (Kraft *et al.* 2007).

***Test d'interaction G×E dans un schéma cas-témoins* (I–cas-témoins)**

A partir du modèle logistique complet (équation 3.8), on teste l'interaction G×E seule par un test du rapport de vraisemblances à 1 ddl dont l'hypothèse nulle est $\beta_{ICT} = 0$.

***Test d'association génétique ajusté sur l'exposition* (G–ajusté)**

Lorsque l'exposition est disponible chez les cas et chez les témoins, le modèle avec ajustement sur la variable E s'écrit :

$$\text{logit } P(M = 1 \mid G, E) = \beta_{0A} + \beta_{EA} E + \beta_{GA} G \qquad (3.9)$$

Un test du rapport de vraisemblances à 1ddl permet de tester l'hypothèse nulle d'absence d'association génétique, $\beta_{GA} = 0$.

***Test d'interaction G×E dans un schéma cas-seuls* (I–cas-seuls)**

Le modèle logistique du schéma cas-seuls s'écrit :

$$\text{logit } P(E = 1 \mid G) = \beta_{0CS} + \beta_{ICS} G \qquad (3.10)$$

Si la maladie est rare et que les facteurs G et E sont indépendants en population, le coefficient de régression β_{ICS} estime l'interaction G×E. Un test du rapport de vraisemblances à 1 ddl dont l'hypothèse nulle est $\beta_{ICS} = 0$ permet de la tester.

***Test d'association génétique marginale* (G–marginal)**

Lorsque l'information sur l'exposition n'est disponible ni chez les cas, ni chez les témoins, il n'est possible que de modéliser l'effet marginal de G dans un modèle de régression logistique qui s'écrit :

$$\text{logit } P(D = 1 \mid G) = \beta_{0M} + \beta_{GM} G \qquad (3.11)$$

Le test d'association entre G et M ($\beta_{GM} = 0$) est testé par un rapport de vraisemblances à 1 ddl.

Effet testé	Disponibilité chez les cas		Disponibilité chez les témoins		Modèle logistique	Hypothèse nulle	Degrés de liberté	Notations dans les figures et les tableaux
	Génotype	Exposition	Génotype	Exposition				
Génétique seul	Oui	Non	Oui	Non	Binomial	$\beta_{GM} = 0$	1	G–marginal
	Oui	Oui	Oui	Oui	Binomial	$\beta_{GA} = 0$	1	G–ajusté
Interaction G×E seule	Oui	Oui	Oui	Oui	Binomial	$\beta_{ICT} = 0$	1	I–cas-témoins
	Oui	Oui	Non	Non	Binomial	$\beta_{ICS} = 0$	1	I–cas-seuls
Génétique et Interaction G×E	Oui	Oui	Oui	Oui	Binomial	$\beta_{GCT} = 0$ et $\beta_{ICT} = 0$	2	GI–binomial
	Oui	Non	Oui	Oui	Multinomial	$\beta_{G1} = 0$ et $\beta_{G2} = 0$	2	GI–MInT

Tableau 3.2 Méthodes disponibles pour différents schémas de disponibilité d'information environnementale et génotypique

β_{GM} est le coefficient de régression génétique dans un modèle où seul le facteur génétique est modélisé : logit $P(M = 1| G) = \beta_{0M} + \beta_{GM} G$

β_{GA} est le coefficient de régression génétique dans un modèle avec ajustement sur la variable environnementale : logit $P(M = 1| G, E) = \beta_{0A} + \beta_{EA} E + \beta_{GA} G$

β_{GCT} et β_{ICT} sont les coefficients de régression génétique et d'interaction dans le modèle saturé : logit $P(M = 1| G, E) = \beta_{0CT} + \beta_{ECT} E + \beta_{GCT} G + \beta_{ICT} GE$

β_{ICS} est le coefficient de régression de l'interaction G×E dans un plan d'échantillonnage cas-seuls : logit $P(E = 1| G) = \beta_{0CS} + \beta_{ICS} G$

β_{G1} et β_{G2} sont les coefficients de régression du modèle multinomial : logit $P(M_P = j| G) = \beta_{0j} + \beta_{Gj} G$

2.2. Modélisation, simulation et critère d'évaluation

Nous avons modélisé la probabilité de maladie conditionnellement aux facteurs génétique et environnemental en utilisant une modélisation logistique telle que présentée dans la partie 1 (page 45) :

$$\text{logit } P(M = 1 | G, E) = \beta_0 + \beta_G G + \beta_E E + \beta_I GE \qquad (3.12)$$

où $\beta_0 = \ln(f_B / 1 - f_B)$, $\beta_G = \ln(OR_G)$, $\beta_E = \ln(OR_E)$ et $\beta_I = \ln(OR_I)$

Les mêmes notations ont été conservées.

Par ailleurs, nous avons également modélisé une corrélation gène-environnement par un coefficient θ_{GE} qui peut varier entre 0 et $+\infty$ tel que précédemment présenté (page 43) :

$$\theta_{GE} = \frac{P(E=1|G=1)/P(E=0|G=1)}{P(E=1|G=0)/P(E=0|G=0)} \qquad (3.13)$$

En appliquant le théorème de Bayes et en utilisant les équations 3.12 et 3.13, les probabilités attendues des différentes catégories C_{ijk} de $G = i$ et $E = j$ sachant $M = k$ sont :

$$P(G=i, E=j|M=k) = \frac{P(M=k|G=i, E=j)P(G=i)P(E=j|G=i)}{P(M=k)} \qquad (3.14)$$

où $P(M = k \mid G = i, E = j)$ est obtenue à partir de l'équation 3.12 et $P(E = j \mid G = i)$ est une fonction de θ_{GE}, f_E et f_G.

Afin d'imiter l'utilisation d'un panel de témoins de référence issu de la population générale, nous avons introduit un biais de classification chez les témoins en supposant que 10 % des témoins sont atteints de la maladie sans qu'on ne le sache. Nous avons utilisé des valeurs de f_G et f_E variant de 0,1 à 0,9 par intervalles de 0,2 ; des valeurs de OR_G, OR_E et OR_I variant de 0,5 à 2 (à 3 pour OR_I) par intervalles de 0,25 et des valeurs de θ_{GE} variant de 1 à 2 par intervalles de 0,25 pour une prévalence de la maladie fixée à f_M = 0,1. Ensuite, pour chaque combinaison de paramètres, nous avons calculé les effectifs attendus dans les différentes catégories C_{ijk} en utilisant l'équation 3.14 et nous avons simulé 1000 réplicats par modèle en fixant la taille d'échantillon à 500 cas et 500 témoins.

L'erreur de première espèce et la puissance des différents tests du tableau 3.2 ont été estimées asymptotiquement en utilisant des distributions de χ^2 non-centrés avec les degrés de liberté correspondant et une erreur de première espèce nominale de 0,01. Elles peuvent également être estimées par la proportion de réplicats simulés pour lesquels le test est significatif à un degré de 0,01 ou moins ($p \leq 0{,}01$). Puisque ces deux méthodes de calcul de l'erreur de première espèce et de la puissance ont donné des résultats équivalents, seuls les résultats des calculs asymptotiques sont présentés ici.

Le carré du biais, la variance et la probabilité de couverture (probabilité que l'intervalle de confiance à 95 % de l'estimation contienne la valeur théorique du paramètre) des différents paramètres de l'effet génétique et de l'interaction G×E estimés par les différentes modèles ont été calculés à partir des données simulées.

Tous les calculs et les simulations ont été réalisés avec un script écrit en R (R Development Core Team 2009) et les analyses statistiques ont été réalisées à l'aide des fonction *logit* et *mlogit* de Stata SE version 10.1 (StataCorp. 1984-2008).

3. Résultats

3.1. Erreur de première espèce

Les erreurs de première espèce obtenues sous l'hypothèse nulle d'absence d'association génétique ($OR_G = 1$) et d'absence d'interaction G×E ($OR_I = 1$) sont présentées sur la figure 3.2 en fonction des autres paramètres (θ_{GE}, f_G, OR_E et f_E). En l'absence de corrélation gène-environnement ($\theta_{GE} = 1$), les erreurs de première espèce sont très proches de la valeur attendue (0,01). Par contre, en présence d'une corrélation, les erreurs de première espèce des tests I–cas-seuls et de GI–MInT sont augmentées et cela d'autant plus que la corrélation est forte et que les fréquences des deux facteurs sont proches de 0,5. Les erreurs de première espèce du test G–marginal sont également augmentées, mais de façon beaucoup plus modérée que pour les deux tests précédents. Les trois autres méthodes qui utilisent l'information environnementale des témoins sont quant à elles robustes à la présence d'une corrélation gène-environnement.

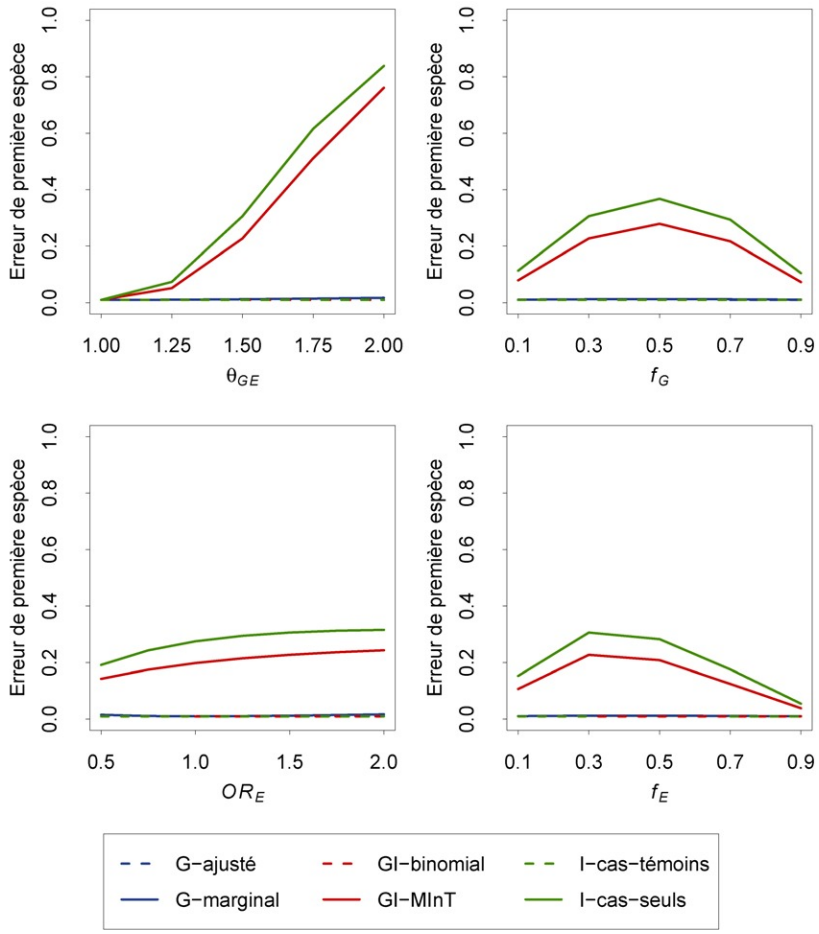


Figure 3.2 Erreur de première espèce en fonction de la corrélation gène-environnement (θ_{GE}), de la fréquence du facteur génétique (f_G), de l'odds ratio environnemental (OR_E) et de la fréquence d'exposition (f_E)

Le modèle de base à partir duquel ces paramètres varient est le suivant : odds ratio d'interaction G×E, OR_I = 1 ; odds ratio génétique, OR_G = 1 ; fréquence du facteur génétique, f_G = 0,3 ; odds ratio environnemental, OR_E = 1,5 ; fréquence de l'exposition, f_E = 0,3 ; corrélation gène-environnement, θ_{GE} = 1,5 ; prévalence de la maladie, f_M = 0,1 ; erreur de première espèce nominale de 0,01 pour un test bilatéral et un échantillon de 500 cas et 500 témoins

3.2. Puissance en l'absence de corrélation gène-environnement

Les figures 3.3 et 3.4 représentent les puissances asymptotiques des différents tests en fonction de l'interaction G×E pour différentes situations en l'absence de corrélation gène-environnement ($\theta_{GE} = 1$).

Le test GI–MInT a une puissance supérieure à tous les autres tests en présence d'une interaction G×E (figure 3.3). Les mêmes tendances sont observées pour des fréquences variables des deux facteurs avec une puissance globalement meilleure de l'ensemble des tests pour des fréquences proches de 0,5 (figure 3.4).

En l'absence d'interaction G×E, GI–MInT a une puissance légèrement inférieure aux tests de l'effet génétique seul (G–marginal et G–ajusté, courbes bleues), ce qui s'explique par le fait que la stratification de l'échantillon de cas en fonction de l'exposition apporte peu d'information et augmente la variance des paramètres et le degré de liberté du test MInT. Cette différence de puissance est au maximum de 11,25 % quand $OR_G = 1{,}5$ et $f_G = 0{,}5$. En contrepartie, le gain de puissance par rapport aux tests de l'association génétique seule est bien plus important, tout particulièrement pour des situations d'interaction G×E pure où les facteurs *E* et *G* n'ont pas d'effets indépendants. Le gain de puissance dans cette situation peut atteindre 42 % et 54 % par rapport aux tests G–marginal et G–ajusté, respectivement. Ajuster sur l'exposition diminue la puissance du test d'association dans la majorité des situations. Le test G–ajusté est plus puissant que le test G–marginal lorsque *E* a un effet qui va dans le sens opposé à celui de *G* (première ligne et première colonne de la figure 3.3).

Par rapport aux deux tests d'interaction G×E (courbes vertes), GI–MInT a une puissance plus importante lorsque le facteur génétique a un effet propre ($OR_G \neq 1$). Lorsque $OR_G = 1$, la puissance du test GI–MInT est identique à celle du test I–cas-seuls. Le test I–cas-témoins qui utilise l'information environnementale des cas et des témoins a la plus faible puissance quelque soit le modèle.

Finalement le fait le plus intéressant mis en évidence dans la figure 3.3 est que, pour tous les modèles étudiés, le test GI–MInT a une puissance supérieure au test GI–binomial qui test aussi conjointement les effets du facteur *G* et de l'interaction G×E (Kraft *et al.* 2007) et qui nécessite de disposer de l'information environnementale des témoins.

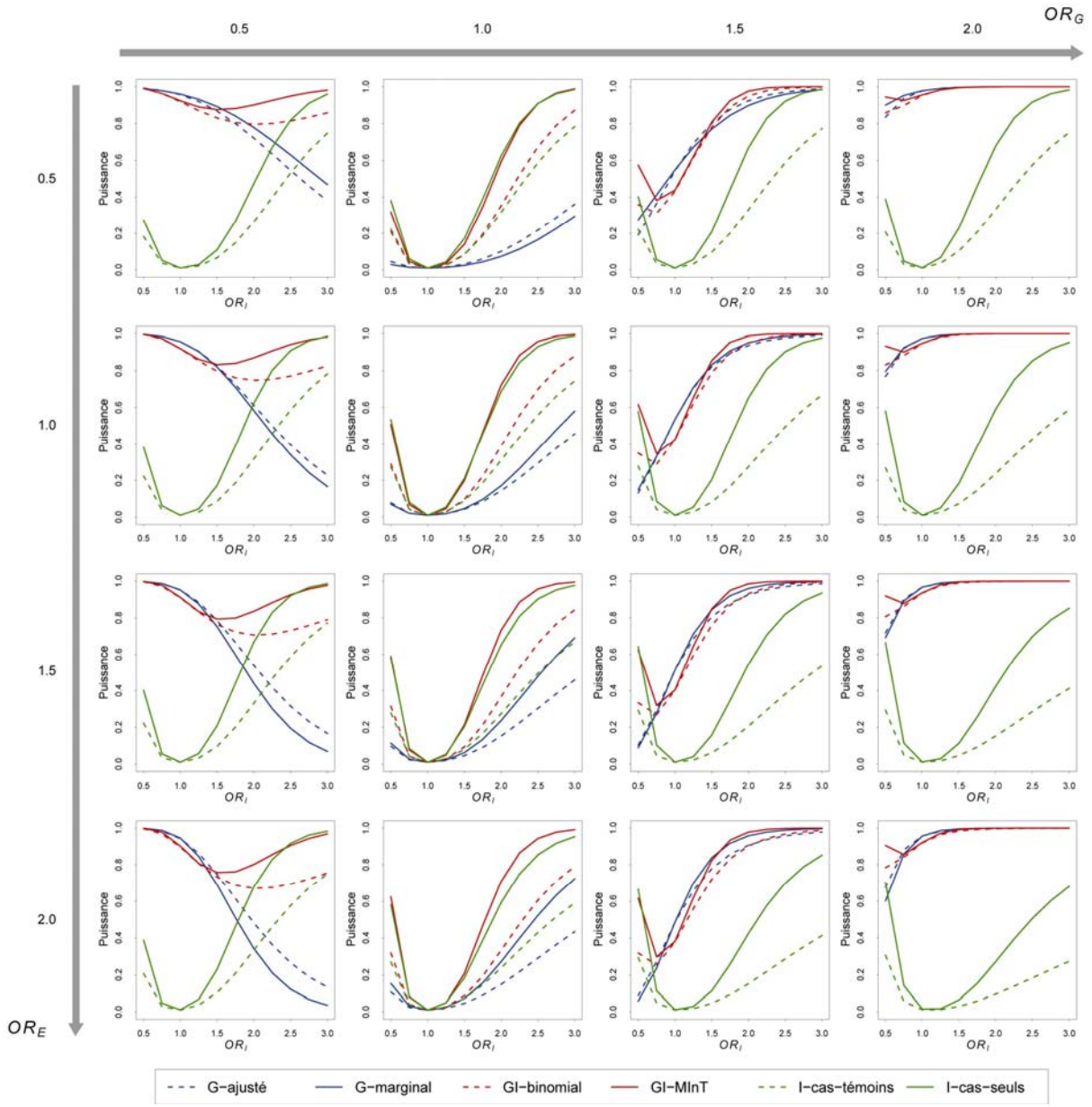


Figure 3.3 Puissance asymptotique en fonction de l'interaction gène-environnement (OR_I) pour différentes valeurs d'odds ratios des facteurs génétique (OR_G) et environnemental (OR_E)

fréquence du facteur génétique, f_G = 0,3 (modèle dominant) ; fréquence de l'exposition, f_E = 0,3 ; corrélation gène-environnement, θ_{GE} = 1 ; prévalence de la maladie, f_M = 0,1 ; erreur de première espèce nominale de 0,01 pour un test bilatéral et un échantillon de 500 cas et 500 témoins

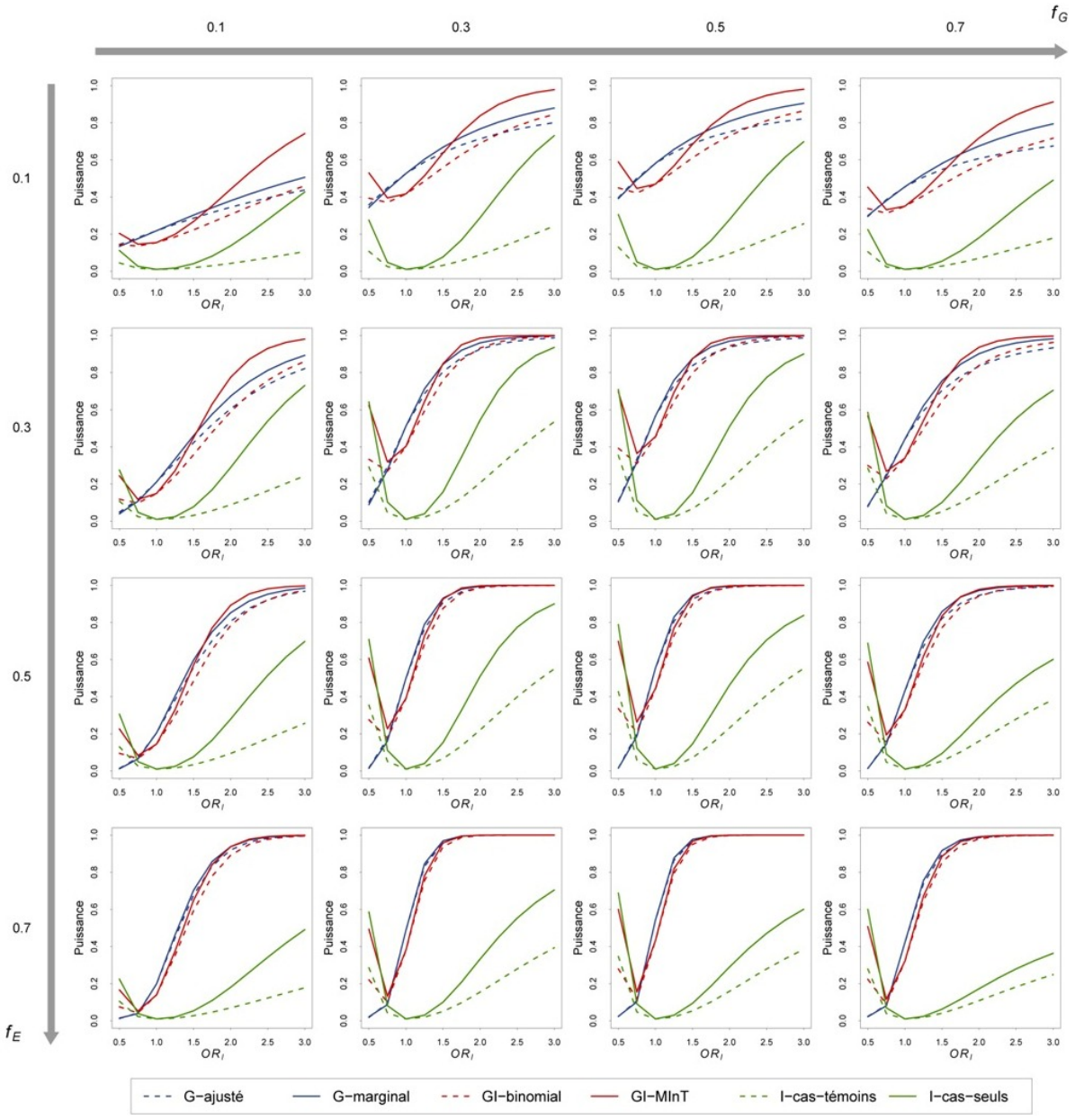


Figure 3.4 Puissance asymptotique en fonction de l'interaction gène-environnement (OR_I) pour différentes valeurs de fréquences des facteurs génétique (f_G) et environnemental (f_E)

odds ratio génétique, OR_G = 1,5 (modèle dominant) ; odds ratio environnemental, OR_E = 1,5 ; corrélation gène-environnement, θ_{GE} = 1 ; prévalence de la maladie, f_M = 0,1 ; erreur de première espèce nominale de 0,01 pour un test bilatéral et un échantillon de 500 cas et 500 témoins

3.3. Puissance en présence de corrélation gène-environnement

Les calculs de puissance pour les modèles incluant une corrélation gène-environnement ($\theta_{GE} > 1$) ont été ajustés pour prendre en compte l'inflation de l'erreur de première espèce en utilisant un seuil corrigé plutôt que le seuil standard du χ^2 centré (qui est de 6,63 et 9,21 pour 1 et 2 ddl respectivement et pour une erreur de première espèce nominale de 0,01). Ces seuils corrigés ont été calculés en utilisant des distributions de χ^2 non-centrés. Le paramètre de non-centralité est égal à la valeur du test pour le modèle correspondant à une même valeur de θ_{GE}, mais sans effet du facteur génétique et sans interaction G×E (sous l'hypothèse nulle).

La figure 3.5 représente ces puissances asymptotiques en fonction de l'interaction G×E et pour différentes valeurs d'OR_G et de θ_{GE}. Lorsque les valeurs de corrélation augmentent, on observe une perte de puissance des deux tests qui avaient une inflation de leurs erreurs de première espèce (I–cas-seuls et GI–MInT) tout particulièrement pour des valeurs d'interactions G×E faibles et pour les interactions G×E antagonistes ($OR_I < 1$). En utilisant comme valeur de référence la puissance en l'absence de corrélation ($\theta_{GE} = 1$), le tableau 3.3 quantifie les variations de puissance en fonction de la corrélation gène-environnement. Pour des valeurs d'interaction G×E fortes, la puissance de tous les tests a plutôt tendance à augmenter lorsque θ_{GE} augmente sauf dans les situations d'interaction antagonistes ($OR_I < 1$) pour les tests I–cas-seuls et GI–MInT où la perte de puissance par rapport à la situation d'absence de corrélation peut être importante.

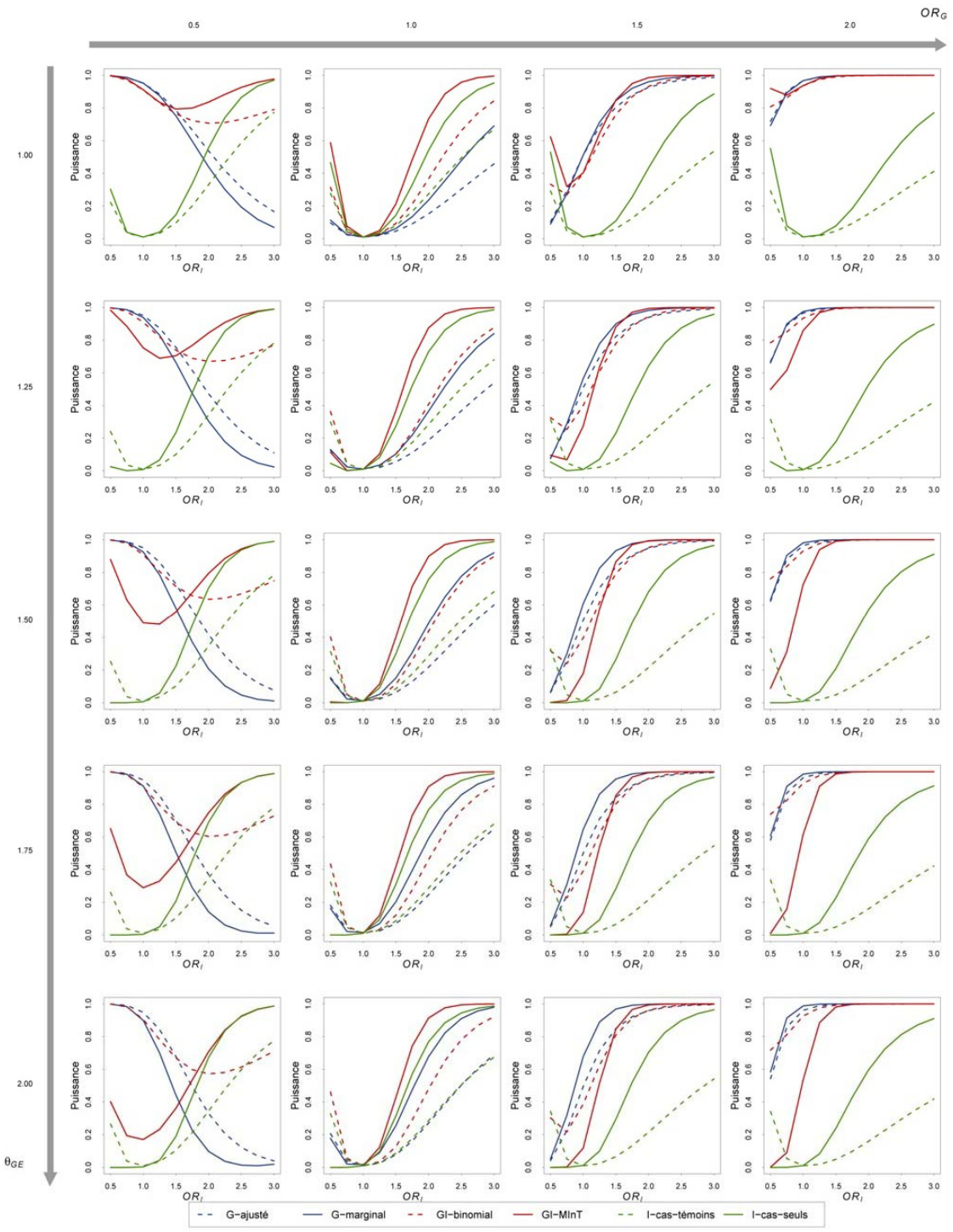


Figure 3.5 Puissance asymptotique en fonction de l'interaction gène-environnement (OR_I), de l'odds ratio génétique (OR_G) et de la corrélation gène-environnement (θ_{GE})

fréquence du facteur génétique, f_G = 0,3 (modèle dominant) ; odds ratio environnemental, OR_E = 1,5 ; fréquence de l'exposition, f_E = 0,3 ; prévalence de la maladie, f_M = 0,1; erreur de première espèce nominale de 0,01 avec un test bilatéral et pour un échantillon de 500 cas et 500 témoins

OR_G	θ_{GE}	G–ajusté		G–marginal		GI–binomial		GI–MInT		I–cas-témoins		I–cas-seuls	
		OR_I		OR_I		OR_I		OR_I		OR_I		OR_I	
		0,5	2,0	0,5	2,0	0,5	2,0	0,5	2,0	0,5	2,0	0,5	2,0
1,0	1,00	**9,68**	**15,20**	**11,35**	**23,62**	**31,43**	**36,67**	**58,86**	**73,30**	**27,77**	**27,39**	**57,84**	**65,25**
	1,25	+ 2,89	+ 3,51	+ 1,23	+ 12,05	+ 4,86	+ 4,05	– 47,43	+ 14,28	+ 2,18	+ 0,95	– 52,78	+ 11,24
	1,50	+ 5,71	+ 6,62	+ 1,68	+ 22,10	+ 8,80	+ 7,11	– 58,37	+ 16,58	+ 3,60	+ 1,34	– 57,81	+ 12,35
	1,75	+ 8,40	+ 9,39	+ 1,68	+ 30,05	+ 12,03	+ 9,49	– 58,84	+ 17,58	+ 4,51	+ 1,37	– 57,83	+ 12,64
	2,00	+ 10,96	+ 11,86	+ 1,46	+ 36,30	+ 14,71	+ 11,37	– 58,85	+ 18,07	+ 5,05	+ 1,19	– 57,83	+ 12,44
1,5	1,00	**10,30**	**92,60**	**8,88**	**96,04**	**33,34**	**93,22**	**62,39**	**98,57**	**29,23**	**20,60**	**64,26**	**54,49**
	1,25	– 2,39	+ 1,39	– 1,63	+ 2,12	– 0,76	+ 1,29	– 52,79	+ 0,83	+ 2,20	+ 0,71	– 58,19	+ 13,65
	1,50	– 4,08	+ 2,27	– 3,25	+ 2,98	– 1,54	+ 2,08	– 62,21	+ 0,84	+ 3,63	+ 1,01	– 64,22	+ 15,69
	1,75	– 5,31	+ 2,88	– 4,60	+ 3,37	– 2,32	+ 2,59	– 62,38	+ 0,82	+ 4,55	+ 1,05	– 64,25	+ 16,60
	2,00	– 6,22	+ 3,31	– 5,65	+ 3,58	– 3,08	+ 2,95	– 62,38	+ 0,79	+ 5,10	+ 0,93	– 64,25	+ 16,84

Tableau 3.3 Variations de puissance en fonction de la corrélation gène-environnement

Les puissances asymptotiques en l'absence de corrélation gène-environnement sont en gras ($\theta_{GE} = 1$). Les autres valeurs représentent les variations par rapport à cette valeur de référence : augmentation pour les valeurs positives ou diminution pour les valeurs négatives. Toutes ces valeurs ont été multipliées par 10^2. OR_G et OR_I sont respectivement les odds ratio génétique et d'interaction G×E du modèle de pénétrance simulé.

fréquence du facteur génétique, $f_G = 0,3$ (modèle dominant) ; odds ratio environnemental, $OR_E = 1,5$; fréquence de l'exposition, $f_E = 0,3$; prévalence de la maladie, $f_M = 0,1$; erreur de première espèce nominale de 0,01 avec un test bilatéral et pour un échantillon de 500 cas et 500 témoins

3.4. Biais, variance et couverture des estimateurs

Le biais, la variance et la probabilité de couverture des différents coefficients de régression qui mesurent l'effet du facteur *G* sont présentés dans le tableau 3.4. Le paramètre β_{G1} du modèle multinomial a des valeurs de biais similaires à celles du paramètre β_{GCT} du modèle logistique complet et inférieures à celles de β_{GM} et β_{GA} des modèles marginal et ajusté. La variance de β_{G1} est inférieure à celle de β_{GCT} mais supérieure à celles de β_{GM} et β_{GA}. Les probabilités de couverture des intervalles de confiance à 95 % des paramètres β_{G1} et β_{GCT} sont très proches de la valeur attendue de 95 % mais diminuent avec des valeurs croissantes de OR_G, tandis que celles de β_{GM} et β_{GA} sont inférieures pour tous les modèles, tout particulièrement pour des valeurs élevées de OR_E et OR_I.

Un estimateur de l'interaction G×E peut être calculé pour le modèle multinomial en faisant le rapport de β_{G2} sur β_{G1}. Cette estimateur $\beta_{IM} = \beta_{G2} / \beta_{G1}$ a exactement les mêmes valeurs de biais, de variance et de probabilité de couverture que le paramètre d'interaction G×E du modèle cas-seuls, β_{ICS}. Par rapport à l'estimateur β_{ICT} du modèle logistique complet, β_{IM} présente des biais plus élevés et des probabilités de couverture plus faibles lorsque OR_I augmente. Sa variance est par contre plus faible que celle de β_{ICT}.

OR_G	OR_I	OR_E	Biais au carré × 10^2				Variance × 10^2				Probabilité de couverture × 10^2			
			β_{GM}	β_{GA}	β_{GCT}	β_{G1}	β_{GM}	β_{GA}	β_{GCT}	β_{G1}	β_{GM}	β_{GA}	β_{GCT}	β_{G1}
0,5	1,0	1,0	0,18	0,17	0,19	0,18	2,30	2,31	3,30	2,88	95,7	95,6	94,6	95,1
		2,0	0,41	0,38	0,20	0,21	2,29	2,35	3,80	3,39	91,3	91,5	93,9	93,9
	2,0	1,0	8,20	7,51	0,22	0,19	2,12	2,14	3,40	2,98	50,0	52,9	94,9	93,9
		2,0	14,44	10,17	0,14	0,13	2,06	2,16	3,97	3,55	23,4	40,9	95,2	96,2
1,0	1,0	1,0	0,00	0,00	0,00	0,00	1,91	1,91	2,74	2,32	95,1	94,9	95,6	94,9
		2,0	0,00	0,00	0,00	0,00	1,91	1,96	3,08	2,67	94,5	94,4	94,9	94,3
	2,0	1,0	5,04	4,39	0,00	0,00	1,84	1,86	2,84	2,43	62,2	67,3	94,8	95,6
		2,0	7,53	4,81	0,01	0,00	1,83	1,91	3,25	2,84	46,9	65,7	94,8	95,0
1,5	1,0	1,0	0,19	0,18	0,21	0,18	1,81	1,81	2,58	2,17	94,0	94,2	92,3	94,2
		2,0	0,26	0,23	0,15	0,13	1,81	1,85	2,88	2,47	93,2	93,6	94,5	95,2
	2,0	1,0	2,42	1,99	0,27	0,26	1,77	1,79	2,69	2,28	79,0	81,5	93,2	93,2
		2,0	3,00	1,57	0,26	0,29	1,77	1,84	3,04	2,63	73,9	84,7	92,7	93,9
2,0	1,0	1,0	0,94	0,92	0,94	0,97	1,77	1,77	2,53	2,12	87,9	88,1	91,0	88,8
		2,0	1,31	1,18	0,58	0,66	1,77	1,81	2,81	2,39	84,8	86,4	92,1	92,2
	2,0	1,0	1,19	0,91	0,67	0,72	1,76	1,77	2,65	2,23	88,7	91,1	92,1	91,8
		2,0	0,92	0,34	0,77	0,78	1,76	1,82	2,95	2,54	88,0	92,5	91,5	90,4

Tableau 3.4 Biais au carré, variance et probabilité de couverture des estimateurs du coefficient de régression génétique

fréquence du facteur génétique, $f_G = 0,3$ (modèle dominant) ; fréquence de l'exposition, $f_E = 0,3$; corrélation gène-environnement, $\theta_{GE} = 1$; prévalence de la maladie, $f_M = 0,1$; échantillon de 500 cas et 500 témoins

OR_G, OR_E et OR_I sont respectivement les odds ratio génétique, environnemental et de l'interaction G×E du modèle de pénétrance simulé.

β_{GM} est le coefficient de régression génétique du modèle marginal; β_{GA} est celui du modèle avec ajustement sur le facteur environnemental; β_{GCT} est celui du modèle complet; β_{G1} est celui du modèle multinomial.

OR_I	OR_G	OR_E	Biais au carré × 10^2			Variance × 10^2			Probabilité de couverture × 10^2		
			β_{ICT}	β_{ICS}	β_{IM}	β_{ICT}	β_{ICS}	β_{IM}	β_{ICT}	β_{ICS}	β_{IM}
0,5	1,0	1,0	0,22	0,21	0,21	11,00	6,41	6,41	94,7	94,7	94,7
		2,0	0,71	0,53	0,53	9,53	4,91	4,91	93,0	93,1	93,1
	2,0	1,0	0,83	0,67	0,67	9,51	4,90	4,90	92,6	93,1	93,1
		2,0	0,84	0,58	0,58	8,45	3,84	3,84	93,9	93,1	93,1
1,0	1,0	1,0	0,04	0,00	0,00	9,17	4,60	4,60	95,5	94,2	94,2
		2,0	0,01	0,00	0,00	8,52	3,91	3,91	95,3	93,8	93,8
	2,0	1,0	0,00	0,00	0,00	8,50	3,91	3,91	95,6	95,4	95,4
		2,0	0,78	0,62	0,62	7,94	3,35	3,35	94,3	91,7	91,7
1,5	1,0	1,0	0,17	0,16	0,16	8,65	4,05	4,05	93,8	95,6	95,6
		2,0	0,29	0,48	0,48	8,26	3,65	3,65	92,7	92,8	92,8
	2,0	1,0	0,53	0,63	0,63	8,25	3,66	3,66	93,4	92,2	92,2
		2,0	3,49	3,97	3,97	7,87	3,26	3,26	88,4	81,0	81,0
2,0	1,0	1,0	0,63	0,80	0,80	8,43	3,82	3,82	94,0	93,6	93,6
		2,0	1,63	2,22	2,22	8,18	3,57	3,57	92,6	87,6	87,6
	2,0	1,0	2,44	2,33	2,33	8,16	3,57	3,57	91,7	86,1	86,1
		2,0	9,64	9,86	9,86	7,84	3,25	3,25	79,8	59,9	59,9

Tableau 3.5 Biais au carré, variance et probabilité de couverture des estimateurs du coefficient d'interaction gène-environnement

fréquence du facteur génétique, $f_G = 0,3$ (modèle dominant) ; fréquence de l'exposition, $f_E = 0,3$; corrélation gène-environnement, $\theta_{GE} = 1$; prévalence de la maladie, $f_M = 0,1$; échantillon de 500 cas et 500 témoins

OR_G, OR_E et OR_I sont respectivement les odds ratio génétique, environnemental et de l'interaction G×E du modèle de pénétrance simulé.

β_{ICT} est le coefficient d'interaction G×E du modèle complet (cas-témoins) ; β_{ICS} est celui du modèle cas-seuls et β_{IM} est celui du modèle multinomial.

Chapitre 3
Comparaison de méthodes de liaison et/ou d'association génétique pour prendre en compte les interactions gène-environnement

En 2006, les données simulées du *Genetic Analysis Workshop* 15 (GAW 15) nous ont offert l'opportunité de réaliser une première comparaison de quatre méthodes destinées à tester les effets génétiques et les interactions G×E. Nous avons initialement sélectionné une méthode de liaison utilisant des paires de germains atteints, le *Mean Interaction Test* (MIT), une modélisation log-linéaire utilisant des données familiales de type trios (MLL) et deux méthodes d'association utilisant des données de population cas-témoins (CT) ou cas-seuls (CS). L'objectif de ce travail était principalement de comparer la puissance, l'erreur de première espèce et les estimations des paramètres entre ces méthodes.

Afin de compléter cette première étude, nous avons également étudié ces données en utilisant l'approche MInT afin de voir si les résultats obtenus avec ces données simulées à partir d'un modèle plus complexe de maladie étaient en accord avec ceux de notre étude de simulation d'un modèle simple. Nous comparons dans ce chapitre uniquement les puissances et erreurs de première espèce des quatre méthodes initialement choisies avec celles du test MInT. L'ensemble des autres résultats se trouve dans l'article présenté en annexe 3 (Kazma *et al.* 2007).

1. Matériel

Les données du problème 3 du GAW 15 ont été simulées à partir d'un modèle complexe (figure 3.6) imitant les effets génétiques et environnementaux observés dans la polyarthrite rhumatoïde (PR). Ces données étaient composées de 100 réplicats contenant des informations sur 1500 familles nucléaires (2 parents génétiquement indépendants et 2 enfants atteints) et 2000 témoins. Nous nous sommes intéressés à l'interaction entre le locus B et le tabagisme. Les covariables épidémiologiques incluaient le tabagisme sur la vie entière. Cette variable a été simulée de telle façon à ce que sa variance soit expliquée par des effets polygéniques additifs à 50 %, par des facteurs environnementaux partagés dans la famille à 40 % et à des facteurs environnementaux non partagés à 10 %. Un modèle à seuil dépendant de l'âge a permis d'attribuer un statut fumeur ou non fumeur à chaque individu. Par ailleurs, le

tabagisme influait sur le risque de maladie indirectement et de façon indépendante du locus B en modulant la variance des Immunoglobulines M (IgM) mais nous n'avons pas pris en compte cet effet dans nos analyses.

Le modèle de risque de PR incluait neuf locus génétiques dont le locus B qui interagissait avec le tabagisme. Ce locus pouvait présenter 2 allèles, *B* et *b*. *B* avait une fréquence de 0,35 et les génotypes *B*||*b* et *B*||*B* multipliaient le risque de PR par 1,5 chez les fumeurs uniquement. Nous avons utilisé l'identité par descendance (IBD) des paires de germains pour le test de liaison et les génotypes des individus à ce locus pour tous les autres tests. Ces données étaient fournies dans les fichiers de « réponse ». Enfin, pour le test MInT, nous avons combiné les variables PR et tabac chez les cas pour créer la variable multinomiale M_P égale à 0 chez les témoins, 1 chez les cas non fumeurs et 2 chez les cas fumeurs.

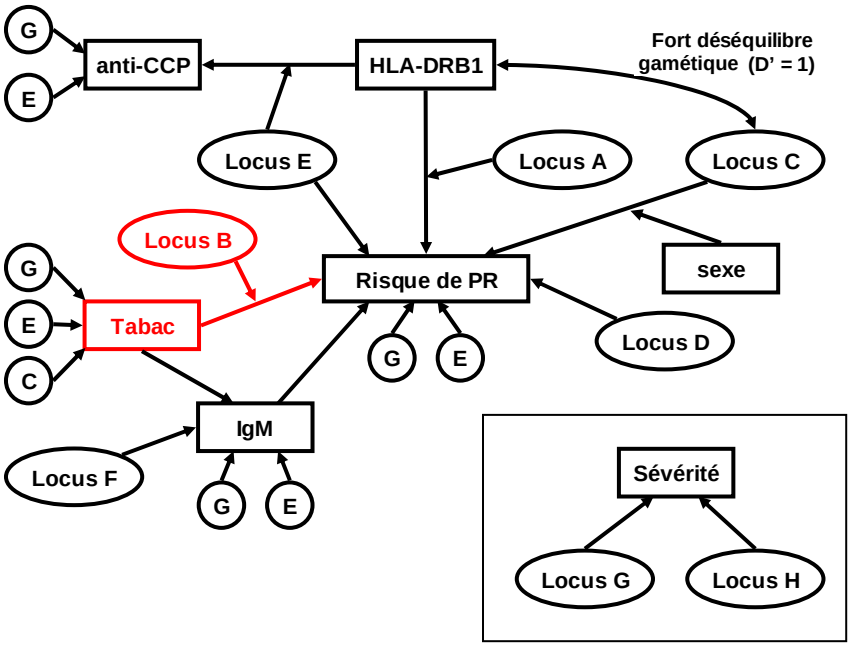


Figure 3.6 Modèle de simulation des données du problème 3, GAW15

G : représente la part de variance due aux effets polygéniques additifs

E : représente la part de variance due à l'environnement partagé en famille

C : représente la part de variance due à des facteurs environnementaux non partagés

PR : Polyarthrite rhumatoïde

Dans chacun des 100 réplicats, nous avons utilisé 1500 paires de germains atteints pour le MIT, 1500 trios (2 parents et 1 enfant atteint) pour le MLL, 1500 cas et 1500 témoins pour les tests d'association cas-témoins et pour MInT et les 1500 cas seulement pour le test d'interaction cas-seul. Les cas ont été obtenus en choisissant le premier germain atteint de chaque fratrie et les 1500 témoins ont été sélectionnés en prenant les 1500 premiers témoins du fichier qui en contenait 2000.

2. Méthodes

Toutes les méthodes ont été présentées dans la partie 1. Voici brièvement les différents tests qui ont été effectués.

2.1. Liaison génétique sur des paires de germains atteints

Le modèle de régression que nous avons utilisé pour le MIT est le suivant

$$\pi_i = \pi + \beta_{EE}(X_{EE} - E(X_{EE})) + \beta_{EU}(X_{EU} - E(X_{EU})) + \varepsilon_i \qquad (3.15)$$

Ce codage permet de contraster les paires de germains respectivement exposés (*EE*) et discordantes (*EU*) pour l'exposition par rapport aux paires non exposées (*UU*). X_{EE} et X_{EU} sont les covariables indicatrices de l'exposition des paires de germains centrées sur leurs moyennes $E(X_{EE})$ et $E(X_{EU})$. Ainsi pour une paire de germains exposés, $X_{EE} = 1$ et $X_{EU} = 0$; pour une paire de germains discordants, $X_{EE} = 0$ et $X_{EU} = 1$ et pour une paire de germains non exposés $X_{EE} = 0$ et $X_{EU} = 0$.

Nous avons effectué un test de liaison génétique seule (équation 1.30), un test d'interaction G×E en présence de liaison génétique (équation 1.31) et un test conjoint de la liaison génétique et de l'interaction G×E (équation 1.32).

2.2. Liaison et association génétique sur des données trios

Le modèle log-linéaire appliqué à des données de trios a été décrit dans la partie 1 (pages 68-70). Nous avons effectué les tests de rapport de vraisemblances permettant de tester l'effet du locus seul (équation 1.43), l'effet conjoint du locus et de l'interaction G×E (équation 1.44) et l'effet de l'interaction G×E seule (équation 1.45).

2.3. Association génétique sur des données de population

A partir des données cas-témoins, les trois tests décrits dans la partie 1 (pages 65-68) ont été réalisés: le test d'association du locus seul (équation 1.36), le test conjoint de l'association au locus et de l'interaction G×E (équation 1.38) et le test de l'interaction G×E seule (équation 1.39). Sur les données cas-seuls, nous avons testé l'interaction G×E à partir de l'équation 1.41 (page 67). Le test MInT a été décrit au chapitre précédent (équation 3.7, page 109) et a été utilisé ici sur les données cas-témoins en ignorant l'exposition des témoins.

3. Résultats

Le tableau 3.6 montre la puissance et l'erreur de première espèce des différents tests effectués. Les erreurs de première espèce ont été estimées à partir des données concernant les 7 locus qui n'interagissaient pas avec le tabac pour les tests d'interaction G×E. Elles varient de 4 % (locus G) à 38 % (locus C) sur les données cas-témoins, de 4 % (locus A, D, G et H) à 16 % (locus C) pour les données cas-seuls et de 3 % (locus A et F) à 23 % (locus C) sur les données cas-témoins sans information sur l'exposition des témoins.

Les trois tests fondés sur la liaison génétique (MIT) ont une puissance faible ce qui est cohérente avec les moyennes des proportions d'allèles IBD dans l'échantillon total et dans chacune des trois catégories d'exposition de paires de germains (*EE*, *EU* et *UU*) présentées dans le tableau 3.7. En effet, ces proportions sont toutes très proches de la valeur attendue sous l'hypothèse d'absence de liaison génétique (0,5).

Le test du MLL avec des données trios a une puissance de 78 % pour détecter l'effet du locus B qui augmente à 87 % lorsque l'interaction G×E est prise en compte. Il y a donc un gain de puissance pour détecter le locus lorsque l'interaction avec le tabac est prise en compte. Pour les données cas-témoins, la puissance de détection du locus B est de 95 % et passe à 98 % lorsque l'interaction est prise en compte. Le test MInT a une puissance de 98 %.

Concernant la détection de l'interaction G×E seule, le test utilisant des données cas-seuls est le plus puissant (95 %) suivi du test sur des données cas-témoins (69 %) et du test d'interaction sur les données cas-témoins sans information sur l'exposition des témoins (53 %). Le test d'interaction G×E en présence de liaison n'avait qu'une puissance de 12 %.

Locus	**B**			**A**	**C**	**D**	**E**	**F**	**G**	**H**
Test	**G**	**GI**	**I**	**Erreur de première espèce (test I)**						
MIT	8	6	12	–	–	–	–	–	–	–
MLL	78	87	53	3	23	6	8	3	6	4
CT	95	98	69	4	16	4	9	8	4	4
CS	–	–	95	13	38	24	18	6	4	5
MInT	–	98	–	–	–	–	–	–	–	–

Tableau 3.6 Puissance et erreurs de première espèce des cinq tests.

MIT : *mean interaction test* ; MLL : modèle log-linéaire ; CT : cas-témoins ; CS : cas-seuls ; MInT : *multinomial interaction test*. Les trois premières colonnes indiquent la puissance en pourcentage de détection de l'effet du locus B seul (G), de l'effet du locus B en prenant en compte son interaction avec le tabagisme (GI) et de l'interaction seule (I). Les sept colonnes de droite indiquent l'erreur de première espèce des tests d'interaction gène-environnement (I) aux 7 locus n'interagissant pas avec le tabac. Ces erreurs de type 1 n'ont pas été estimées pour le test MInT car nous avons montré que le test d'interaction de ce modèle était identique au schéma CS. Elles n'ont pas non plus été calculées pour le MIT car les puissances de ce test étant très faibles, nous n'avons pas jugé nécessaire de le faire.

	π	π_{UU}	π_{EU}	π_{EE}
moyenne	0,502	0,500	0,501	0,503
écart-type	0,008	0,018	0,018	0,013
minimum	0,485	0,464	0,455	0,480
maximum	0,525	0,543	0,543	0,539

Tableau 3.7 Proportion d'allèles IBD chez les paires de germains atteints

Moyenne, écart-type, valeurs minimum et maximum des proportions d'allèles IBD chez les 1500 paires de germains atteints sur les 100 réplicats étudiés. π est la proportion de l'échantillon total et π_{UU}, π_{EU} et π_{EE} sont les proportions chez les paires de germains avec 0, 1 ou 2 germains exposés, respectivement.

4. Discussion

L'objectif initial de cette étude était de comparer les méthodes fondées sur la liaison génétique et celles fondées sur l'association pour prendre en compte ou détecter une interaction G×E. Sous le modèle particulier qui a été simulé dans ces données, les méthodes d'association utilisant des échantillons populationnels ou familiaux sont bien plus puissantes que la méthode utilisant la liaison. Les puissances de la méthode MIT sont très faibles que ce soit pour détecter l'effet du facteur génétique, l'interaction G×E ou les deux conjointement. Ceci est en accord avec les résultats de Gauderman et Siegmund (2000) qui montraient que pour un coefficient d'interaction inférieur à 3 (ou supérieur à 1/3), les tests fondés sur la liaison sont peu puissants mais également avec les résultats comparant les méthodes de liaison et d'association pour tester le facteur génétique seulement (Risch et Merikangas, 1996).

La prise en compte de l'interaction G×E dans les tests d'association améliore la puissance de détection du facteur génétique, mais ce gain de puissant est assez limité (3 % pour le test sur les données cas-témoins et 9 % pour le test sur les données cas-témoins sans exposition chez les témoins). Ceci peut être expliqué par les niveaux relativement élevés des puissances des tests ne prenant pas en compte l'interaction. Les génotypes à risques (*B*||*B* et *B*||*b*) avaient une fréquence, f_G de 0,58 et le tabagisme avait une fréquence, f_E de 0,46 chez les témoins. Ces fréquences élevées peuvent également expliquer le gain de puissance peut important apporté par la prise en compte de l'interaction G×E (Selinger-Leneman *et al.* 2003).

Le test MInT a pour ce modèle une puissance de détection du locus B en prenant en compte son interaction avec le tabac de 98 %. Le modèle qui a été simulé consistait en une interaction pure sans effets indépendants des facteurs *G* et *E* sous un modèle dominant. D'après nos simulations présentées au chapitre précédent, dans ce cas de figure, le test d'interaction G×E sur des données pour lesquelles on dispose de l'information sur l'exposition des cas et des témoins et le test MInT ont des puissances quasiment égales qui sont supérieures aux puissances du test conjoint (Kraft *et al.* 2007), qui a lui-même une puissance supérieure au test du facteur génétique seul. Ici, ces trois tests ont des puissances équivalentes (entre 95 et 98 %) car les effectifs utilisés (1500 cas et 1500 témoins) sont trop grands pour discriminer ces tests. Par ailleurs, ces données ont été simulées à partir d'un modèle plus complexe que celui que nous avions utilisé pour nos simulations.

Chapitre 4
Réactions cutanées sévères et interactions gène-médicament

Le registre Européen des réactions cutanés sévères secondaires aux médicaments (RegisCAR : *European Registry of Severe Cutaneous Adverse Reactions to Drugs*) a été mis en place par un réseau de chercheurs travaillants sur ce type de réactions telles que le syndrome de Stevens-Johnson ou le syndrome de Lyell (http://regiscar.uni-freiburg.de/). Ce réseau maintient un registre multinational qui représente le plus gros échantillon de cas dans le monde avec à la fois des informations épidémiologiques relatives à la sévérité de la maladie, aux antécédents de prises médicamenteuses et des informations génétiques pour près de 600 malades. Ne disposant pas d'un échantillon de témoins spécifiquement collecté, l'étude pangénomique a été réalisée avec l'échantillon de cas provenant de RegiSCAR et un sous-échantillon du panel de témoin européen EuroPa génotypé par le Centre National de Génotypage Français (Heath *et al.* 2008). Au travers de cet exemple, nous montrons l'intérêt de l'approche multinomiale pour prendre en compte une interaction G×E lorsque le facteur d'exposition, ici le médicament, n'est pas documenté chez les témoins.

Après avoir introduit le contexte général des réactions cutanées sévères consécutives à la prise de médicaments, nous présentons les échantillons qui ont été analysées, puis les résultats de l'analyse pangénomique. Cette étude a été réalisée en utilisant MInT afin de tester l'association des SNPs avec la maladie en prenant en compte une potentielle hétérogénéité en fonction du médicament incriminé. Nous terminons cette partie par une discussion sur les avantages et les limites de cette approche et nous y discutons les résultats de cette application.

1. Contexte

Les effets secondaires des traitements médicamenteux sont un problème de santé publique majeur tant en terme de morbidité que de mortalité (Lazarou *et al.* 1998) et les lésions cutanées en sont l'une des formes les plus fréquentes représentant pour certaines molécules jusqu'à 10 % des réactions observées (Svensson *et al.* 2001; Wolverton 1997). Le syndrome de Stevens-Johnson (SSJ) et le syndrome de Lyell (SL) sont des réactions cutanées très sévères caractérisées par le développement d'un exanthème aigu limité (dans le SSJ) ou diffus (dans le SL), de lésions bulleuses avec des décollements et une érosion cutanéo-

muqueuse (Roujeau et Stern 1994). Le SSJ et le SL sont considérés comme étant deux formes différentes de la même maladie, le SL ou nécrolyse épidermique toxique étant la forme la plus sévère des deux (Auquier-Dunant *et al.* 2002). L'incidence de ces deux syndromes réunis est estimée à environ 2 cas par million d'individus par an (Rzany *et al.* 1996). Il s'agit donc d'une maladie rare mais qui est associée à une morbidité et une mortalité (20 à 25 %) importantes et qui nécessite un traitement lourd et intensif. Le développement d'un SSJ ou d'un SL n'est pas spécifiquement associé à une seule molécule ou à une seule classe de médicaments mais peut survenir suite à la prise de centaines de médicaments différents. Certains sont cependant plus souvent associés avec la maladie. Ces médicaments « à risque » incluent sulfonamides (particulièrement sulfamethoxazole), allopurinol, carbamazepine, lamotrigine, phénobarbital, phénytoine, anti-inflammatoires non stéroïdiens de la catégorie des oxicam et neviparine (Mockenhaupt *et al.* 2008; Roujeau et Stern 1994). En Europe, l'allopurinol est la molécule la plus fréquemment associée à ces syndromes (Lonjou *et al.* 2008). Cependant, seul un petit nombre d'individus exposés à ces médicaments « à risque » développent la maladie et une susceptibilité génétique est fortement suspectée (Melsom 1999; Pellicano *et al.* 1992; Pirmohamed et Park 2001; Roujeau *et al.* 1987).

Une association avec HLA a été documentée depuis une vingtaine d'années (Roujeau *et al.* 1986; Roujeau *et al.* 1987) et plus récemment, des études ont montré que certains allèles de HLA-B entraînent une augmentation du risque de la maladie avec certains médicaments. Par exemple, des associations spécifiques ont été mises en évidence entre l'allèle HLA-B*5701 et un SSJ/SL induit par l'abacavir (Hetherington *et al.* 2002; Mallal *et al.* 2002), entre l'allèle HLA-B*1502 et un SSJ/SL induit par la carbamazepine (Chung *et al.* 2004) et finalement entre l'allèle HLA-B*5801 et un SSJ/SL induit par l'allopurinol (Hung *et al.* 2005). Ces deux dernières études, réalisées dans la population de Chinois Han, montrent des associations très fortes. En effet, l'allèle HLA-B*1502 était porté par la totalité des 44 cas de SSJ/SL induit par la carbamazépine et de façon similaire, l'allèle HLA-B*5801 était présent chez tous ceux dont le SSJ/SL était induit par l'allopurinol. L'investigation des associations de HLA-B dans un échantillon Européen n'a pas retrouvé l'association entre l'allèle HLA-B*1502 et le SSJ/SL induit par la carbamazépine mais a mis en évidence celle, bien qu'incomplète, entre HLA-B*5801 et le SSJ/SL induit par l'allopurinol (Lonjou *et al.* 2008; Lonjou *et al.* 2006).

Plusieurs autres gènes candidats impliqués dans la réponse immunitaire, l'inflammation ou le métabolisme des médicaments ont aussi été étudiés sans qu'aucune de ces associations ne soit confirmée (Abe 2008; Miller 2008; Pirmohamed 2006; Pirmohamed *et al.* 2007). Plutôt que de se limiter à une stratégie gène candidat, une étude d'association pangénomique pourrait apporter de nouvelles perspectives sur la susceptibilité génétique de ces syndromes et permettre d'identifier des gènes que l'on n'aurait *a priori* pas soupçonnés (Pirmohamed 2006).

2. Le projet RegisCAR

2.1. Matériel

Un total de 563 cas (226 hommes et 337 femmes, sexe ratio = 0,67) ont été inclus dans cette étude. Ils ont été collectés dans le cadre du projet RegiSCAR dans 6 pays : 331 cas en Allemagne, 1 cas en Autriche, 184 cas en France, 14 cas en Israël, 26 cas en Italie et 7 cas aux Pays-Bas. Tous les sujets avaient été diagnostiqués pour un SSJ (268 cas, 47,6 %), une forme intermédiaire FI (181 cas, 32,2 %) ou un SL (114 cas, 20,2 %) validés par le comité d'expert de RegiSCAR. Pour chaque individu, un consentement éclairé a été obtenu et un échantillon de sang a été prélevé pour extraction d'ADN au Centre d'Étude des Polymorphismes Humains – Fondation Jean Dausset (CEPH) (http://www.cephb.fr). Les 563 patients ont été génotypés sur une puce Illumina 317K au Centre National de Génotypage Français (CNG) (http://www.cng.fr).

Les témoins ont été sélectionnés à partir du Panel de Référence Européen EuroPa collecté par le CNG (Heath *et al.* 2008). Ce panel inclut 5847 individus non apparentés provenant de 13 pays européens génotypés sur la puce Illumina 317K. Étant donné que la majorité des patients recrutés par le projet RegiSCAR provenait de France et d'Allemagne, seuls les 1881 individus du panel EuroPa originaires de ces deux pays (653 d'Allemagne et 1228 de France) ont été inclus dans l'analyse.

2.2. Contrôle de qualité

Après un contrôle de qualité strict réalisé en utilisant le logiciel Plink version 1.05 (Purcell *et al.* 2007), un total de 495 cas et 1881 témoins ont été conservés pour lesquels les génotypes de 268 818 SNPs autosomiques étaient disponibles. Les critères du contrôle de qualité étaient un taux de génotypage individuel supérieur à 95 %, un taux de données manquantes par SNPs inférieur à 5 % avec une différence non significative entre les cas et les témoins, et l'absence de monomorphisme des SNPs chez les cas et les témoins. La stratification de population dans l'échantillon a été étudiée par une Analyse en Composante Principale (ACP) (Patterson *et al.* 2006; Price *et al.* 2006) des données génotypiques des 495 cas et 1881 témoins pour un sous-panel de 35 232 SNPs sélectionnés en minimisant le déséquilibre gamétique. À l'issue de cette analyse, 71 cas ont été exclus, pour lesquels une origine Africaine ou Asiatique est fortement suspectée.

2.3. Caractéristiques de l'échantillon

L'analyse finale a donc été réalisée sur les 424 cas de SSJ/SL et les 1881 témoins. Parmi ces 424 cas, 59,9 % étaient des femmes, 47,2 % avaient un SSJ, 34,4 % une forme intermédiaire FI et 18,4 % un SL (tableau 3.8). Il n'y avait pas de différence entre les distributions phénotypiques des hommes et des femmes mais par contre des différences ont été retrouvées en fonction du médicament et du pays de recrutement. Il y a un excès de SSJ parmi les cas induits par l'allopurinol par rapport au reste des cas ($\chi^2 = 7{,}89$; $p = 0{,}019$) et un excès de SL parmi les cas recrutés en France par rapport à ceux recrutés en Allemagne ($\chi^2 = 23{,}68$; $p = 7{,}2 \times 10^{-6}$). Les prises médicamenteuses précédant la réaction cutanée ont été documentées pour tous les malades. Un médicament a été considéré « inducteur » de la maladie lorsque sa prise avait débuté moins de 42 jours et avait eu lieu entre 4 et 10 jours avant le début des symptômes de la maladie et qu'aucun autre médicament potentiellement inducteur n'a été pris pendant la même période. Le médicament retrouvé le plus fréquemment dans cet échantillon était l'allopurinol chez 54 cas (12,7 %). Les autres médicaments ont été retrouvés dans un nombre beaucoup plus petit de cas (21 cas pour la carbamazepine ou 19 pour la phénytoine par exemple). L'analyse a donc été restreinte à l'allopurinol pour éviter des problèmes liés aux petits effectifs.

	SSJ		FI		SL		Total
Total	200	(47,2)	146	(34,4)	78	(18,4)	424
Femmes	121	(47,6)	86	(33,9)	47	(18,5)	254
Hommes	79	(46,5)	60	(35,3)	31	(18,2)	170
Allopurinol	33	(61,1)	19	(35,2)	2	(3,7)	54
Non allopurinol	167	(45,1)	127	(34,3)	76	(20,6)	370
France	41	(37,3)	32	(29,1)	37	(33,6)	110
Allemagne	145	(34,2)	97	(22,9)	35	(8,3)	277
Autres pays	14	(3,3)	17	(4,0)	6	(1,4)	37

Tableau 3.8 Description du sous-échantillon de 424 cas sélectionnés pour l'analyse

Le nombre de cas (et le pourcentage) pour les différents phénotypes de la maladie est présenté pour l'échantillon total et en fonction du sexe, de la prise d'allopurinol et du pays où les cas on été recrutés. SSJ : Syndrome de Stevens-Johnson ; FI : Forme intermédiaire ; SL : Syndrome de Lyell

3. Étude d'association pangénomique utilisant MInT

L'association a été testée à chaque marqueur en utilisant un modèle de régression logistique multinomial implémenté dans Stata SE version 10.1 (StataCorp. 1984-2008) en ajustant sur les deux premières composantes principales de l'ACP obtenue par Eigenstrat (Price *et al.* 2006). Le statut maladie a été recodé en trois catégories : 0 pour les témoins, 1 pour les cas induits par allopurinol et 2 pour les cas non-induits par allopurinol. À chaque SNP, un modèle additif a été considéré et deux odds ratios ont été calculés modélisant l'effet du SNP chez les malades induits par allopurinol (OR_1) et chez les malades induits par un autre médicament (OR_2). L'hypothèse nulle testée par MInT est l'absence d'effet du SNP dans chacune des deux strates : $OR_1 = OR_2 = 1$. L'effet marginal des SNPs (OR_G) a également été estimé par un modèle de régression logistique tel que présenté page 66 en ajustant sur les deux premières composantes principales de l'ACP. Le test de l'effet marginal du SNP dont l'hypothèse nulle est $OR_G = 1$ et le test d'interaction G×E dont l'hypothèse nulle est $OR_1 = OR_2$ ont été effectués. Ce dernier test est équivalent au test d'interaction G×E avec un échantillonnage cas-seuls.

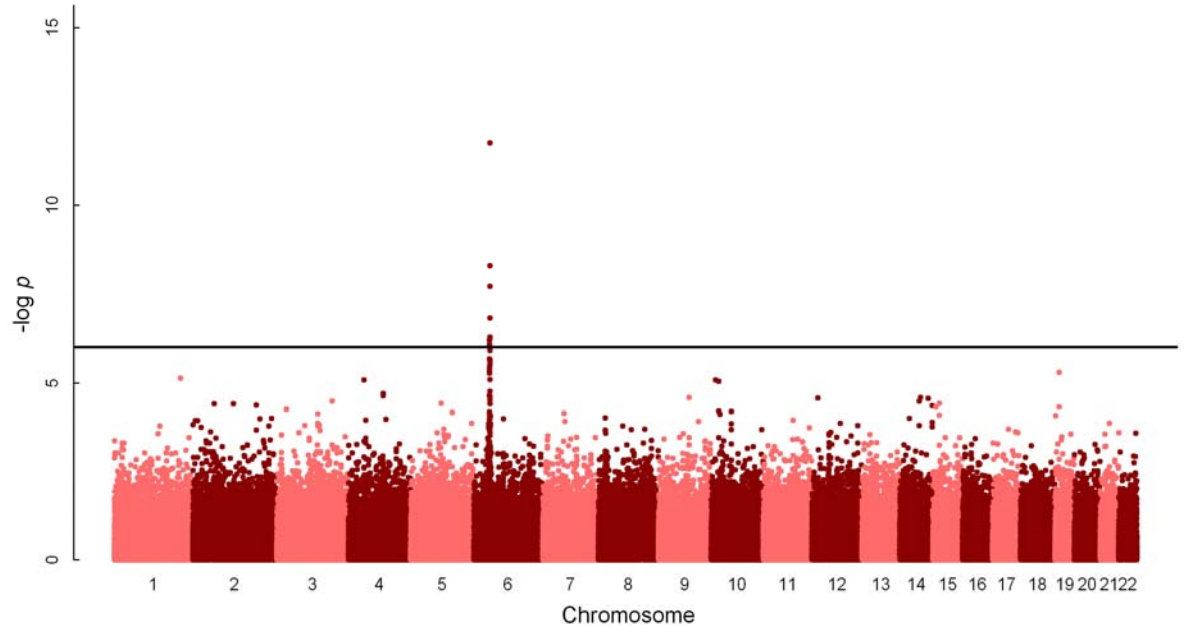


Figure 3.7 Résultats de l'étude d'association pangénomique utilisant MInT

Les valeurs de –log (*p*) du test MInT ajusté sur les deux premières composantes principales pour les 268 818 SNPs autosomiques sont représentées pour les 22 chromosomes.
La ligne horizontale représente un seuil de valeur de *p* à 10^{-6}.

En prenant un seuil de significativité pour la valeur de *p* à 10^{-6}, l'analyse pangénomique des 424 malades SSJ/SL et des 1881 témoins a identifié huit SNPs significativement associés à la maladie (figure 3.7). Ces huit SNPs sont tous situés dans la région HLA localisée sur le bras court du chromosome 6 (tableau 3.9). Le SNP le plus significatif, rs9469003, est localisé à la position 31 515 807 à environ 85kb en amont du locus HLA-B. Les odds ratios d'association de ce SNP chez les malades induits par allopurinol (OR_1) et chez les malades induits par un autre médicament (OR_2) sont de 4,03 (IC à 95 % = [2,72 – 5,97]) et de 1,49 (IC à 95 % = [1,22 – 1,82]) respectivement. L'odds ratio marginal de ce SNP en ne prenant pas en compte une hétérogénéité en fonction du médicament inducteur est de 1,73 (IC à 95 % = [1,44 – 2,08]). La valeur *p* du test MInT est de $1{,}6 \times 10^{-12}$ alors que celles des tests de l'effet du SNP seul et de l'interaction G×E seule sont de $1{,}6 \times 10^{-9}$ et $3{,}1 \times 10^{-6}$ respectivement.

SNP Allèle à risque (Autre allèle)	Position (Build36)	q_t	OR_1	IC à 95 %	OR_2	IC à 95 %	p_{MInT}	p_I	OR_G	IC à 95 %	p_G	p_{HW}	Annotations
rs9469003 C (T)	31515807	0,15	4,03	2,72 – 5,97	1,49	1,22 – 1,82	$1,6 \times 10^{-12}$	$3,1 \times 10^{-6}$	1,73	1,44 – 2,08	$1,6 \times 10^{-9}$	0,003	*HCP5* 5'UTR
rs3094188 A (C)	31250224	0,63	2,94	1,79 – 5,00	1,47	1,25 – 1,75	$4,9 \times 10^{-9}$	$5,6 \times 10^{-3}$	1,59	1,34 – 1,88	$2,9 \times 10^{-8}$	0,06	*POU5F1* 5'UTR
rs3130501 G (A)	31244432	0,74	2,78	1,51 – 5,00	1,67	1,35 – 2,04	$1,8 \times 10^{-8}$	0,09	1,74	1,43 – 2,13	$1,7 \times 10^{-8}$	0,58	*POU5F1* Intron
rs3130931 C (T)	31242867	0,69	3,03	1,72 – 5,26	1,43	1,19 – 1,72	$1,4 \times 10^{-7}$	$6,3 \times 10^{-3}$	1,54	1,29 – 1,84	$7,8 \times 10^{-7}$	0,66	*POU5F1* Intron
rs2428486 A (G)	31462083	0,36	2,60	1,75 – 3,85	1,27	1,08 – 1,50	$5,1 \times 10^{-7}$	$5,4 \times 10^{-4}$	1,39	1,19 – 1,62	$2,2 \times 10^{-5}$	0,01	*MICA* 5'UTR
rs2844529 C (T)	31461572	0,36	2,60	1,75 – 3,85	1,27	1,08 – 1,50	$5,3 \times 10^{-7}$	$5,4 \times 10^{-4}$	1,39	1,19 – 1,62	$2,3 \times 10^{-5}$	0,01	*MICA* 5'UTR
rs2844665 C (T)	31114834	0,62	2,08	1,31 – 3,33	1,47	1,25 – 1,75	$6,2 \times 10^{-7}$	0,14	1,54	1,30 – 1,82	$2,7 \times 10^{-7}$	0,77	*C6orf205* 3'UTR
rs3815087 A (G)	31201566	0,21	2,23	1,51 – 3,29	1,43	1,20 – 1,71	$8,4 \times 10^{-7}$	$3,4 \times 10^{-2}$	1,53	1,29 – 1,80	$2,9 \times 10^{-7}$	0,08	*PSORS1C1* UTR

Tableau 3.9 Liste des SNPs significatifs au seuil de $p \leq 10^{-6}$ avec l'approche MInT comparant les groupes de malades induits par allopurinol et de malades induits par un autre médicament au groupe de témoins

Tous les SNPs sont localisés sur le chromosome 6. q_t est la fréquence de l'allèle à risque chez les témoins. OR_1 et OR_2 sont les odds ratios génotypiques en supposant un modèle additif respectivement pour les cas induits par allopurinol et pour les cas induits par un autre médicament. OR_G est l'odds ratio génotypique en supposant un modèle additif et IC 95 % est son intervalle de confiance à 95 %. p_{MInT} correspond au test conjoint ($OR_1 = OR_2 = 1$). p_G correspond au test de l'effet marginal du facteur génétique ($OR_G = 1$). p_I correspond au test d'interaction G×E dont l'hypothèse nulle est $OR_1 = OR_2$. UTR : Région non traduite (*Untranslated Region*). p_{HW} correspond à la valeur p du test d'adéquation aux proportions de Hardy-Weinberg.

Les sept autres SNPs significatifs sont localisés dans une région plus proche du télomère (extrémité du chromosome) à une distance d'environ 250 kb de rs9469003. Ils présentent tous un OR_1 supérieur à OR_2, mais leurs tests d'interaction G×E ne sont pas significatifs au seuil de 10^{-6}. En particulier, deux de ces SNPs (rs2428486 et rs2844529) qui sont en fort déséquilibre gamétique chez les cas et chez les témoins (r^2 = 0,99) ont un test MInT significatif (de l'ordre de 5×10^{-7}) alors que ni leur test d'interaction G×E, ni leur test marginal de l'effet du SNP ne sont significatifs. Par ailleurs, le test de l'effet marginal des SNPs et le test d'interaction G×E n'ont détecté aucun autre SNP.

Discussion

Le modèle de régression logistique multinomiale est un modèle simple et efficace pour comparer les cas exposés et non exposés par rapport à l'échantillon de témoins lorsque l'information sur *E* n'est disponible que chez les cas. Il permet de tester conjointement l'association génétique et l'interaction G×E en combinant au sein du même modèle la mesure marginale de l'effet du facteur *G* et la mesure de l'interaction G×E en n'utilisant que l'exposition des cas. En quelque sorte, le modèle multinomial est la fusion du modèle G–marginal et du modèle I–cas-seuls. Nous montrons ici qu'il est possible de conserver une puissance satisfaisante en répondant aux deux préoccupations suivantes : détecter l'effet d'un facteur *G* même s'il n'y a pas d'interaction G×E et ne pas le manquer à cause d'une interaction avec un facteur *E* particulier. En effet, notre approche a une puissance similaire ou à peine moins importante qu'un test de l'effet de *G* seul en l'absence d'interaction G×E et une puissance au moins égale au test de l'interaction G×E avec un schéma cas-seuls en présence d'une interaction G×E pure (sans effet indépendant de *G*), des situations où chacun de ces deux tests a la meilleur puissance respectivement.

Pour l'ensemble des modèles d'interaction G×E explorés, le test GI–MInT a une meilleure puissance que le test GI–Binomial (Kraft *et al.* 2007) qui utilise les statuts d'exposition des cas et des témoins. Ce résultat inattendu suggère que tant que l'indépendance entre les facteurs *G* et *E* est assurée, l'information d'exposition des témoins n'est pas indispensable pour explorer l'effet du facteur *G* et de son interaction avec *E*. Cependant, cette information devient importante en présence d'une corrélation gène-environnement afin de se prémunir de l'inflation de l'erreur de première espèce que nous avons observée avec les tests GI–MInT et I–cas-seuls. Le défaut de robustesse du schéma cas-seuls en présence d'une corrélation gène-environnement a été largement décrit dans la littérature et est l'une des limitations qui a parfois refreiné son utilisation pour détecter des interactions G×E. Cependant, connaissant ce problème, nous pensons que le test MInT est une alternative à ne pas négliger lorsque l'information sur l'exposition n'est pas disponible chez les témoins comme dans les études d'association réalisées avec des panels de témoins de référence. On peut alors soit se reposer sur les résultats d'études précédentes dans la même population pour exclure une corrélation sous-jacente du facteur *G* avec le facteur exogène étudié, soit étudier l'association entre ces deux facteurs dans une seconde étape au sein d'un échantillon de

témoins plus spécifique pour lequel l'information sur l'exposition est collectée. Cette stratégie est aussi discutée dans la littérature du schéma cas-seuls (Gatto *et al.* 2004; Goldstein et Andrieu 1999; Schmidt et Schaid 1999).

Sous l'hypothèse d'indépendance entre *G* et *E*, les estimateurs du facteur *G* et de l'interaction G×E du modèle multinomial ont une variance plus faible, une précision similaire et donc des probabilités de couverture meilleures que celles obtenues par les estimateurs des modèles qui utilisent le statut d'exposition des témoins.

Tout au long de cette étude, nous avons uniquement exploré des facteurs de risque positivement corrélés ($\theta_{GE} \geq 1$). Les scénarios où les facteurs *G* et *E* sont négativement corrélés, mais interagiraient positivement sur la maladie semblent très peu probables. Nous nous sommes aussi essentiellement intéressés à des modèles génétiques dominants, mais on s'attendrait à observer les mêmes tendances avec des valeurs de puissance globalement inférieures pour des modèles additifs et récessifs. Nous avons également exploré les situations d'interaction G×E où le génotype à risque devient protecteur quand le statut d'exposition change « *flip-flop interaction* ». Dans cette situation, le test GI–MInT est particulièrement intéressant et surpasse toutes les autres méthodes. De telles situations sont rarement décrites dans la littérature, probablement parce que ce n'est pas une situation commune, mais aussi peut-être parce que les approches logistiques traditionnelles n'ont jusque là pas eu assez de puissance pour les détecter.

Umbach et Weinberg (1997) ont proposé une approche fondée sur des modèles de régression log-linéaire qui est semblable au modèle multinomiale que nous proposons. Leur modèle exploite l'hypothèse d'indépendance entre les facteurs *G* et *E* pour modéliser en plus de l'interaction G×E environnement, l'effet indépendant du facteur *E* (ou *G*) avec un schéma cas-témoins où le génotype (ou l'exposition) n'est pas disponible chez les témoins. Leur argumentation en faveur de leur modèle log-linéaire est qu'un modèle logistique ne permet pas d'imposer l'indépendance entre les facteurs *G* et *E* de façon explicite. Cependant, notre méthode a le même schéma d'échantillonnage et un paramétrage semblable. La différence principale entre le modèle log-linéaire et le modèle logistique réside dans le choix du type d'estimateur de risque qui est respectivement le risque relatif et l'odds ratio (Gauderman 2002). L'utilisation d'un modèle logistique est donc plus appropriée à l'étude de données cas-témoins.

Dans l'application que nous avons présentée au chapitre 4, nous avons pu réaliser une étude d'association pangénomique en prenant en compte l'interaction gène-médicament avec un panel de témoins de référence sans information sur leurs prises médicamenteuses. Pour le SNP le plus significatif (rs9469003), la valeur de *p* du test MInT est largement supérieure à celle du test marginal. En effet, l'hétérogénéité est importante car l'association de ce SNP avec la maladie était de 4,03 dans le groupe induit par allopurinol alors qu'elle n'était que de 1,49 dans le groupe non-induit par allopurinol. D'ailleurs, le test d'interaction G×E mesurant l'hétérogénéité entre ces deux odds ratios était également significative. À l'échelle d'une étude pangénomique, le test de l'effet marginal ou le test d'interaction G×E n'ont pas permis de détecter d'autres SNPs par rapport à l'approche MInT, alors que deux des SNPs trouvés avec MInT n'étaient significatifs avec aucun de ces deux autres tests. Cela montre bien une fois de plus que ce test permet d'englober ces deux situations avec un potentiel gain de puissance.

Ces résultats vont dans le sens de ceux déjà publiés et qui ont mis en évidence une association avec des variants du gène HLA-B avec les réactions cutanées sévères secondaires à l'Allopurinol (Hung *et al.* 2005; Lonjou *et al.* 2008). Mis à part les huit SNPs localisés à proximité du gène HLA-B, l'étude pangénomique n'a pas permis de mettre en évidence d'autres associations ailleurs sur le génome. Pour expliquer l'absence de nouvelles associations, on peut évoquer la taille de l'échantillon de cas qui est limitée par la faible prévalence de la maladie. En effet, un échantillon de 500 cas et 2000 témoins ne permet de mettre en évidence que des odds ratios génétiques de plus de 1,8 avec une puissance de 0,80 (Burton *et al.* 2009). Cette estimation suppose en outre l'utilisation de témoins sélectionnés rigoureusement pour l'étude et l'absence d'erreur de génotypage. Comme nous l'avions présenté au premier chapitre, l'utilisation de témoins non spécifiques à l'étude peut entrainer une perte de puissance car certains témoins peuvent avoir eu la maladie au cours de leur vie. Mais les SSJ/SL étant des syndromes rares dans la population, on peut supposer que la perte de puissance est minime.

Conclusion

Si l'importance d'étudier les interactions gène-environnement dans les maladies complexes est souvent mise en avant, peu d'interactions ont jusqu'à présent été confirmées dans différentes études (Hunter 2005; Manolio et al. 2008; Murcray et al. 2009; Selinger-Leneman et al. 2003; Thomas 2010). Par le simple fait de stratifier l'échantillon, la détection des interactions gène-environnement nécessite des effectifs plus grands que ceux nécessaires à la détection des effets marginaux de même ordre de grandeur. Le manque d'outils statistiques puissants et surtout adaptés aux données dont nous disposons constitue donc un obstacle majeur à l'inclusion des interactions dans les analyses génétiques. Par ailleurs, la nécessité de collecter des données environnementales rigoureusement mesurées chez l'ensemble des individus de l'étude semble être une autre difficulté, notamment lorsque les facteurs potentiellement intéressants sont nombreux.

Le travail décrit dans ce manuscrit apporte deux outils statistiques originaux par rapport aux méthodes dont nous disposions au commencement de cette thèse, mais il propose également une stratégie globale cohérente pour étudier les interactions gène-environnement dans les études génétiques des maladies complexes. Dans une première étape, l'utilisation de la méthode EURECA sur des données épidémiologiques de familles de personnes malades permet de concevoir l'étude génétique en sélectionnant les variables environnementales importantes à prendre en compte dans l'analyse et en orientant l'échantillonnage en fonction de critères épidémiologiques particuliers. La seconde étape que nous proposons consiste à ne collecter qu'un échantillon de cas et à utiliser un panel de témoins de référence pour réaliser soit une étude de gènes candidats, soit une étude pangénomique, en utilisant le test MInT pour prendre en compte les facteurs environnementaux qui ont été sélectionnés à l'étape précédente.

Comme nous l'avons vu, EURECA est une méthode dont la finalité n'est pas de détecter l'interaction gène-environnement en tant que telle, mais plutôt de discriminer parmi de nombreux facteurs environnementaux ceux qui sont les plus pertinents à prendre en compte. Le critère que nous avons choisi est la différence entre les risques de récurrence familiale stratifiés en fonction de l'exposition du proposant, mesurée par l'odds ratio de récurrence (*ORR*). Même si cette information peut être biaisée par la présence d'une corrélation familiale du facteur environnemental, nous avons montré qu'il est possible de prendre en compte ce biais dans la construction du test statistique. La méthode EURECA a été développée, évaluée et appliquée à un échantillon de paires de germains, mais il serait

également intéressant d'évaluer ses propriétés lorsqu'un échantillon de paires d'apparentés autres que des germains est utilisé. En effet, même si l'on s'attend à une perte de puissance du test lorsque l'apparentement est plus éloigné, celle-ci pourrait être contrebalancée par la possibilité de collecter des échantillons de taille plus grande. Il serait ainsi possible d'exploiter l'ensemble de la famille d'un proposant, mais la corrélation induite par l'utilisation du même proposant pour plusieurs paires devra alors être prise en compte en utilisant par exemple des méthodes de permutations.

Bien évidemment, ce test demeure un outil parmi d'autres, mais qui est particulièrement utile pour effectuer un tri lorsque l'on ne dispose d'aucun indice permettant de présumer *a priori* qu'un facteur environnemental est important à prendre en compte. En particulier, il est évident que si d'autres arguments indiquent qu'un facteur environnemental appartient à une voie biologique impliquée dans la maladie étudiée, ce facteur devra être pris en compte dans l'analyse, quel que soit son *ORR*. Restreindre la recherche d'interactions gène-environnement aux facteurs génétiques dont les protéines codées interagissent avec des facteurs environnementaux au sein d'une même voie biologique est une alternative intéressante, tout comme celle consistant à ne tester que des variants altérant la fonction d'une protéine en particulier ou régulant sa production. Cependant de telles approches nécessitent de combiner les connaissances multidisciplinaires issues de la génétique moléculaire, de l'épidémiologie, de la bioinformatique, mais aussi de la physiologie et d'autres domaines (Rebbeck et al. 2004). C'est tout un pan de la recherche qui n'en est qu'à ses premiers balbutiements et qui vise à intégrer les informations issues de l'étude de SNPs, de l'ontologie des gènes, des voies biologiques, mais également des facteurs environnementaux en relation avec la maladie, des paramètres métaboliques et des informations toxicologiques. Des approches intégratives automatisées utilisant des bases de données bibliographiques ont été proposées (Jensen et al. 2006; Raychaudhuri et al. 2009), mais sont sujettes aux biais de publication qui sont particulièrement importants dans la littérature des interactions gène-environnement (Thomas 2010).

L'approche multinomiale pour tester conjointement l'association génétique et l'interaction gène-environnement dans les études avec des témoins de référence sans mesure de l'exposition est une méthode facile à mettre en œuvre dans la plupart des logiciels statistiques disponibles (Stata, SAS, R). Ce modèle est flexible puisqu'il permet également de traiter différents sous-phénotypes d'une même maladie (Morris et al. 2009), des variables

environnementales polytomiques, ou même des combinaisons de ces deux situations. Tout comme pour une régression logistique standard, il est possible d'ajuster sur des covariables spécifiques telles que l'âge, le sexe, la catégorie de recrutement ou les axes d'une analyse en composantes principales pour corriger une stratification de population, comme nous l'avons fait dans l'analyse des données RegiSCAR. Ce test est une sorte d'hybride entre un test de l'effet marginal du facteur génétique et un test d'interaction gène-environnement avec un échantillon composé uniquement de cas. Il bénéficie donc de la puissance respective de ces deux tests dans ces deux situations.

En focalisant l'ensemble des ressources disponibles sur la mesure des facteurs environnementaux et le génotypage de l'échantillon de cas uniquement et en utilisant un panel de témoins de référence dont les génotypes sont disponibles sans se soucier de leur exposition, il est possible pour un coût identique de doubler la taille de l'échantillon de cas par rapport à une étude cas-témoin standard. Cependant, comme nous l'avons vu, l'utilisation de témoins de référence n'est pas dénuée d'inconvénients méthodologiques et il n'est pas toujours possible au cours de l'analyse statistique de remédier aux biais dont on pourrait se prémunir par une sélection rigoureuse des témoins en amont. C'est le prix à payer pour l'économie réalisée. Par ailleurs dans certaines situations, le recrutement d'un nombre suffisamment important de cas peut être difficile, comme ce fut le cas pour les réactions cutanées sévères à la suite d'une prise médicamenteuse qui sont des syndromes rares.

L'hypothèse d'indépendance des deux facteurs génétique et environnemental qui permet de s'affranchir de la mesure d'exposition chez les témoins peut ne pas toujours être valide. Mais s'il existe une association entre ces deux facteurs dans la population, deux cas de figures sont possibles. Le facteur génétique peut être responsable d'une prédisposition accrue à l'exposition au facteur environnemental et en tant que tel, ce n'est pas une découverte totalement anodine. C'est le cas par exemple du gène *FTO* (*fat mass and obesity associated gene*) dont l'association retrouvée avec le diabète de type 2 est due à sa corrélation en population générale avec un IMC élevé, l'obésité étant un facteur de risque connu du diabète de type 2 (Dina et al. 2007; Frayling et al. 2007; Freathy et al. 2008). La seconde situation correspond à une corrélation des deux facteurs en population reliée à une stratification de population pour ces deux facteurs. Il s'agit du même problème que celui rencontré dans les études d'association classiques quand, dans l'échantillon étudié, il existe des sous-groupes d'individus avec des prévalences de la maladie et des fréquences alléliques différentes sauf

qu'ici ce sont les fréquences des facteurs génétique et environnemental qui varient dans ces sous-groupes. Si le facteur environnemental étudié entraîne un risque accru de maladie, tout sous-groupe d'individus ayant une prévalence de ce facteur plus élevée sera surreprésenté dans l'échantillon de cas et une interaction entre ce facteur environnemental et les marqueurs génétiques ayant des fréquences alléliques différentes dans ce sous-groupe pourra alors être mise en évidence à tort. Dans cette situation, disposer de l'information environnementale des témoins permet de détecter ce problème de stratification.

Plusieurs extensions de la méthode multinomiale peuvent être envisagées. Tout d'abord, en considérant la situation inverse où l'on dispose d'un échantillon de cas et de témoins pour lesquels l'information sur l'environnement est disponible, comme dans une étude d'épidémiologie traditionnelle, et que l'on veuille prendre en compte l'interaction avec les facteurs génétiques dans l'analyse. Il est possible de limiter le coût de l'étude en ne génotypant que les cas et en réalisant le test MInT qui contrasterait l'association du facteur environnemental chez les cas porteurs d'un allèle (ou d'un génotype) de susceptibilité et chez les cas non porteurs d'un allèle (ou d'un génotype) de susceptibilité par rapport à l'échantillon de témoins. D'autre part, il serait également intéressant d'étendre cette approche à des phénotypes quantitatifs tout en conservant des facteurs discrets. Ainsi, dans le cas d'une variable environnementale dichotomique, l'approche consisterait à réaliser deux régressions linéaires et à tester simultanément les deux paramètres correspondants à l'effet du facteur génétique dans chacun des deux groupes exposés et non exposés.

Vers où vont se diriger les recherches sur les interactions gène-environnement dans les années à venir ? La réponse à cette question est étroitement liée à l'évolution des technologies moléculaires. En effet, avec la disponibilité récente de puces permettant de détecter les variants du nombre de copies (CNV), les études d'association génétique ont cherché avec parfois un certain succès à trouver des associations entre le nombre de copies et les maladies (Ionita-Laza et al. 2009; McCarroll 2008a, b; Shlien et Malkin 2009). L'idée que des réarrangements de l'ADN modifiant le nombre de copies d'un gène interagissent avec des facteurs environnementaux semble plausible et pourtant aucune des études sur les CNV réalisées jusqu'ici n'ont inclut de variable environnementale. La détection des interactions CNV-environnement pourrait également guider les recherches fonctionnelles vers de nouvelles voies biologiques.

Références bibliographiques

(1831) XVII. Epidemiologia Espanola, o Historia Cronologica de las pestes, Contagios, Epidemias y Epizootias que han Acaecida en Espana, desde la venida de los Cartagineses, hasta el ano 1801. Con noticia de algunas otras enfermidades de esta especie que han sufrido los Espanoles en otros Reynos, y de los Autores Nacionales que han escrito sobre esta materia, asi en la peninsula como fuera da ella. The American Journal of the Medical Sciences: 449-456

Abe R (2008) Toxic epidermal necrolysis and Stevens-Johnson syndrome: soluble Fas ligand involvement in the pathomechanisms of these diseases. J Dermatol Sci 52: 151-9

Anderson EC, Novembre J (2003) Finding haplotype block boundaries by using the minimum-description-length principle. Am J Hum Genet 73: 336-54

Andrieu N, Goldstein AM (1996) Use of relatives of cases as controls to identify risk factors when an interaction between environmental and genetic factors exists. Int J Epidemiol 25: 649-57

Andrieu N, Goldstein AM, Thomas DC, Langholz B (2001) Counter-matching in studies of gene-environment interaction: efficiency and feasibility. Am J Epidemiol 153: 265-74

Auquier-Dunant A, Mockenhaupt M, Naldi L, Correia O, Schroder W, Roujeau JC (2002) Correlations between clinical patterns and causes of erythema multiforme majus, Stevens-Johnson syndrome, and toxic epidermal necrolysis: results of an international prospective study. Arch Dermatol 138: 1019-24

Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7: 781-91

Bartsocas CS, Leslie RD (2002) Genetics of diabetes mellitus. Am J Med Genet 115: 1-3

Bellenguez C, Ober C, Bourgain C (2009) Linkage analysis with dense SNP maps in isolated populations. Hum Hered 68: 87-97

Blackwelder WC, Elston RC (1985) A comparison of sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 2: 85-97

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32: 314-31

Bouyer J, Hémon D, Cordier S, Derriennic F, Stücker I, Stengel B, Clavel J (1995) Chapitre 3. Mesure d'association entre la maladie et un facteur de risque. Épidémiologie. Principes et méthodes quantitatives. Les éditions INSERM, Paris, pp 57-88

Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, Elliott P (2009) Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. Int J Epidemiol 38: 263-73

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL (2002) A human genome diversity cell line panel. Science 296: 261-2

Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361: 598-604

Centers for Disease Control and Prevention, National diabetes surveillance system, US Department of Health and Human Services (2004) Diabetes Data for Colorado. http://apps.nccd.cdc.gov/ddtstrs/statePage.aspx?state=Colorado.

Chung WH, Hung SI, Hong HS, Hsih MS, Yang LC, Ho HC, Wu JY, Chen YT (2004) Medical genetics: a marker for Stevens-Johnson syndrome. Nature 428: 486

Clayton DG (2009) Prediction and interaction in complex disease genetics: experience in type 1 diabetes. PLoS Genet 5: e1000540

Clerget-Darpoux F, Bonaïti-Pellié C (1993) An exclusion map covering the whole genome: a new challenge for genetic epidemiologists? Am J Hum Genet 52: 442-3

Cordell HJ, Olson JM (2000) Correcting for ascertainment bias of relative-risk estimates obtained using affected-sib-pair linkage data. Genet Epidemiol 18: 307-21

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29: 229-32

deCODE Genetics Biobank Iceland. http://www.decode.com/

Dina C, Meyre D, Gallina S, Durand E, Korner A, Jacobson P, Carlsson LM, Kiess W, Vatin V, Lecoeur C, Delplanque J, Vaillant E, Pattou F, Ruiz J, Weill J, Levy-Marchal C, Horber F, Potoczna N, Hercberg S, Le Stunff C, Bougneres P, Kovacs P, Marre M, Balkau B, Cauchi S, Chevre JC, Froguel P (2007) Variation in FTO contributes to childhood obesity and severe adult obesity. Nat Genet 39: 724-6

Dizier MH, Selinger-Leneman H, Genin E (2003) Testing linkage and gene x environment interaction: comparison of different affected sib-pair methods. Genet Epidemiol 25: 73-9

Dobson AJ (2002) Nominal and Ordinal Logistic Regression. An introduction to generalized linear models, Second edn. Chapman & Hall/CRC, Boca Raton, pp 135-150

Doll R, Hill AB (1954) The mortality of doctors in relation to their smoking habits; a preliminary report. Br Med J 1: 1451-5

Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden DW, Smith DR, Lander ES, *et al.* (1987) A genetic linkage map of the human genome. Cell 51: 319-37

Elbein SC (1997) The genetics of human noninsulin-dependent (type 2) diabetes mellitus. J Nutr 127: 1891S-1896S

Elston RC, Yelverton KC (1975) General models for segregation analysis. Am J Hum Genet 27: 31-45

Estonian Genome Project. http://www.geenivaramu.ee/index.php?lang=eng

Feingold J, Martinez M (1998) Hérédité monogénique. In: Feingold J, Fellous M, Solignac M (eds) Principes de génétique humaine. Hermann, Paris, pp 69-104

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7: 85-97

Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, McCarthy MI (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316: 889-94

Freathy RM, Timpson NJ, Lawlor DA, Pouta A, Ben-Shlomo Y, Ruokonen A, Ebrahim S, Shields B, Zeggini E, Weedon MN, Lindgren CM, Lango H, Melzer D, Ferrucci L, Paolisso G, Neville MJ, Karpe F, Palmer CN, Morris AD, Elliott P, Jarvelin MR, Smith GD, McCarthy MI, Hattersley AT, Frayling TM (2008) Common variation in the FTO gene alters diabetes-related metabolic traits to the extent expected given its effect on BMI. Diabetes 57: 1419-26

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, Lee C (2006) Copy number variation: new insights in genome diversity. Genome Res 16: 949-61

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward

R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296: 2225-9

Gatto NM, Campbell UB, Rundle AG, Ahsan H (2004) Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias. Int J Epidemiol 33: 1014-24

Gauderman WJ (2002) Sample size requirements for matched case-control studies of gene-environment interaction. Stat Med 21: 35-50

Gauderman WJ, Siegmund KD (2001) Gene-environment interaction and affected sib pair linkage analysis. Hum Hered 52: 34-46

Generation Scotland. http://www.generationscotland.org/

Genetic Alliance Biobank (USA). http://www.biobank.org/

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M (2007) What is a gene, post-ENCODE? History and updated definition. Genome Res 17: 669-81

Goldstein AM, Andrieu N (1999) Detection of interaction involving identified genes: available study designs. J Natl Cancer Inst Monogr: 49-54

Goldstein AM, Dondon MG, Andrieu N (2006) Unconditional analyses can increase efficiency in assessing gene-environment interaction of the case-combined-control design. Int J Epidemiol 35: 1067-73

Guo SW (1998) Inflation of sibling recurrence-risk ratio, due to ascertainment bias and/or overreporting. Am J Hum Genet 63: 252-8

Guo SW (2002) Sibling recurrence risk ratio as a measure of genetic effect: caveat emptor! Am J Hum Genet 70: 818-9

Hardy GH (1908) Mendelian Proportions in a Mixed Population. Science 28: 49-50

Harley JB, Moser KL, Neas BR (1995) Logistic transmission modeling of simulated data. Genet Epidemiol 12: 607-12

Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, Fabianova E, Foretova L, Georges M, Janout V, Kabesch M, Krokan HE, Elvestad MB, Lissowska J, Mates D, Rudnai P, Skorpen F, Schreiber S, Soria JM, Syvanen AC, Meneton P, Hercberg S, Galan P, Szeszenia-Dabrowska N, Zaridze D, Genin E, Cardon LR, Lathrop M (2008) Investigation of the fine structure of European populations with applications to disease association studies. Eur J Hum Genet 16: 1413-29

Hetherington S, Hughes AR, Mosteller M, Shortino D, Baker KL, Spreen W, Lai E, Davies K, Handley A, Dow DJ, Fling ME, Stocum M, Bowman C, Thurmond LM, Roses AD

(2002) Genetic variations in HLA-B region and hypersensitivity reactions to abacavir. Lancet 359: 1121-2

Hill WG, Robertson A (1968) The effects of inbreeding at loci with heterozygote advantage. Genetics 60: 615-28

Hippocrate (1996) Airs, Eaux, Lieux. Rivages poche : Petite bibliothèque, Paris

Holle R, Happich M, Lowel H, Wichmann HE (2005) KORA--a research platform for population based health research. Gesundheitswesen 67 Suppl 1: S19-25

Hung SI, Chung WH, Liou LB, Chu CC, Lin M, Huang HP, Lin YL, Lan JL, Yang LC, Hong HS, Chen MJ, Lai PC, Wu MS, Chu CY, Wang KH, Chen CH, Fann CS, Wu JY, Chen YT (2005) HLA-B*5801 allele as a genetic marker for severe cutaneous adverse reactions caused by allopurinol. Proc Natl Acad Sci U S A 102: 4134-9

Hunter DJ (2005) Gene-environment interactions in human diseases. Nat Rev Genet 6: 287-98

Ionita-Laza I, Rogers AJ, Lange C, Raby BA, Lee C (2009) Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. Genomics 93: 22-6

Jensen LJ, Saric J, Bork P (2006) Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 7: 119-29

Kazma R, Bonaïti-Pellié C, Norris JM, Génin E (2010) On the use of sibling recurrence risks to select environmental factors liable to interact with genetic risk factors. Eur J Hum Genet 18: 88-94

Kazma R, Dizier MH, Guilloud-Bataille M, Bonaïti-Pellié C, Génin E (2007) Power comparison of different methods to detect genetic effects and gene-environment interactions. BMC Proc 1 Suppl 1: S74

Khoury MJ (1994) Case-parental control method in the search for disease-susceptibility genes. Am J Hum Genet 55: 414-5

Khoury MJ, Adams MJ, Jr., Flanders WD (1988a) An epidemiologic approach to ecogenetics. Am J Hum Genet 42: 89-95

Khoury MJ, Beaty TH, Cohen BH (1993a) Epidemiologic approaches to familial aggregation. Fundamentals of genetic epidemiology. Oxford University Press, New York, Oxford, pp 164-199

Khoury MJ, Beaty TH, Cohen BH (1993b) Genetic approaches to familial aggregation II. Segregation analysis. Fundamentals of genetic epidemiology. Oxford University Press, New York, Oxford, pp 233-283

Khoury MJ, Beaty TH, Cohen BH (1993c) Scope and Strategies of Genetic Epidemiology. Fundamentals of Genetic Epidemiology. Oxford University Press, New York, Oxford, pp 3-25

Khoury MJ, Beaty TH, Liang KY (1988b) Can familial aggregation of disease be explained by familial aggregation of environmental risk factors? Am J Epidemiol 127: 674-83

Khoury MJ, Flanders WD, Lipton RB, Dorman JS (1991) Commentary: the affected sib-pair method in the context of an epidemiologic study design. Genet Epidemiol 8: 277-82

Kleinbaum DG, Klein M (2002) Polytomous logistic regression. Logistic regression., Second edn. Springer-Verlag, New-York, pp 267-292

Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, Nathan DM (2002) Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med 346: 393-403

Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ (2007) Exploiting gene-environment interaction to detect genetic associations. Hum Hered 63: 111-9

Krawczak M, Nikolaus S, von Eberstein H, Croucher PJ, El Mokhtari NE, Schreiber S (2006) PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. Community Genet 9: 55-61

Kriska AM, Knowler WC, LaPorte RE, Drash AL, Wing RR, Blair SN, Bennett PH, Kuller LH (1990) Development of questionnaire to examine relationship of physical activity and diabetes in Pima Indians. Diabetes Care 13: 401-11

Kriska AM, Saremi A, Hanson RL, Bennett PH, Kobes S, Williams DE, Knowler WC (2003) Physical activity, obesity, and the incidence of type 2 diabetes in a high-risk population. Am J Epidemiol 158: 669-75

Kupper LL, Hogan MD (1978) Interaction in epidemiologic studies. Am J Epidemiol 108: 447-53

Lazarou J, Pomeranz BH, Corey PN (1998) Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. Jama 279: 1200-5

Leal SM, Ott J (2000) Effects of stratification in the analysis of affected-sib-pair data: benefits and costs. Am J Hum Genet 66: 567-75

Lewontin RC (1964) The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics 49: 49-67

Lewontin RC (1988) On measures of gametic disequilibrium. Genetics 120: 849-52

Li C, Sacks L (1954) The derivation of the joint distribution and correlation between relatives by the use of stochastic matrices. Biometrics 10: 347-360

Lindström S, Yen YC, Spiegelman D, Kraft P (2009) The impact of gene-environment dependence and misclassification in genetic association studies incorporating gene-environment interactions. Hum Hered 68: 171-81

Lonjou C, Borot N, Sekula P, Ledger N, Thomas L, Halevy S, Naldi L, Bouwes-Bavinck JN, Sidoroff A, de Toma C, Schumacher M, Roujeau JC, Hovnanian A, Mockenhaupt M (2008) A European study of HLA-B in Stevens-Johnson syndrome and toxic epidermal necrolysis related to five high-risk drugs. Pharmacogenet Genomics 18: 99-107

Lonjou C, Thomas L, Borot N, Ledger N, de Toma C, LeLouet H, Graf E, Schumacher M, Hovnanian A, Mockenhaupt M, Roujeau JC (2006) A marker for Stevens-Johnson syndrome …: ethnicity matters. Pharmacogenomics J 6: 265-8

Lynch J, Helmrich SP, Lakka TA, Kaplan GA, Cohen RD, Salonen R, Salonen JT (1996) Moderately intense physical activities and high levels of cardiorespiratory fitness reduce the risk of non-insulin-dependent diabetes mellitus in middle-aged men. Arch Intern Med 156: 1307-14

Maestri NE, Beaty TH, Hetmanski J, Smith EA, McIntosh I, Wyszynski DF, Liang KY, Duffy DL, VanderKolk C (1997) Application of transmission disequilibrium tests to nonsyndromic oral clefts: including candidate genes and environmental exposures in the models. Am J Med Genet 73: 337-44

Mallal S, Nolan D, Witt C, Masel G, Martin AM, Moore C, Sayer D, Castley A, Mamotte C, Maxwell D, James I, Christiansen FT (2002) Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. Lancet 359: 727-32

Manolio TA, Brooks LD, Collins FS (2008) A HapMap harvest of insights into the genetics of common disease. J Clin Invest 118: 1590-605

Manolio TA, Rodriguez LL, Brooks L, Abecasis G, Ballinger D, Daly M, Donnelly P, Faraone SV, Frazer K, Gabriel S, Gejman P, Guttmacher A, Harris EL, Insel T, Kelsoe JR, Lander E, McCowin N, Mailman MD, Nabel E, Ostell J, Pugh E, Sherry S, Sullivan PF, Thompson JF, Warram J, Wholley D, Milos PM, Collins FS (2007) New models of collaboration in genome-wide association studies: the Genetic Association Information Network. Nat Genet 39: 1045-51

McCarroll SA (2008a) Copy-number analysis goes more than skin deep. Nat Genet 40: 5-6

McCarroll SA (2008b) Extending genome-wide association studies to copy-number variation. Hum Mol Genet 17: R135-42

Melsom RD (1999) Familial hypersensitivity to allopurinol with subsequent desensitization. Rheumatology (Oxford) 38: 1301

Miller JW (2008) Of race, ethnicity, and rash: the genetics of antiepileptic drug-induced skin reactions. Epilepsy Curr 8: 120-1

Mockenhaupt M, Viboud C, Dunant A, Naldi L, Halevy S, Bouwes Bavinck JN, Sidoroff A, Schneck J, Roujeau JC, Flahault A (2008) Stevens-Johnson syndrome and toxic epidermal necrolysis: assessment of medication risks with emphasis on recently marketed drugs. The EuroSCAR-study. J Invest Dermatol 128: 35-44

Morris AP, Lindgren CM, Zeggini E, Timpson NJ, Frayling TM, Hattersley AT, McCarthy MI (2009) A powerful approach to sub-phenotype analysis in population-based genetic association studies. Genet Epidemiol

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7: 277-318

Murcray CE, Lewinger JP, Gauderman WJ (2009) Gene-environment interaction in genome-wide association studies. Am J Epidemiol 169: 219-26

Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G, Vollenweider P, Oksenberg JR, Hauser SL, Stirnadel HA, Kooner JS, Chambers JC, Jones B, Mooser V, Bustamante CD, Roses AD, Burns DK, Ehm MG, Lai EH (2008) The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 83: 347-58

Nelson TL, Fingerlin TE, Moss LK, Barmada MM, Ferrell RE, Norris JM (2007) Association of the peroxisome proliferator-activated receptor gamma gene with type 2 diabetes mellitus varies by physical activity among non-Hispanic whites from Colorado. Metabolism 56: 388-93

Ottman R (1995) Gene-environment interaction and public health. Am J Hum Genet 56: 821-3

Ottman R (1996) Gene-environment interaction: definitions and study designs. Prev Med 25: 764-70

Pan XR, Li GW, Hu YH, Wang JX, Yang WY, An ZX, Hu ZX, Lin J, Xiao JZ, Cao HB, Liu PA, Jiang XG, Jiang YY, Wang JP, Zheng H, Zhang H, Bennett PH, Howard BV (1997) Effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance. The Da Qing IGT and Diabetes Study. Diabetes Care 20: 537-44

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2: e190

Pellicano R, Silvestris A, Iannantuono M, Ciavarella G, Lomuto M (1992) Familial occurrence of fixed drug eruptions. Acta Derm Venereol 72: 292-3

Piegorsch WW, Weinberg CR, Taylor JA (1994) Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 13: 153-62

Pirmohamed M (2006) Genetic factors in the predisposition to drug-induced hypersensitivity reactions. Aaps J 8: E20-6

Pirmohamed M, Arbuckle JB, Bowman CE, Brunner M, Burns DK, Delrieu O, Dix LP, Twomey JA, Stern RS (2007) Investigation into the multidimensional genetic basis of drug-induced Stevens-Johnson syndrome and toxic epidermal necrolysis. Pharmacogenomics 8: 1661-91

Pirmohamed M, Park BK (2001) Genetic susceptibility to adverse drug reactions. Trends Pharmacol Sci 22: 298-305

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-9

Purcell S (2002) Variance components models for gene-environment interaction in twin analysis. Twin Res 5: 554-71

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-75

R Development Core Team (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Raychaudhuri S, Plenge RM, Rossin EJ, Ng AC, Purcell SM, Sklar P, Scolnick EM, Xavier RJ, Altshuler D, Daly MJ (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet 5: e1000534

Rebbeck TR, Spitz M, Wu X (2004) Assessing the function of genetic variants in candidate gene association studies. Nat Rev Genet 5: 589-97

Risch N (1990a) Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46: 222-8

Risch N (1990b) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 46: 242-53

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 13: 1516-7

Rothman KJ, Greenland S, Walker AM (1980) Concepts of interaction. Am J Epidemiol 112: 467-70

Roujeau JC, Bracq C, Huyn NT, Chaussalet E, Raffin C, Duedari N (1986) HLA phenotypes and bullous cutaneous reactions to drugs. Tissue Antigens 28: 251-4

Roujeau JC, Huynh TN, Bracq C, Guillaume JC, Revuz J, Touraine R (1987) Genetic susceptibility to toxic epidermal necrolysis. Arch Dermatol 123: 1171-3

Roujeau JC, Stern RS (1994) Severe adverse cutaneous reactions to drugs. N Engl J Med 331: 1272-85

Rzany B, Mockenhaupt M, Baur S, Schroder W, Stocker U, Mueller J, Hollander N, Bruppacher R, Schopf E (1996) Epidemiology of erythema exsudativum multiforme majus, Stevens-Johnson syndrome, and toxic epidermal necrolysis in Germany (1990-1992): structure and results of a population-based registry. J Clin Epidemiol 49: 769-73

Schaid DJ (1999) Case-parents design for gene-environment interaction. Genet Epidemiol 16: 261-73

Schmidt S, Schaid DJ (1999) Potential misinterpretation of the case-only study to assess gene-environment interaction. Am J Epidemiol 150: 878-85

Selinger-Leneman H, Genin E, Norris JM, Khlat M (2003) Does accounting for gene-environment (GxE) interaction increase the power to detect the effect of a gene in a multifactorial disease? Genet Epidemiol 24: 200-7

Serre JL (1997) Chapitre 2. Le modèle général de Hardy-Weinberg. Génétique des populations. Modèles de base et applications. Nathan, Paris, pp 43-70

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308-11

Shlien A, Malkin D (2009) Copy number variations and cancer. Genome Med 1: 62

Snow SJ (2008) John Snow: the making of a hero? Lancet 372: 22-3

Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59: 983-9

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52: 506-16

StataCorp. (1984-2008) STATA/SE 10.1. 10.1 edn. Statacorp LP., College Station, Texas, USA

Stücker I, Bonaïti-Pellié C, Hémon D (1993) Epidemiology of lung cancer: interaction between genetic susceptibility and environmental risk factors. In: Hirsch A, Goldberg M, Martin J-P, Masse R (eds) Prevention of respiratory diseases, vol 68: Lung biology in health and disease. Marcel Dekker, New York, Basel, Hong Kong, pp 149-165

Svensson CK, Cowen EW, Gaspari AA (2001) Cutaneous drug reactions. Pharmacol Rev 53: 357-79

Sweedish National Biobank Program. http://www.biobanks.se/

The International HapMap Consortium (2003) The International HapMap Project. Nature 426: 789-96

The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299-320

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-61

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-78

Thomas D (2010) Gene-environment-wide association studies: emerging approaches. Nat Rev Genet 11: 259-272

Thomas DC (2000) Case-parents design for gene-environment interaction by Schaid. Genet Epidemiol 19: 461-3

Thompson WD (1991) Effect modification and the limits of biological inference from epidemiologic data. J Clin Epidemiol 44: 221-32

Tuomilehto J, Lindstrom J, Eriksson JG, Valle TT, Hamalainen H, Ilanne-Parikka P, Keinanen-Kiukaanniemi S, Laakso M, Louheranta A, Rastas M, Salminen V, Uusitupa M (2001) Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. N Engl J Med 344: 1343-50

UK Biobank. www.ukbiobank.ac.uk

Umbach DM, Weinberg CR (1997) Designing and analysing case-control studies to exploit independence of genotype and exposure. Stat Med 16: 1731-43

Umbach DM, Weinberg CR (2000) The use of case-parent triads to study joint effects of genotype and exposure. Am J Hum Genet 66: 251-61

Wahrendorf J (1981) Approaches to the detection of interaction effects. In: Bithell J, Coppi R (eds) Perspectives in medical statistics. Academic Press, London, pp 1-20

Walter SD, Holford TR (1978) Additive, multiplicative, and other models for disease risks. Am J Epidemiol 108: 341-6

Waterloo Maple Inc. (2005) Maple Software, version 10. 10 edn, Ontario, Canada

Weijnen CF, Rich SS, Meigs JB, Krolewski AS, Warram JH (2002) Risk of diabetes in siblings of index cases with Type 2 diabetes: implications for genetic studies. Diabet Med 19: 41-50

Wichmann HE, Gieger C, Illig T (2005) KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen 67 Suppl 1: S26-30

Witte JS, Gauderman WJ, Thomas DC (1999) Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. Am J Epidemiol 149: 693-705

Wolverton SE (1997) Update on cutaneous drug reactions. Adv Dermatol 13: 65-84

Yang Q, Khoury MJ (1997) Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. Epidemiol Rev 19: 33-43

Zaykin D, Kozbur D (2009) Shared control design in genome-wide association studies. (Poster P151) 18th annual meeting of the International Genetic Epidemiology Society, Kahuku, Hawaii, USA. October 18-20, 2009.

Annexes

Annexe 1

Kazma R, Bonaiti-Pellié C, Norris J, Génin E.
On the use of sibling recurrence risks to select environmental factors liable to interact with genetic risk factors.
Eur J Hum Genet 2010, 18(1):88-94.

On the use of sibling recurrence risks to select environmental factors liable to interact with genetic risk factors

Rémi Kazma[*,1,2], Catherine Bonaïti-Pellié[3,1], Jill M. Norris[4], Emmanuelle Génin[2,5]

[1] Univ. Paris-Sud, Faculté de Médecine, Le Kremlin Bicêtre, France

[2] Inserm, UMR-S946, Fondation Jean Dausset – CEPH, Paris, France

[3] Inserm, UMR-S535, Villejuif, France

[4] Department of Preventive Medicine and Biometrics, University of Colorado Denver, Denver, Colorado, USA

5 Univ. Paris-Diderot, Paris, France

* Correspondence: R Kazma, Inserm UMR-S946, Fondation Jean Dausset – CEPH, 27 rue Juliette Dodu, Paris 75010, France. Tel: +33153725027; Fax: +33153725049; E-mail: remi.kazma@inserm.fr

Running title: GxE interaction and sibling recurrence risk

Abstract

Gene-environment interactions are likely to be involved in the susceptibility to multifactorial diseases but are difficult to detect. Available methods usually concentrate on some particular genetic and environmental factors. In this paper, we propose a new method to determine whether or not a given exposure is susceptible to interact with unknown genetic factors. Rather than focusing on a specific genetic factor, the degree of familial aggregation is used as a surrogate for genetic factors. A test comparing the recurrence risks in sibs according to the exposure of indexes is proposed and its power is studied for varying values of model parameters. The Exposed versus Unexposed Recurrence Analysis (*EURECA*) is valuable for common diseases with moderate familial aggregation, only when the role of exposure has been clearly outlined. Interestingly, accounting for a sibling correlation for the exposure increases the power of *EURECA*. An application on a sample ascertained through one index affected with type 2 diabetes is presented where gene-environment interactions involving obesity and physical inactivity are investigated. Association of obesity with type 2 diabetes is clearly evidenced and a potential interaction involving this factor is suggested in Hispanics (p=0.045), whereas a clear gene-environment interaction is evidenced involving physical inactivity only in Non-Hispanic Whites (p=0.028). The proposed method might be of particular interest prior to genetic studies to help determine the environmental risk factors that will need to be accounted for to increase the power to detect genetic risk factors and to select the most appropriate samples to genotype.

Keywords: diabetes mellitus, type 2; epidemiologic research design; familial aggregation; genetic predisposition to disease; environmental exposure.

Introduction

If gene-environment (GxE) interactions are expected to play an important role in multifactorial disease susceptibility[1] genetic and environmental factors are most often evaluated independently rather than jointly. Joint analysis and GxE interaction testing is usually performed in a second step once the observed effects of each factor has been evidenced[2-4]. Using such a strategy, we are likely to miss important genetic or environmental factors which effects could only be detected when accounting for the other factor[5,6]. This was clearly evidenced in the study by Selinger-Leneman *et al.*[6] where it was shown that the power to detect a genetic risk factor interacting with an environmental risk factor might be considerably reduced when the environmental exposure of individuals is not accounted for. However, this was very dependent on the environmental risk factor prevalence, on its effect on the disease and on its interaction with the genetic factor. In some situations, accounting for the environmental exposure was even detrimental in terms of power. This first study called for the need to develop methods to select environmental factors that might be involved in GxE interaction and should therefore be accounted for in genetic studies.

The problem of selecting environmental exposures to account for in genetic studies becomes even more crucial when performing genome-wide association studies with hundreds of thousands of markers. Indeed, in this context, for each exposure to study, there is such a huge number of tests to perform that one wants to make sure that only relevant exposures are accounted for. The development of methods to select these relevant environmental factors will probably be the first step in order to test for GxE interactions at the genome-wide levels.

In their previous work, Selinger-Leneman *et al.*[6] have shown that selecting environmental factors based solely on their observed effects is not an efficient strategy and it might be useful to find a statistical tool to determine if they are likely to interact with genetic risk factors. This, however, should be done prior to the genetic analysis and thus involves the use of methods that do not require genotyping data. One such method was proposed by

Purcell[7] for twin data and relies on variance component modeling. Apart from the fact that it requires twin data, the method also requires exposure status of both sibs which is not always easy to obtain. Our proposed method also uses familial aggregation of the disease as a surrogate for the genetic factors but exposure in indexes only. Indeed, as suggested by Stücker *et al.*[8], familial aggregation of disease would be different for exposed and unexposed indexes if the environmental factor studied is involved in GxE interaction. A rational for this property is that, in presence of GxE interaction, exposed indexes have not the same distribution of genotypes as unexposed indexes. Their sibs will consequently have a different probability of having the disease from those of unexposed indexes.

In this paper, we used this idea of difference in sibling recurrence risks based on index's exposure to propose a test aimed at selecting environmental factors that are prone to interact with the genetic component of a multifactorial disease and propose a simple statistical test. We study the statistical properties of this test under different models and apply it on a type 2 diabetes (T2D) sample.

Materials and Methods

To evidence a difference in the recurrence risk for siblings of exposed and unexposed individuals, we need data on a sample of sib pairs ascertained through an affected index (sib 1). The variable of interest is the affection status of the other sib (sib 2) and the explicative variable is the exposure status of sib 1. The data can be presented in a contingency table such as table 1.

Odds Ratio of Recurrence and Exposed versus Unexposed Recurrence Analysis

Let K_S be the sibling recurrence risk defined as the probability of sib 2 being affected given sib 1 is affected[9] and K_{SE} and $K_{S\bar{E}}$ these risks when sib 1 is exposed and unexposed respectively to a given environmental factor *E*. To measure the difference between these two stratified risks, an Odds Ratio of Recurrence (*ORR*) can be calculated by analogy with an Odds Ratio (*OR*):

$$ORR = \frac{K_{SE} \times \left(1 - K_{S\bar{E}}\right)}{K_{S\bar{E}} \times \left(1 - K_{SE}\right)} \qquad (1)$$

Deriving the above recurrence risks as a function of observed numbers in the contingency table (table 1), the *ORR* can be expressed as:

$$ORR = \frac{ad}{bc}$$

In contrast to the *OR* of an environmental factor where exposure and disease statuses are measured in the same individual, in the *ORR*, exposure is measured in the affected index and the disease status is measured in the sib.

In the presence of a GxE interaction involving environmental factor E, we expect the *ORR* to be different from 1. To test for "*ORR* = 1", we propose to perform a 1 degree of freedom (df) chi-square test on the contingency table crossing sib 1's exposure with sib 2's affection status (table 1) or the asymptotically similar Wald test based on the logistic

regression parameter estimate and its variance. This test will be referred to as the Exposed versus Unexposed Recurrence Analysis (*EURECA*) test.

Properties of the *ORR* and of the *EURECA* test under different models

In order to study the behavior of the *ORR* and the statistical properties of *EURECA*, we considered a model of interaction involving a single gene (G) and a single environmental factor (E) even though the method practically only uses environmental information. We computed the expected numbers in each cell of the contingency table and derived the different recurrence risks (table 1) under the different models of gene-environment interaction defined by the parameters presented in table 2. A disease *D* with population prevalence f_D is considered. It is assumed that *D* is causally associated only with an environmental factor *E* and a genetic factor *G*.

The *E* factor is dichotomous with population frequency f_E and a main effect size on *D* measured by the exposure relative risk, RR_E. To model the possibility for a familial clustering of *E*, as in Khoury *et al.*[10], we define the conditional probability of sib 2 being exposed given the exposure status of sib 1 as:

$$P(\text{sib 2 } E+ \mid Y_1) = (1 - C_E) \times f_E + C_E \times Y_1 \qquad (2)$$

where Y_1 is a dummy variable that takes the value 1 when sib 1 is exposed and 0 otherwise, and C_E is the environmental correlation between the sibs. Thus, when $C_E = 0$, sib 2's exposure status is independent from sib 1's exposure status and its probability is always equal to the prevalence of *E* in the general population, f_E. When $C_E = 1$, correlation between sibs for exposure is complete and sib 2's exposure probability is equal to 1 when sib 1 is exposed and 0 when sib 1 is unexposed.

The *G* factor corresponds to a predisposing genetic factor localized on an autosomal biallelic genetic locus. The allele that confers predisposition to disease is noted *A* and has a

population frequency of q, whereas the other allele a has a population frequency of $1–q$. Frequencies of the different possible genotypes (AA, Aa and aa) are supposed to follow Hardy-Weinberg proportions in the population (*i.e.*, q^2, $2q(1–q)$, $(1–q)^2$, respectively). The main effect of the G factor is measured by the genotypic relative risk (RR_G) which corresponds to the ratio of the disease risk in carriers of the predisposing genotype(s) to the risk in non-carriers of the predisposing genotype(s) among unexposed individuals. In all situations, we compared dominant and recessive genetic models for a given frequency f_G of predisposing genotype(s), with $f_G = q^2$ under a recessive model and $f_G = q^2+2q(1–q)$ under a dominant model.

Let B designate the baseline risk *i.e.* the probability of disease for a non-carrier and unexposed individual. The interaction between E and G is measured by an interaction coefficient I, which corresponds to a departure from a multiplicative model when both E and G are present. In the absence of interaction, the risk of an individual exposed and carrier of the predisposing genotype is the product of B, RR_E and RR_G. In the presence of interaction, this risk is multiplied by the interaction coefficient I (table 2). The conditional risks of disease given genotype and exposure status and the numbers of the contingency table cells were derived using the ITO matrix method of Li and Sacks[11] modified in order to account for the environmental factor. Computations were done with the Maple 10 software[12] and explanations are given in the Supplementary materials.

Type I error and power of the *EURECA* test were asymptotically estimated considering a sample of 1000 sib-pairs by use of 1 df non-central chi-square distributions. Alternatively, we calculated the required number of sib-pairs to reach a power of 0.80 with a type I error rate of 0.05.

Application to type 2 diabetes

The Gene ENvironment Interactions (GENI) study[13] collected phenotypic and environmental data of type 2 diabetic subjects and their families living in the San Luis Valley and the Denver metropolitan area in Colorado (USA). Among 452 pedigrees (3090 nuclear families) ascertained through one index sib affected with T2D, we extracted 2699 index-sib pairs for which data was available in the index for at least one of the two studied exposures: obesity and physical inactivity. Of those pairs, 1734 were Hispanics (H) and 965 were Non-Hispanic Whites (NHW). Subjects previously diagnosed by a physician as having T2D and treated with oral hypoglycemic agents or insulin were considered affected. For subjects that did not report having T2D or subjects untreated for T2D, diabetic status was determined by an oral glucose tolerance test using American Diabetes Association criteria (1997). For diabetic subjects, self reported body mass index (BMI) at the time of diagnosis was used. BMI was calculated at recruitment time for other subjects. Individuals having a BMI value exceeding 30 kg/m^2 were classified as obese. Physical activity assessment was done once during the study using a previously validated questionnaire self-administered by the subjects[14]. Energy expenditure was assessed as metabolic equivalent task (MET) units. The MET is the ratio of the metabolic rate during exercise to the metabolic rate at rest[15]. The average METs per week (prior to the diagnosis of T2D for affected individuals) was calculated for each study participant. The METs variable was divided into sex-specific tertiles, and a dichotomous variable was created distinguishing individuals in the lower tertile ("low physical activity") from those in the upper two tertiles.

We carried all the analysis separately for the two population strata (H and NHW) because the two exposures distributions were significantly heterogeneous. We first evaluated the observed main effect of each exposure using conditional logistic regression applied on discordant sib pairs for the T2D affection status. The numbers of available subjects were 198

H and 116 NHW for obesity and 458 H and 309 NHW for physical inactivity. Exposure frequency was measured in the control samples (unaffected sibs) and used as an estimate of exposure prevalence in population.

For each exposure, we randomly selected one sib for each index in order to compute contingency table numbers and global and stratified recurrence risks (K_S, K_{SE} and $K_{S\bar{E}}$). The numbers of available pairs were 267 H and 321 NHW for obesity and 246 H and 268 NHW for physical inactivity. We derived an *ORR* for each exposure and applied the *EURECA* test of interaction using a logistic regression model. In order to account for correlated pairs belonging to the same pedigree, we computed the standard error of the logistic regression parameter using a robust sandwich estimator clustered by family as implemented in Stata/SE 10.1[16]. When exposure of the random sib was available, the pairs were also used to calculate a correlation coefficient between sib pairs for each exposure variable using equation 2.

Results

Behavior of the Odds Ratio of Recurrence under different disease models

To evaluate the pertinence of using the *ORR* as an indicator of the presence of a GxE interaction, we investigated the variations of the *ORR* under different models first without correlation between siblings for *E* ($C_E = 0$). As expected, we observe that, in presence of an interaction, the values of the *ORR* increase with increasing values of the interaction coefficient *I*, but they also depend on the other model parameters. Impacts of these parameters are shown in figure 1 for the exposure parameters (f_E and RR_E) and in figure 2 for the genetic parameters (f_G and RR_G). For a given value of *I*, *ORR* is greater for high values of f_E and RR_E (figure 1) and small values of f_G. When prevalence of the predisposing genotype(s) increases ($f_G = 0.2$), the changes in *ORR* seen with varying RR_G tend to disappear and even reverse when interaction values are elevated (figure 2). *ORR* is higher for a dominant as compared to a recessive model at fixed f_G.

Since environmental correlation between sibs might induce a possible confusion with a GxE interaction, we looked into variations of *ORR* values for different values of C_E, when $I = 1$ and $I = 5$ (figure 3). We observe that under the null hypothesis ($I = 1$), the *ORR* value (referred to as ORR_0) is always equal to 1 in situations where there is no correlation of the *E* factor ($C_E = 0$) or when there is no effect of *E* ($RR_E = 1$). On the other hand, in the presence of an effect of *E* (*i.e.*, $RR_E \neq 1$) associated with a correlation between sibs for this factor (*i.e.*, $C_E \neq 0$), the ORR_0 values are inflated. In presence of a sibling correlation for *E*, the estimates obtained with a GxE interaction ($I = 5$, in figure 3) should thus be tested against the value of ORR_0 rather than against 1. The null hypothesis of the test becomes "$ORR = ORR_0$". The value of ORR_0 depends on the disease prevalence, on the environmental parameters and to a lesser extent on the genetic parameters. In order to estimate ORR_0, we thus need to obtain

some estimates these different parameters. Disease prevalence is often known from previous studies in similar populations. The environmental parameters (C_E, f_E, RR_E) can be estimated using the studied sample when data on the environmental exposure of siblings is available (see the type-2 diabetes example here). If this is not the case, results from previous studies on the effect of the environmental could be used. Only the genetic model is not known. We propose to calculate ORR_0 for different genetic model parameters (f_G, RR_G) and then to use as ORR_0 the value the closest to the observed ORR. This "worst case scenario" ensures a robust inference on the test (see example in the section Results, Application to type 2 diabetes). In order to compute the expected ORR_0, the Maple source code of "*EURECA*" is available from the corresponding author upon request. More theoretical derivation of the ORR_0 computation is also given in the Supplementary materials.

Properties of the Exposed versus Unexposed Recurrence Analysis test

In figure 4, the power of the *EURECA* test of "$ORR = ORR_0$" is reported for varying levels of interaction and C_E under dominant and recessive models. As expected, the power increases with increasing value of I but more interestingly, this increases depends on C_E and is larger for high C_E values than for low C_E values. Alternatively, table 3 reports the number of sib pairs that are needed to reach a power of 0.80 with a type I error rate of 0.05 for increasing values of I and C_E under a plausible disease model (frequent factors, $f_G = 0.1$ and $f_E = 0.2$; with moderate effects, $RR_G = 2$ and $RR_E = 2$). In situations with no C_E and small I, sample sizes are very high and thus unlikely to be recruited. But considering situations with elevated interaction coefficients ($I > 3$) and with high correlation for exposure in sibs ($C_E > 0$), sample sizes are more reasonable.

Considering the same frequencies with a sibling correlation of 0.25 and varying values of RR_E and RR_G, the required sample sizes are shown in figure 5. As expected, these sizes are smaller when *G* and *E* have strong effects, but they seem to be more sensitive to *G* than to *E*.

All the previous results considered a disease prevalence (f_D) of 0.10. Variations in power as a function of interaction and disease prevalence are presented in figure 6. In summary, it shows that the best performances of this test are obtained with common rather than with rare diseases. When the disease is rare, the sibling recurrence risk (K_S) is low and the difference between the exposed and unexposed index strata due to the GxE interaction is harder to detect.

Application to type 2 diabetes

The results of the T2D application are presented in table 4. For each population strata (H and NHW) and each studied exposure, we show first the environmental parameter estimates: OR_E (exposure's odds ratio), f_E and C_E, and then the proposed GxE interaction analysis: *ORR* and *EURECA* test. In order to account for C_E, we calculated the *ORR* expected under the null hypothesis (ORR_0). The Center for Disease Control and Prevention (CDC) 2001 diabetes data for the state of Colorado provided diabetes prevalence (f_D = 4.5 %)[17]. Based on this estimate and using the environmental parameters calculated previously on the T2D data, we computed expected ORR_0 values for a wide range of genetic parameters (f_G = 0.01-0.5, RR_G =0.5-10). An interval of variation of ORR_0 was obtained in this way. To ensure robustness of the test, we considered the "worst-case scenario" and compared the observed *ORR* to the value of ORR_0 that was the closest to the observed *ORR*.

In H, obesity has an *ORR* equal to 0.67 (95 % CI: 0.40, 1.11). Remarkably in this stratum, obesity has a strong significant observed effect of 2.48 (95 % CI: 1.18, 5.22), which, associated with a C_E of 0.22 and a f_E of 0.29, gives an expected ORR_0 varying between 1.25

and 1.27. In this example, we used 1.25 (closest value to the observed *ORR* of 0.67) to perform the *EURECA* test and obtained a *p*-value of 0.045. In NHW, obesity has also a significant observed effect with an *OR* of 3.87 (95 % CI: 1.54, 9.65) but the interaction test is not significant.

Considering physical inactivity, the interaction test is significant in the NHW sample only ($p = 0.028$) and the *ORR* is 2.13 (95 % CI: 1.08, 4.19). This exposure has no significant observed effect and does not aggregate in sib-pairs, which is a situation where the proposed test usually lacks power to detect the GxE interaction (as shown in table 3 and figure 5).

Since the sex distributions of indexes and of sibs were homogenous between the groups of exposed and unexposed indexes, this variable should not interfere with the *EURECA* test.

Discussion

Contrasting the sibling recurrence risks based on the exposure status of the index is a simple and attractive approach to select environmental factors involved in a GxE interaction. We propose to measure this contrast by computing an Odds Ratio of Recurrence (*ORR*) and show that the *ORR* is a good indicator of a GxE interaction. This *ORR* is not a direct measure of interaction but rather a measure of the difference between recurrence risks in exposed and unexposed indexes. For example, using the low physical activity in NHW result in table 4, the risk of T2D in a NHW individual is multiplied on average by a factor of 2.13 when his affected index sib has a low physical activity compared to an individual whose affected index sib has a high physical activity. At this level of information, discriminating between an underlying genetic component interacting with the exposure and the familial clustering of this exposure associated to the disease is quite difficult[18], but our results show that it is possible, provided that the effect and familial correlation of the environmental factor is well documented.

In the context of a dichotomous environmental variable, the interests of using the *ORR*, instead of the ratio of recurrence risks, resides in applying a logistic regression as done in most epidemiologic studies, but the same approach can be easily extended to multiclass or continuous environmental factors using the classic general linear models. The use of continuous variables when available would probably increase the power but would also make the assumption of a linear relation. To test for the difference of the *ORR* with a null hypothesis value (ORR_0), we derive a statistical test, the Exposed versus Unexposed Recurrence Analysis (*EURECA*) test. To use this test, we need to define the value of ORR_0. We have derived analytically a formula to compute ORR_0 based on exposure parameter and disease prevalence estimates. These estimates are often easily obtained from the data sample and from the literature. To ensure robustness of the test, we suggest accounting for the impact

on ORR_0 of possible variations in these estimates by deriving a range of variation of ORR_0 and to consider in the test the ORR_0 value the closest to the observed *ORR*. Note that the loss of power due to the uncertainty of the genetic parameters should be minimal since the ORR_0 variations would usually be small as in the illustrative example (from 0.01 to 0.03). Interestingly in our example, we found even under this "worst-case scenario", it is possible to show that observed *ORR* for some exposure significantly differs from ORR_0. This is in good agreement with the results of Khoury *et al.*[10] showing that the degree of familial aggregation of most common diseases cannot be entirely explained by a familial clustering of environmental risk factors even if we assume an extreme clustering of the environmental factor.

The study of the statistical properties of *EURECA* has shown that the test is appropriate to test for common diseases rather than rare ones (figure 6). Interestingly, even when the tests are corrected for the exposure correlation in siblings, powers were found to be higher for elevated values than for lower values of C_E. We hypothesize that the sibling correlation actually has a confounding effect on one part, but also emphasizes the existing difference in recurrence risks between strata of indexes due to the GxE interaction. We only tested positive correlation coefficients between siblings for exposure since it is probably the most common situation in familial studies.

GxE interactions are difficult to detect and often require very large sample sizes. In an effort to increase the power to detect GxE interaction, new methods have been developed that are based on particular sampling designs. Among these methods are the log-linear modeling method that uses case-parent trio data and compares genotype distribution of exposed and unexposed cases conditional on parental genotypes[19], methods that use counter-matching designs to enrich the sample with rare exposure or genetic factors[20] or case-control-combined designs with both population and familial controls[21]. A common feature of all these different

methods to detect GxE interactions is their need to have a complete knowledge of the exposure statuses and genotypes of the studied subjects. Among the methods that use familial aggregation of the disease as a surrogate for the latent genetic factor, Purcell[7] proposed to apply variance components models in twin studies to evidence GxE interactions with an environmental factor measured in both twins. What distinguishes the method we propose here is the type of information used in order to assess the GxE interaction. This method relies on the exposure of the index case and information on the familial recurrence of the disease. There is no need to have a measure of exposure in the sibs and for easily recognizable diseases, their affection status might be obtained from indexes. Large sample sizes can thus be obtained at a minimal cost. It is true however that if sibs could also be examined, familial recurrence will certainly be better estimated. It will also be possible to assess, directly from the data rather than from the literature, potential environmental correlation between sibs.

The use of the sib recurrence information as surrogate for genetic risk factors has the advantage of requiring no *a priori* hypothesis on the genetic model underlying disease susceptibility. It also permits to test for the involvement in the disease of genetic factors located anywhere on the genome at no cost in terms of multiple testing. This is an important point as the issue of multiple testing in GxE interaction studies considering thousands of genetic markers coupled with tens of exposures remains to be resolved. On the counterpart, this approach only stipulates a specific environmental factor and tests for its interaction with the genetic component implicated in disease risk increase. As compared to other methods that use both genotypic and environmental information, this method could lack power to detect some interaction with a specified genetic factor. But it provides an easy way to screen for environmental factors potentially implicated in GxE interactions when genotypes are not available.

Association between T2D and obesity was significant both in H and NHW, as previously evidenced in many cross-sectional and longitudinal studies[22]. Concerning interaction, *EURECA* was significant only in H ($p = 0.045$) with a particular model of interaction where the interaction effect is in opposite direction compared to the main exposure effect. In an earlier study of recurrence risk estimation in T2D families, analogous results were found and elevated recurrence risk ratios were found in siblings of non-obese as compared to obese patients[23]. This kind of interaction illustrates the situations where the GxE interaction is a nuisance element that has to be accounted for in order to better detect a main effect[5,6]. Regarding, low physical activity which had no significant observed effect in any of the two populations, the interaction test was significant in NHW ($p = 0.028$) but not in H. A previous study that applied family-based association tests and generalized estimating equations models showed a GxE interaction between the peroxisome proliferator-activated receptor γ gene and low physical activity in H too[13]. Ascertainment of indexes through multiplex families as in the case of the GENI study could make it difficult to extrapolate results to the general population of diabetic patients. Indeed, an enrichment in disease susceptibility alleles is expected in these families and thus sibling recurrence risk estimates are likely to be increased as compared to those expected in the general population[23]. However, it should not create an erroneous heterogeneity between exposed and unexposed indexes strata unless there is a correlation between sibs for the environmental factor that is not correctly accounted for. In this example, the results are likely to encourage further studies to select non obese subjects in H populations, in order to search for genetic factors implicated in T2D, whereas studying NHW populations, we would be more interested in searching for an interaction with low physical activity. This illustrates how one can use the *ORR* point estimates, their confidence intervals and corresponding *p*-values to rank among many

environmental factors those that should be selected in priority to test for a GxE interaction in following genetic studies.

In conclusion, this paper demonstrates that valuable amount of familial information can be exploited towards detecting GxE interactions that underpin multifactorial disease susceptibility. This method is proposed as a strategy that can be used prior to genetic studies to help plan these studies. It can help investigators identify environmental factors liable to interact with genetic factors and that will need to be accounted for in the analysis but could also be used in the study design to select subcategories of the population to enhance genetic factor detection.

Acknowledgments

The authors wish to thank Marie-Claude Babron for her comments on the manuscript. RK's doctoral work is funded by a grant from the Presidency of the University Paris-Sud.

References

1 Lander ES, Schork NJ: Genetic dissection of complex traits. *Science* 1994; **265**: 2037-2048.

2 Andrieu N, Goldstein AM: Epidemiologic and genetic approaches in the study of gene-environment interaction: an overview of available methods. *Epidemiol Rev* 1998; **20**: 137-147.

3 Ottman R: Gene-environment interaction: definitions and study designs. *Prev Med* 1996; **25**: 764-770.

4 Yang Q, Khoury MJ: Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. *Epidemiol Rev* 1997; **19**: 33-43.

5 Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ: Exploiting gene-environment interaction to detect genetic associations. *Hum Hered* 2007; **63**: 111-119.

6 Selinger-Leneman H, Genin E, Norris JM, Khlat M: Does accounting for gene-environment (GxE) interaction increase the power to detect the effect of a gene in a multifactorial disease? *Genet Epidemiol* 2003; **24**: 200-207.

7 Purcell S: Variance components models for gene-environment interaction in twin analysis. *Twin Res* 2002; **5**: 554-571.

8 Stücker I, Bonaïti-Pellié C, Hémon D: Epidemiology of lung cancer: interaction between genetic susceptibility and environmental risk factors.; in: Hirsch A, Goldberg M, Martin J-P, et al (eds): Prevention of respiratory diseases. New York, Basel, Hong Kong, Marcel Dekker, 1993, pp 149-165.

9 Risch N: Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet* 1990; **46**: 222-228.

10 Khoury MJ, Beaty TH, Liang KY: Can familial aggregation of disease be explained by familial aggregation of environmental risk factors? *Am J Epidemiol* 1988; **127**: 674-683.

11 Li C, Sacks L: The derivation of the joint distribution and correlation between relatives by the use of stochastic matrices. *Biometrics* 1954; **10**: 347-360.

12 Maple 10: Maplesoft. Ontario, Canada, Waterloo Maple Inc., 2006.

13 Nelson TL, Fingerlin TE, Moss LK, Barmada MM, Ferrell RE, Norris JM: Association of the peroxisome proliferator-activated receptor gamma gene with type 2 diabetes mellitus varies by physical activity among non-Hispanic whites from Colorado. *Metabolism* 2007; **56**: 388-393.

14 Kriska AM, Knowler WC, LaPorte RE *et al*: Development of questionnaire to examine relationship of physical activity and diabetes in Pima Indians. *Diabetes Care* 1990; **13**: 401-411.

15 Lynch J, Helmrich SP, Lakka TA *et al*: Moderately intense physical activities and high levels of cardiorespiratory fitness reduce the risk of non-insulin-dependent diabetes mellitus in middle-aged men. *Arch Intern Med* 1996; **156**: 1307-1314.

16 STATA/SE 10.1. College Station, Texas, USA, Statacorp, 1984-2008.

17 Centers for disease Control and Prevention, National diabetes surveillance system, US Department of Health and Human Services. Diabetes Data for Colorado. http://apps.nccd.cdc.gov/ddtstrs/statePage.aspx?state=Colorado.

18 Mac Mahon B: Epidemiologic approaches to family resemblance; in: Morton N, Chung C (eds): Genetic Epidemiology. New York, Academic Press, 1978, pp 3-11.

19 Umbach DM, Weinberg CR: The use of case-parent triads to study joint effects of genotype and exposure. *Am J Hum Genet* 2000; **66**: 251-261.

20 Andrieu N, Goldstein AM, Thomas DC, Langholz B: Counter-matching in studies of gene-environment interaction: efficiency and feasibility. *Am J Epidemiol* 2001; **153**: 265-274.
21 Goldstein AM, Dondon MG, Andrieu N: Unconditional analyses can increase efficiency in assessing gene-environment interaction of the case-combined-control design. *Int J Epidemiol* 2006; **35**: 1067-1073.
22 Kriska AM, Saremi A, Hanson RL *et al*: Physical activity, obesity, and the incidence of type 2 diabetes in a high-risk population. *Am J Epidemiol* 2003; **158**: 669-675.
23 Weijnen CF, Rich SS, Meigs JB, Krolewski AS, Warram JH: Risk of diabetes in siblings of index cases with Type 2 diabetes: implications for genetic studies. *Diabet Med* 2002; **19**: 41-50.

Table 1 Distribution of the sample of sib pairs in cross table according to exposure of sib 1 and disease status of sib 2.

The sibling recurrence risk over the whole sample (K_S) and sibling recurrence risks stratified on sib 1's exposure (K_{SE} and $K_{S\bar{E}}$) can be derived from the observed numbers (a, b, c and d). The Odds Ratio of Recurrence (ORR) is equal to:

$$ORR = \frac{K_{SE} \times \left(1 - K_{S\bar{E}}\right)}{K_{S\bar{E}} \times \left(1 - K_{SE}\right)} = \frac{ad}{bc}$$

		Sib 1 (affected)		
		exposed	unexposed	
Sib 2	affected	a	b	$a+b$
	unaffected	c	d	$c+d$
		$a+c$	$b+d$	N
		$K_{SE} = a/(a+c)$	$K_{S\bar{E}} = b/(b+d)$	$K_S = (a+b)/\mathrm{N}$

Table 2 Probability of disease given exposure and genotype statuses according to genetic and environmental model parameters.

		Genotype	
		G– ($1–f_G$)	*G*+ (f_G)
Exposure	*E*– ($1–f_E$)	*B*	$B.RR_G$
	E+ (f_E)	$B.RR_E$	$B.RR_G.RR_E.I$

E+: exposed; *E*–: unexposed; f_E: proportion of exposed individuals in population; *G*+: carrier of the predisposing genotype(s); *G*–: non-carrier of the predisposing genotype(s); f_G: proportion of carriers of the predisposing genotype(s) in population; *B*: baseline risk; RR_E: exposure relative risk; RR_G: genotypic relative risk; *I*: multiplicative interaction coefficient for individuals both exposed and carrier of the predisposing genotype.

Table 3 Sample size (number of sib pairs) required to obtain a power of 0.80 with a type I error rate of 0.05 as a function of the interaction coefficient (I) and the environmental correlation between sibs (C_E).

Fixed parameters: disease prevalence: $f_D = 0.1$; frequency of predisposing genotype(s): $f_G = 0.1$; genotypic relative risk: $RR_G = 2$; frequency of exposure: $f_E = 0.2$; exposure relative risk: $RR_E = 2$.

		Recessive model					Dominant model				
	I	1	2	3	5	10	1	2	3	5	10
C_E	0	∞	72 417	13 885	2946	720	∞	41 516	8061	1754	448
	0.25	∞	12 250	2568	619	181	∞	8418	1745	423	126
	0.5	∞	5371	1182	307	99	∞	3875	841	219	70
	0.75	∞	3172	725	199	67	∞	2345	529	145	47
	1	∞	2163	511	147	51	∞	1625	379	108	35

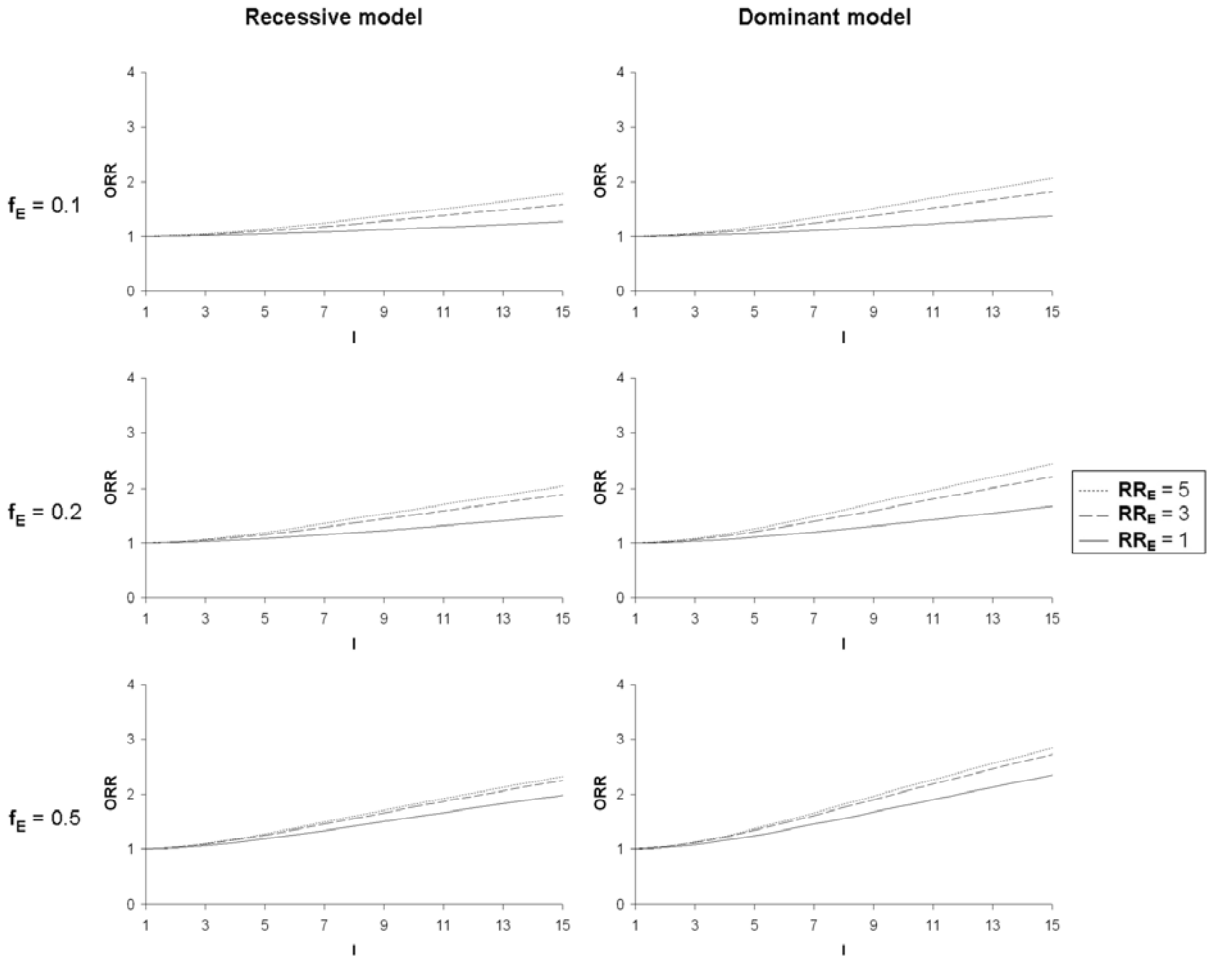
Table 4 Results of the application on type 2 diabetes data.

Environmental factor	**Obesity**		**Low physical activity**	
Stratum	**H**	**NHW**	**H**	**NHW**
OR_E	2.48	3.87	1.13	0.93
95 % CI *of* OR_E	1.18, 5.22	1.54, 9.65	0.72, 1.77	0.53, 1.65
f_E	0.29	0.37	0.31	0.23
C_E	0.22	0.14	– 0.02	0.07
ORR_0	1.25*, 1.27	1.22*, 1.25	0.99, 1.00*	0.99, 1.00*
ORR	0.67	1.03	1.14	2.13
95 % CI *of ORR*	0.40, 1.11	0.53, 1.99	0.62, 2.08	1.08, 4.19
EURECA	4.03	0.25	0.15	4.78
p-value	0.045	0.617	0.70	0.028

H: Hispanics; NHW: Non-Hispanic Whites; OR_E: Odds Ratio estimate of the environmental factor; f_E: Estimated frequency of the environmental factor; C_E: Estimated sibling correlation for the environmental factor; ORR_0: Interval of variation of the expected Odds Ratio of Recurrence under the null hypothesis; ***** Closest bounding value used to perform the test; *ORR*: Odds Ratio of Recurrence; CI: confidence interval; *EURECA*: Exposed versus Unexposed Recurrence Analysis.

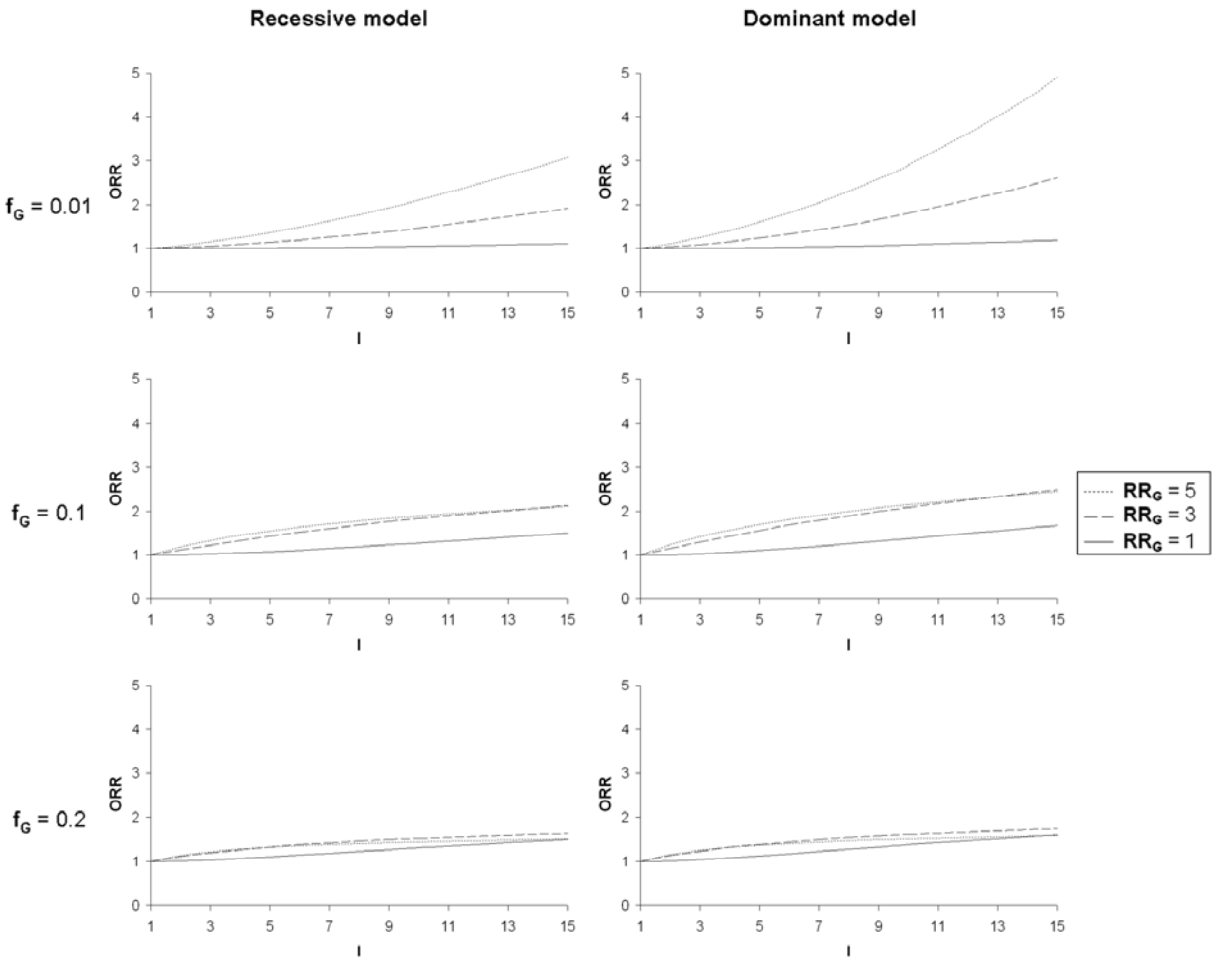
Figures

Figure 1 Odds Ratio of Recurrence (*ORR*) as a function of the gene-environment interaction coefficient (*I*) for varying exposure prevalences (f_E), varying exposure relative risks (RR_E) and for a recessive and a dominant genetic model.



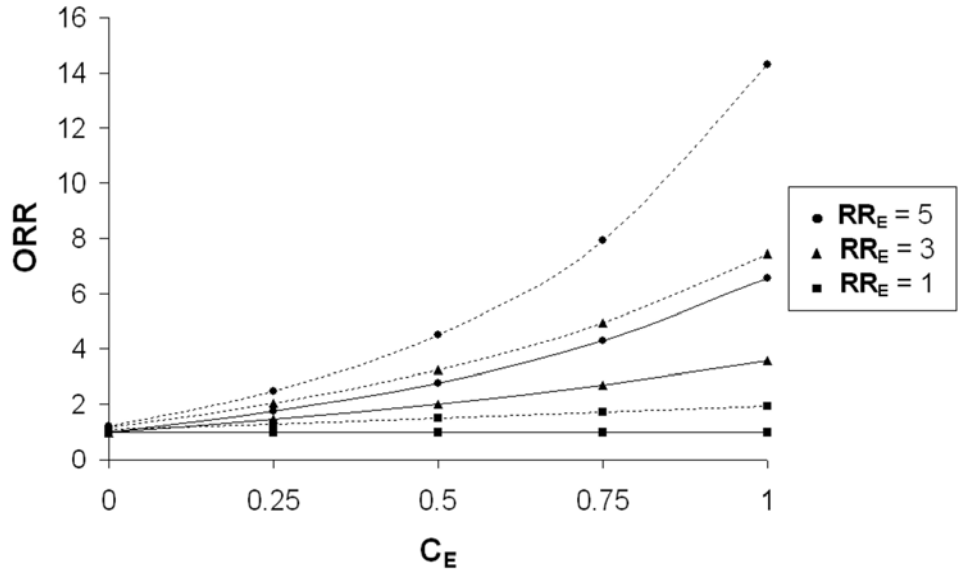
Fixed parameters: disease prevalence: $f_D = 0.1$; frequency of predisposing genotype(s): $f_G = 0.1$; genotypic relative risk: $RR_G = 1$; sibling correlation for the environmental factor: $C_E = 0$.

Figure 2 Odds recurrence ratio (*ORR*) as a function of the gene-environment interaction coefficient (*I*) for varying frequencies of predisposing genotype(s) (f_G), varying genotypic relative risks (RR_G) and for a recessive and a dominant genetic model.



Fixed parameters: disease prevalence: $f_D = 0.1$; frequency of exposure: $f_E = 0.2$; exposure relative risk: $RR_E = 1$; sibling correlation for the environmental factor: $C_E = 0$.

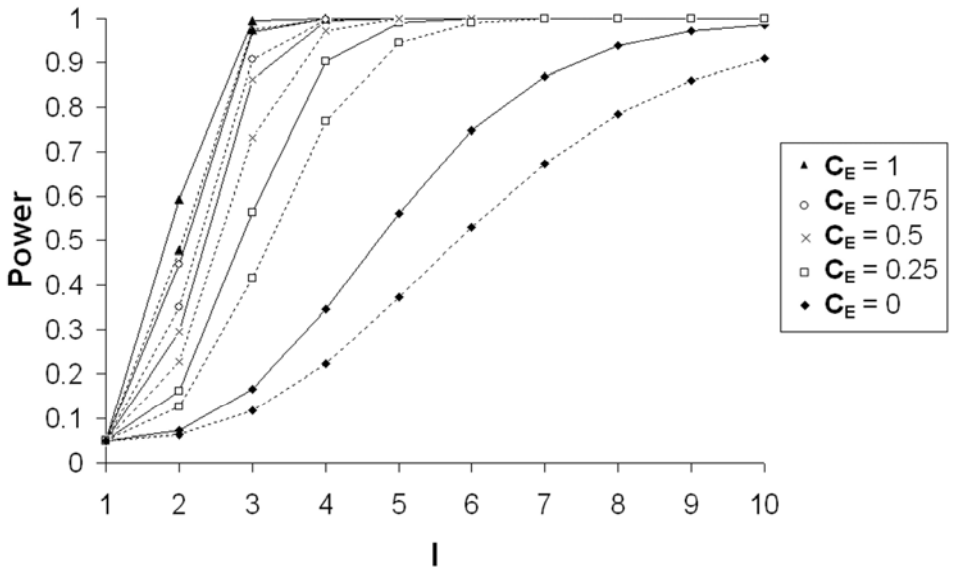
Figure 3 Odds Ratio of Recurrence (*ORR*) as a function of the sibling correlation for the environmental factor (C_E) and the environmental factor relative risk (RR_E).



Fixed parameters: disease prevalence: $f_D = 0.1$; frequency of exposure: $f_E = 0.2$; frequency of predisposing genotype(s): $f_G = 0.1$; genotypic relative risk: $RR_G = 1$.

Solid curves represent null hypothesis scenarios and dotted curves represent corresponding situations with a gene-environment interaction (I) of 5.

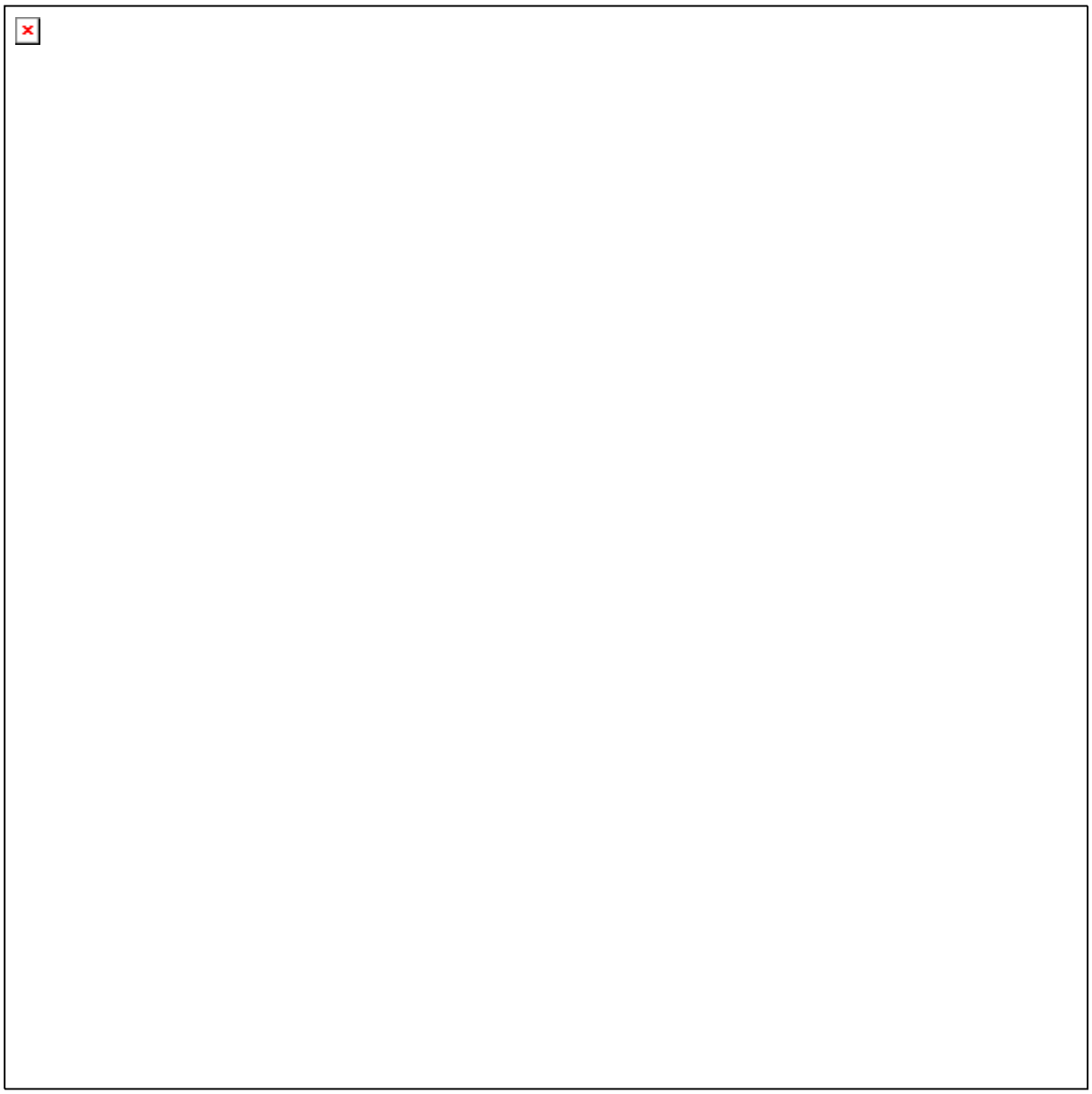
Figure 4 Power of the *EURECA* test as a function of the interaction coefficient *I* and the sibling correlation for exposure C_E, after accounting for inflated type I error rates due to C_E, considering a sample size of 1000 sib-pairs.



Fixed parameters: disease prevalence: $f_D = 0.1$; frequency of exposure: $f_E = 0.2$; exposure relative risk: $RR_E = 2$; frequency of predisposing genotype(s): $f_G = 0.1$; genotypic relative risk: $RR_G = 2$.

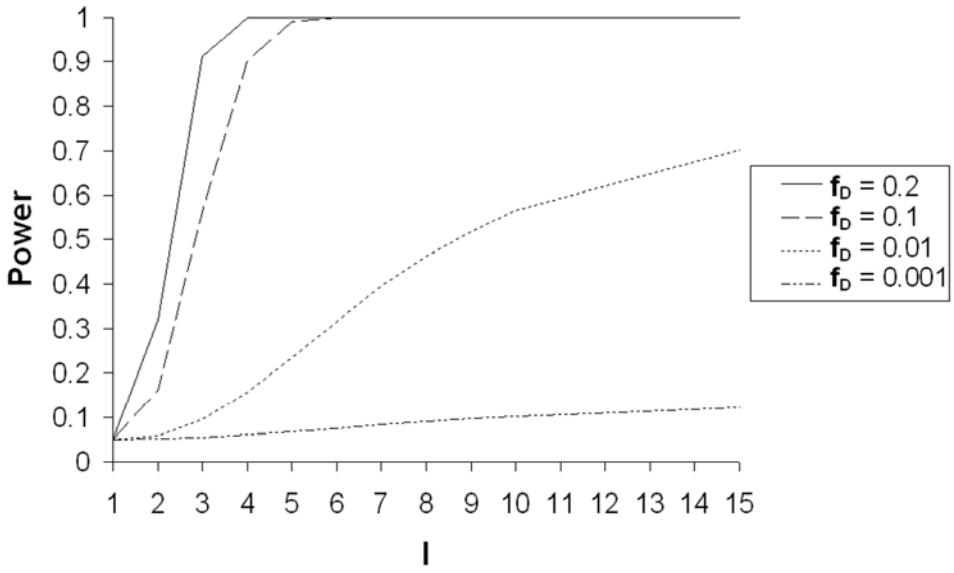
Dotted curves represent computations for recessive models and solid curves for dominant models.

Figure 5 Required sample size (number of sib pairs) to obtain a power of 0.80 with a type I error rate of 0.05 as a function of the interaction coefficient (I) for different exposure (RR_E) and genotypic (RR_G) relative risks.



Fixed parameters: disease prevalence: $f_D = 0.1$; frequency of exposure: $f_E = 0.2$; frequency of predisposing genotype(s): $f_G = 0.1$; sibling correlation for the environmental factor: $C_E = 0.25$. Dotted curves represent computations for recessive models and solid curves for dominant models.

Figure 6 Power of the *EURECA* test as a function of the interaction coefficient *I* and the disease prevalence f_D, after accounting for inflated type I error rates due to C_E, considering a sample size of 1000 sib-pairs.



Fixed parameters: frequency of exposure: $f_E = 0.2$; exposure relative risk: $RR_E = 2$; frequency of predisposing genotype(s): $f_G = 0.1$; genotypic relative risk: $RR_G = 2$; sibling correlation for the environmental factor: $C_E = 0.25$.

Supplementary materials

Computation of contingency table's observed number given model parameters

An individual can have one of three possible genotypes (for a biallelic genetic locus) and one of two exposure statuses (for a dichotomous environmental variable). Thus any two siblings may have 36 (6 × 6) possible combinations of genotypes and exposure statuses whose joint probabilities can be obtained by modifying the ITO matrix method of Li and Sacks (Li and Sacks 1954) to account for exposure probabilities (table S1). In table S2 are shown probabilities for an individual to be affected or unaffected given his genotype and his exposure status, expressed according to model parameters.

We first calculate joint probabilities of having an exposed (or unexposed) mandatorily affected sib 1 and an affected (or unaffected) sib 2 used to calculate contingency table 1 numbers. The indices "*i*" and "*j*" refer to the cells of the U_{ij} matrix presented in table S1:

$$P_a = P(\text{sib 1 } D+ \ E+ \text{ and sib 2 } D+) = \sum_{i=1}^{3}\sum_{j=1}^{6} U_{ij} \times P(D+|i) \times P(D+|j) \quad (3)$$

$$P_b = P(\text{sib 1 } D+ \ E- \text{ and sib 2 } D+) = \sum_{i=4}^{6}\sum_{j=1}^{6} U_{ij} \times P(D+|i) \times P(D+|j) \quad (4)$$

$$P_c = P(\text{sib 1 } D+ \ E+ \text{ and sib 2 } D-) = \sum_{i=1}^{3}\sum_{j=1}^{6} U_{ij} \times P(D+|i) \times P(D-|j) \quad (5)$$

$$P_d = P(\text{sib 1 } D+ \ E- \text{ and sib 2 } D-) = \sum_{i=4}^{6}\sum_{j=1}^{6} U_{ij} \times P(D+|i) \times P(D-|j) \quad (6)$$

where *D*+ is the event of being affected with disease *D* and *E*+ is the event being exposed to environmental factor *E*, "*i*" represents possible genotype-exposure combinations for sib 1 and "*j*" possible genotype-exposure combinations for sib 2.

Their sum is equal to the *a priori* probability of disease in sib 1:

$$P_{total} = P(\text{sib 1 } D+) = P_a + P_b + P_c + P_d \quad (7)$$

Finally, using equations 3 to 7, we determine contingency table 1 observed numbers:

$a = \text{N} \times P_a / P_{total}$

$b = \text{N} \times P_b / P_{total}$

$c = \text{N} \times P_c / P_{total}$

$d = \text{N} \times P_d / P_{total}$

where N is the total number of sib-pairs.

Computation of ORR_0:

The ORR_0 formula derive from the ORR considering a model with no interaction ($I = 1$). It is a difficult formula to write on a single page since it depends on 6 parameters: the susceptibility genotype(s) frequency (f_G) and relative risk (RR_G), the environmental factor frequency (f_E) and relative risk (RR_E), the environmental correlation between sibs (C_E) and the disease prevalence in population (f_D). But according to the type 2 diabetes application results, the ORR_0 seems to depend predominantly on the environmental parameters (f_E, RR_E, and C_E) and on the disease prevalence (f_D) and to a lesser extent on the genetic parameters.

Considering no effect of the genetic factor ($RR_G = 1$), the ORR_0 can be expressed as:

$$ORR_0 = \frac{C_E(1-RR_E)+(1-RR_E)f_E(1-C_E)-1}{(1-RR_E)f_E(1-C_E)-1} \times \frac{f_E(1-RR_E)[f_D(1-C_E)-1]-f_D+1}{f_D C_E(1-RR_E)+f_E(1-RR_E)[f_D(1-C_E)-1]-f_D+1}$$

But in order to obtain the expected range of values of ORR_0 as presented in table 4, computations were done for different values for the genetic model: f_G from 0.01 to 0.5 and RR_G ranging from 0.5 to 10.

Table S1 Sib-sib joint probabilities matrix U_{ij} for genotypic and exposure distributions modified from the ITO matrix method of Li and Sacks(Li and Sacks 1954).

			Sib 1						Total
			E1+			*E1–*			
			AA $i=1$	*Aa* $i=2$	*aa* $i=3$	*AA* $i=4$	*Aa* $i=5$	*aa* $i=6$	
Sib 2	*E2+*	*AA* $j=1$	$1/4\,q^2(1+q)^2 \times f_{E1+}f_{E2+/E1+}$	$1/2\,q^2(1-q^2) \times f_{E1+}f_{E2+/E1+}$	$1/4\,q^2(1-q)^2 \times f_{E1+}f_{E2+/E1+}$	$1/4\,q^2(1+q)^2 \times f_{E1-}f_{E2+/E1-}$	$1/2\,q^2(1-q^2) \times f_{E1-}f_{E2+/E1-}$	$1/4\,q^2(1-q)^2 \times f_{E1-}f_{E2+/E1-}$	$q^2 f_{E2+}$
		Aa $j=2$	$1/2\,q^2(1-q^2) \times f_{E1+}f_{E2+/E1+}$	$q(1-q)(1+q(1-q)) \times f_{E1+}f_{E2+/E1+}$	$1/2\,q(1-q)^2(2-q) \times f_{E1+}f_{E2+/E1+}$	$1/2\,q^2(1-q^2) \times f_{E1-}f_{E2+/E1-}$	$q(1-q)(1+q(1-q)) \times f_{E1-}f_{E2+/E1-}$	$1/2\,q(1-q)^2(2-q) \times f_{E1-}f_{E2+/E1-}$	$2q(1-q)f_{E2+}$
		aa $j=3$	$1/4\,q^2(1-q)^2 \times f_{E1+}f_{E2+/E1+}$	$1/2\,q(1-q)^2(2-q) \times f_{E1+}f_{E2+/E1+}$	$1/4(1-q)^2(2-q)^2 \times f_{E1+}f_{E2+/E1+}$	$1/4\,q^2(1-q)^2 \times f_{E1-}f_{E2+/E1-}$	$1/2\,q(1-q)^2(2-q) \times f_{E1-}f_{E2+/E1-}$	$1/4(1-q)^2(2-q)^2 \times f_{E1-}f_{E2+/E1-}$	$(1-q)^2 f_{E2+}$
	E2–	*AA* $j=4$	$1/4\,q^2(1+q)^2 \times f_{E1+}f_{E2-/E1+}$	$1/2\,q^2(1-q^2) \times f_{E1+}f_{E2-/E1+}$	$1/4\,q^2(1-q)^2 \times f_{E1+}f_{E2-/E1+}$	$1/4\,q^2(1+q)^2 \times f_{E1-}f_{E2-/E1-}$	$1/2\,q^2(1-q^2) \times f_{E1-}f_{E2-/E1-}$	$1/4\,q^2(1-q)^2 \times f_{E1-}f_{E2-/E1-}$	$q^2 f_{E2-}$
		Aa $j=5$	$1/2\,q^2(1-q^2) \times f_{E1+}f_{E2-/E1+}$	$q(1-q)(1+q(1-q)) \times f_{E1+}f_{E2-/E1+}$	$1/2\,q(1-q)^2(2-q) \times f_{E1+}f_{E2-/E1+}$	$1/2\,q^2(1-q^2) \times f_{E1-}f_{E2-/E1-}$	$q(1-q)(1+q(1-q)) \times f_{E1-}f_{E2-/E1-}$	$1/2\,q(1-q)^2(2-q) \times f_{E1-}f_{E2-/E1-}$	$2q(1-q)f_{E2-}$
		aa $j=6$	$1/4\,q^2(1-q)^2 \times f_{E1+}f_{E2-/E1+}$	$1/2\,q(1-q)^2(2-q) \times f_{E1+}f_{E2-/E1+}$	$1/4(1-q)^2(2-q)^2 \times f_{E1+}f_{E2-/E1+}$	$1/4\,q^2(1-q)^2 \times f_{E1-}f_{E2-/E1-}$	$1/2\,q(1-q)^2(2-q) \times f_{E1-}f_{E2-/E1-}$	$1/4(1-q)^2(2-q)^2 \times f_{E1-}f_{E2-/E1-}$	$(1-q)^2 f_{E2-}$
Total (ω_i)			$q^2 f_{E1+}$	$2q(1-q)f_{E1+}$	$(1-q)^2 f_{E1+}$	$q^2 f_{E1-}$	$2q(1-q)f_{E1-}$	$(1-q)^2 f_{E1-}$	1

A: allele that confers susceptibility to disease; *E*1*X*: exposure status of sib 1 (*X* = + if exposed and *X* = – if unexposed); *E*2*Y*: exposure status of sib 2 (*Y* = + if exposed and *Y* = – if unexposed); *q*: frequency of allele A in population; f_{E1X}: frequency of exposed status *X* in sib 1 (equal to the same frequencies as in population); $f_{E2Y/E1X}$: frequency of exposed status *Y* in sib 2 given exposure status *X* of sib 1 (equal to same frequencies as in population if exposure correlation coefficient for sib-pairs is null, $C_E = 0$

Table S2 Probabilities for an individual to be affected ($P(D+ \mid k)$) or unaffected ($P(D- \mid k)$) for the six possible combinations of genotype and exposure statuses.

	$E+$			$E-$		
	AA $k = 1$	Aa $k = 2$	aa $k = 3$	AA $k = 4$	Aa $k = 5$	aa $k = 6$
$P(D+ \mid k)$	P_{EG}	P_E[1] P_{EG}[2]	P_E	P_G	P_B[1] P_G[2]	P_B
$P(D- \mid k)$	$1-P_{EG}$	$1-P_E$[1] $1-P_{EG}$[2]	$1-P_E$	$1-P_G$	$1-P_B$[1] $1-P_G$[2]	$1-P_B$

[1] if autosomal recessive transmission

[2] if autosomal dominant transmission

A: allele that confers susceptibility to disease; $E+$: exposed; $E-$: unexposed; B: baseline risk; RR_E: exposure relative risk; RR_G: genotypic relative risk; I: Interaction coefficient; P_B: probability of disease in non-exposed and non-carrier of susceptibility genotype individual ($P_B = B$); P_E: probability of disease in exposed and non-carrier of susceptibility genotype individual ($P_E = B \times RR_E$); P_G: probability of disease in non-exposed and carrier of susceptibility genotype individual ($P_G = B \times RR_G$); P_{EG}: probability of disease in exposed and carrier of susceptibility genotype individual ($P_{EG} = B \times RR_E \times RR_G \times I$).

Annexe 2

Kazma R, Babron MC, Génin E.
Genetic association and gene-environment interaction: A new method to overcome the lack of exposure information in controls
Am J Epidemiol 2010, en révision

Genetic association and gene-environment interaction:

A new method to overcome the lack of exposure information in controls

ABBREVIATIONS

D, disease variable; d.f., degree of freedom; *E*, environmental factor; *G*, genetic factor; GxE, gene-environment; GWAS, Genome-Wide Association Scan; LRT, Likelihood Ratio Test; *OR*, odds-ratio; SNP, Single Nucleotide Polymorphism

KEYWORDS:

gene-environment interaction; reference control panel; multinomial logistic regression; epidemiologic research design; genetic predisposition to disease; gene-environment correlation

RUNNING HEAD:

GxE interaction with reference control panels

WORD COUNT:

abstract: 195, main manuscript: 3497, 4 figures and 4 tables.

ABSTRACT

The use of a reference control panel in genome-wide association studies is an interesting solution to reduce costs. In such designs, relevant environmental factors are usually collected only in cases adding a difficulty to deal with potential gene-environment interactions when testing for genetic association. Yet, under certain circumstances, neglecting an existing interaction with the environment may be detrimental in term of power to detect the genetic factor.

In this paper, a novel method based on a multinomial logistic regression model is proposed to overcome the lack of environmental exposure information in controls, by contrasting both exposed and unexposed cases against the control sample. For each case group, a genetic effect size parameter is estimated and the genetic association and the gene-environment interaction are tested jointly. The performance of our method is evaluated through asymptotic computations and simulations of cases and population controls under different models.

In presence of a gene-environment interaction, our approach outperforms all the other available methods that test for genetic association and gene-environment interaction either separately or jointly. Interestingly, it even has a better power than the joint test that needs full knowledge of the environmental information in both cases and controls.

INTRODUCTION

Most prevalent human diseases (cancers, cardio-vascular diseases, asthma, neuropsychiatric diseases, diabetes …) are usually associated to multiple genetic and environmental factors that act jointly rather than independently.(1) With the advent of the human genome sequencing and the HapMap project depicting distribution patterns of Single Nucleotide Polymorphisms on the whole genome, a plethora of Genome-Wide Association Scans (GWAS) of multifactorial diseases has been conducted leading to the successful identification of more than a hundred association signals in less than 5 years.(2) In the last few years, huge samples of genotypes from various populations over the world have been collected. These genotype data are usually available to the scientific community can be used as controls in association studies. Using such samples as reference control panels is a very tempting solution to reduce costs and has already been implemented in very large consortium studies(3-7) where the same individuals serve as controls against several sets of cases with different diseases. The drawback of this design is that the relevant environmental factors for a specified disease are usually collected only in cases and not in controls making it difficult to address the gene-environment (GxE) interaction topic.

When dealing with GxE interaction in genetic studies of complex diseases, two complementary rather than opposite strategies can be outlined. The first strategy consists in simultaneously testing for genetic association and GxE interaction arguing that neglecting an existing interaction with an environmental factor may hinder the detection of the genetic factor.(8) Kraft *et al* have shown that under a wide variety of GxE interaction models, such a combined test outperforms a marginal association test in terms of power.(9) But this strategy necessitates collecting exposure measurements on both cases and controls. The second strategy consists in testing for the GxE interaction alone arguing that after an initial GWAS that tests for the genetic effect, such an approach would be independent and identify new

genetic variants involved in pure GxE interaction.(10) In that context, a logistic regression based case-control design can model explicitly the GxE interaction, but it requires full information on genetic and environmental data of both samples. When only the exposures of cases are available, a case-only design can be implemented by regression of the genotype (or allele) counts on the environmental variable.(11, 12) The case-only design is notably more powerful than the case-control design. However it can only estimate and test for the GxE interaction term and it is prone to false-positives should a gene-environment correlation exist in the underlying population. Regardless of the choice to test for GxE interaction or not, most widespread methods in genetic epidemiology are based on the binomial form of the logistic model, where the outcome is binary and the dependant covariates are observed in all subjects. The absence of environmental information in controls makes such methods unfeasible in practice. The development of an alternative approach in such situation would trigger new perspectives for GxE interaction studies at the genome-wide level.

The purpose of this paper is to present a novel method based on a multinomial logistic regression model that contrasts both exposed and unexposed case samples against a control sample with no information collected on the environmental exposure. We evaluate the performance (type I error-rates and power) of this approach and compare it to the different genetic association and GxE interaction tests available under various scenarios of GxE interaction. The bias, variance and coverage probability of the estimated genetic parameters are also compared between the multinomial modeling and the other approaches.

MATERIAL AND METHODS

Disease penetrance model and notations

Let us consider a disease phenotype D, a genetic risk factor G and an environmental risk factor E. The three variables are assumed to be dichotomous, 0 denoting absence and 1 presence of the disease or risk factor. The genetic factor is an autosomal biallelic genetic locus with a susceptibility allele S with population frequency q. We chose to model all along the study a dominant genetic effect. Therefore, the variable G corresponds to having at least one copy of the S allele and the frequency of the susceptibility genotypes is $f_G=q^2+2q(1-q)$ under the assumption that genotypes are in Hardy-Weinberg proportions in the studied population. The main effect size of G is measured by the genotypic odds-ratio OR_G equal to the ratio of the odds of G in cases and controls:

$$OR_G = \frac{P(G=1|D=1)/P(G=0|D=1)}{P(G=1|D=0)/P(G=0|D=0)} \quad (1)$$

The main effect of E on the disease is measured by the environmental odds ratio OR_E equal to the ratio of the odds of E in cases and controls:

$$OR_E = \frac{P(E=1|D=1)/P(E=0|D=1)}{P(E=1|D=0)/P(E=0|D=0)} \quad (2)$$

A possible gene-environment correlation is introduced by considering a coefficient θ_{GE} as in Lindström *et al*.(13)

$$\theta_{GE} = \frac{P(E=1|G=1)/P(E=0|G=1)}{P(E=1|G=0)/P(E=0|G=0)} \quad (3)$$

The base probability of E=1 given G=0 is denoted f_E. When θ_{GE}=1 (independence between G and E), $P(E=1|G=1)=P(E=1|G=0)=f_E$. The gene-environment correlation θ_{GE} can then vary between 0 and $+\infty$ measuring the degree of association between G and E in the population.

Considering the joint distributions of G and E, the odds-ratio associated to the presence of both G and E is measured by OR_{GE}:

$$OR_{GE} = \frac{P(E=1,G=1|D=1)/P(E=0,G=0|D=1)}{P(E=1,G=1|D=0)/P(E=0,G=0|D=0)} \quad (4)$$

In absence of GxE interaction and under a multiplicative model at the odds scale (or an additive model at the log-odds scale), OR_{GE} is expected to be equal to the product of OR_G and OR_E. Using equations (1), (2) and (4), the interaction term OR_I that measures the departure from this condition is expressed as:

$$OR_I = \frac{OR_{GE}}{OR_G \times OR_E} \quad (5)$$

Multinomial logistic regression model

The multinomial (or polychotomous) logistic regression model (14, 15) is a logistic regression model extended for nominal outcome variable with more than two categories. Widely used in traditional epidemiologic studies, it has recently been shown to be also a powerful approach in the genetic study of disease sub-phenotypes.(16) It is usually used to contrast subcategories of phenotypes against the control group and can therefore be extended to our particular context where environmental information is not collected in the control group (where D=0). For that purpose, the D and E dichotomous variables are combined into a single three-class multinomial variable D_M that takes the value of 0 in controls, 1 in unexposed cases and 2 in exposed cases. The coding scheme for the three categories can be swapped as the multinomial model is non-ordinal, but for convenient interpretation of parameters, it is preferable to use the control group (D_M=0) as the reference group. The multinomial model can be defined as:

$$\text{logit } P(D_M=j|G)=\log[P(D_M=j|G)/P(D_M=0|G)]=\beta_{0j}+\beta_{Gj}G \quad (6)$$

with j taking the values 1 and 2. The two corresponding logit equations are used simultaneously to estimate the two genetic parameters, β_{G1} and β_{G2} that maximize the overall

likelihood. The β_{G1} and β_{G2} parameters represent the genetic odds-ratios in unexposed and exposed cases as compared to the whole control group, respectively. A combined test (referred to as Multinomial-GI) of the genetic association and GxE interaction can be carried out by testing the null hypothesis $\beta_{G1}=\beta_{G2}=0$ with a 2-df Likelihood Ratio Test (LRT) comparing the extended multinomial model in equation 6 with the nested model constraining both β_{Gj} parameters to be equal to 0.

Comparison with binomial logistic regression models

The multinomial logistic regression approach was compared to different binomial regression approaches commonly used in the literature to test for association or GxE interaction. These approaches differ by the exposure information they require: in both cases and controls, in cases only or in no individual. They also differ in the hypotheses they test: *G* effect and GxE interaction, GxE interaction only or *G* effect only (see Table 1 for a summary of all methods).

Combined genetic and GxE interaction test (referred to as Binomial-GI) – When information on *E* is available in both cases and controls, the full logistic model writes: logit $P(D=1|G,E)=\beta_0+\beta_E E+\beta_{GCC}G+\beta_{ICC}GE$. As suggested by Kraft *et al*, (9) the association between *G* and *D* and the GxE interaction can be tested for using a combined 2-df LRT of the null hypothesis $\beta_{GCC}=\beta_{ICC=}0$.

GxE interaction test in a case-control design (referred to as Case-Control-I) – A second possibility is to test for the GxE interaction (β_{ICC}) in the full model with a 1-df LRT of the null hypothesis $\beta_{ICC=}0$.

Adjusted genetic association test (referred to as Adjusted-G) – A third possibility consists in only adjusting on *E* without accounting for GxE interaction. The adjusted logistic model is logit $P(D=1|G,E)=\beta_0+\beta_E E+\beta_{GA}G$ and the null hypothesis of no association between *G* and *D* ($\beta_{GA}=0$) is tested with a 1-df LRT.

GxE Interaction test in a case-only design (referred to as Case-Only-I) – Piegorsch *et al*(12) presented this design as a more powerful alternative to test for GxE interaction using only a sample of cases. The procedure consists in considering the exposure as the dependant variable on which the genotype variable is regressed. The corresponding case-only logistic model is logit P(E=1|G)=β_0+$\beta_{ICO}G$ and similarly the GxE interaction (β_{ICO}=0) is tested with a 1-df LRT.

Marginal genetic association test (referred to as Marginal-G) – When information on E is available neither in cases nor in controls, we can only model the marginal effect of G in a logistic regression model: logit P(D=1|G)=β_0+$\beta_{GM}G$. The null hypothesis of no association between G and D (β_{GM}=0) is tested with a 1-df LRT.

Computations, simulations and evaluation criteria

The disease probability conditional on G and E was modeled using a logit function as follow:

$$\text{logit P}(D=1|G, E)=\beta_0+\beta_G G+\beta_E E+\beta_I GE \quad (7)$$

where β_0=log(B/1–B) with B, the baseline risk of disease (i.e., the probability of disease given G=0 and E=0), β_G=log(OR_G), β_E=log(OR_E) and β_I=log(OR_I).

Using Bayes' theorem and equations 1-3, 5 and 7, the expected probabilities of all the categories of G=i and E=j conditional on D=k are:

$$\text{P}(G=i,E=j|D=k)=\text{P}(D=k|G=i,E=j)\times\text{P}(G=i)\times\text{P}(E=j|G=i)/\text{P}(D=k) \quad (8)$$

where P(D=k|G=i,E=j) is derived from equation 7 and P(E=j|G=i) is a function of θ_{GE}, f_E and f_G. In order to mimic a reference control panel design, a misclassification bias was considered in controls, assuming that f_D=10% of the controls are in fact affected individuals.

We considered values of f_G and f_E ranging from 0.1 to 0.9 by intervals of 0.2, values of OR_G, OR_E and OR_I ranging from 0.5 to 2.0 (to 3.0 for OR_I) by intervals of 0.25 and values of θ_{GE} ranging from 1.0 to 2.0 by intervals of 0.25 and for a disease prevalence f_D=0.1. Then, for

each set of parameters, we computed the expected cell counts using equation 8 and simulated 1,000 replicates considering a study with 500 cases and 500 controls.

Asymptotic type I error and power of the different tests in Table 1 are estimated by use of non-central χ^2-distributions with corresponding df at a nominal type I error-rate of 0.01. They are also estimated by the proportion of simulated replicates with a *P*-value lower or equal to 0.01. Since both asymptotic and simulation-based results were similar, only the asymptotic type I errors and powers are reported here. The squared bias, the variance and the coverage probability (probability that the 95% confidence interval of the estimates contains the theoretical value) of the different parameters of the *G* and GxE interaction effects estimated by the different models were computed on the simulated datasets.

All computations and simulations were done using a script written in R(17) and statistical analyses were done using the “logit” and “mlogit” functions of StataSE(18).

RESULTS

Type I error-rates

Type I error-rates obtained under the null hypothesis of no *G* effect and no GxE interaction are presented in Figure 1 as a function of the other parameters (θ_{GE}, f_G, OR_E and f_E). In absence of gene-environment correlation (θ_{GE}=1), all tests have a type I error-rate consistent with its nominal value of 0.01, whereas in presence of correlation (θ_{GE}>1), type I errors of the Case-Only-I and the Multinomial-GI tests are both increased and the magnitude of the inflation depends mainly on the values of θ_{GE} (the more it deviates from 1, the higher), f_G and f_E (with a maximum inflation for values between 0.3 and 0.5). Type I errors of the Marginal-G test are also slightly inflated, but do not compare to the inflation of the other two methods. The maximal inflation is of 3.3% for a model with θ_{GE}=2, OR_E=2, f_G=0.5 and f_E=0.3. The other three methods that use full information on the exposure in both cases and controls are robust to the presence a gene-environment correlation.

Power in absence of gene-environment correlation (θ_{GE}=1)

Figures 2 and 3 represent asymptotic powers of the different tests as a function of GxE interaction under different scenarios and in absence of any gene-environment correlation (θ_{GE}=1). The Multinomial-GI test (solid red lines) performs better or as well as all other tests in the presence of a GxE interaction (Figure 2) with an overall better performance of all tests for frequencies of *G* and *E* close to 0.5 (Figure 3).

Compared to the tests of *G* effect only (Marginal-G or Adjusted-G in blue in Figure 2), a slight loss of power of the Multinomial-GI test is found in the absence of GxE interaction as expected since stratifying the case sample brings little information and increases both the variance of parameters and the df of the test. The difference in power is however at most 11.25% when OR_G=1.5 and f_G=0.5. In contrast, in presence of GxE interaction effects, the

gain in power is much higher, particularly for pure interaction effects in absence of main effects of *G* and *E*. In such situation the Multinomial-GI test improves the power by 42% and 54% compared to the Marginal-G and Adjusted-G tests, respectively. Adjusting on the exposure globally decreases power in all situations except when either the *E* effect (OR_E) or the GxE interaction (OR_I) are in opposite direction with the *G* effect (see first line and first column of Figure 2).

Compared to the two GxE interaction tests (green lines), the Multinomial-GI test performs much better when *G* has a main effect ($OR_G \neq 1$). When $OR_G=1$, it performs similarly to the Case-Only-I test. The Case-Control-I test that uses exposure information in both samples has the lowest power over all models.

Finally, the most striking point evidenced in Figure 2 is that, over all studied models, the Multinomial-GI test outperforms the Binomial-GI test that simultaneously test for the *G* and GxE interaction effects (Kraft *et al*(9)) and requires the exposure information in controls.

Power in presence of a gene-environment correlation ($\theta_{GE}>1$)

Power computations under scenarios that included a gene-environment correlation ($\theta_{GE}>1$) were adjusted for the inflated type I error-rates by using a corrected threshold instead of the central χ^2 threshold (threshold is 6.63 and 9.21 for a central χ^2 with respectively 1 and 2 df and 0.01 nominal type I error-rate). These corrected thresholds were computed using a non-central χ^2 distribution with a non-centrality parameter equal to the value of the test under the corresponding scenario with the same θ_{GE} value, but with no *G* and GxE interaction effects (null hypothesis).

With increasing values of θ_{GE}, an important decrease in power can be observed for the two methods that had an inflated type I error (Case-Only-I and Multinomial-GI tests) especially for small GxE interactions and under flip-flop scenarios where $OR_I<1$ (Figure 4). Using the

power when θ_{GE}=1 as reference value, Table 2 quantifies variations in power for increasing θ_{GE}. For strong GxE interaction, all six tests have an increase in power when θ_{GE} increases but there is a high power loss for both the Multinomial-GI and the Case-Only-I tests under flip-flop scenarios (OR_I<1).

Bias and variance of genetic and GxE interaction estimators

The bias, the variance and the coverage probability of the different genetic parameters are reported in Table 3. The β_{G1} parameter of the multinomial model has bias values very similar to the parameter β_{GCC} of the full logistic model and lower than the β_{GM} and β_{GA} bias values of the marginal and adjusted models. The variance of β_{G1} is lower than that of β_{GCC} but higher than that of β_{GM} and β_{GA}, β_{GM} having the lowest variance of all genetic parameters. The coverage probabilities of both β_{G1} and β_{GCC} confidence intervals are very close to the expected 95% value but decreases with increasing OR_G whereas those of β_{GM} and β_{GA} are lower, particularly for elevated values of OR_E and OR_I.

A GxE interaction coefficient estimator can be derived from the ratio of both parameters of the multinomial model: β_{G2}/β_{G1}. This estimator has exactly the same bias, variance and coverage probability patterns as the parameter derived from the logistic model of the case-only design (β_{ICO}). Compared to the estimator of the full model using the case-control design (β_{ICC}), β_{ICO} and β_{G2}/β_{G1} have a higher bias and lower coverage probability when OR_I increases and a lower variance over all models (Table 4).

DISCUSSION

The multinomial logistic regression is a simple and efficient model to compare exposed and unexposed cases versus controls when the exposure information is available in cases only. It allows a combined test of genetic association and GxE interaction that merges together a marginal test of genetic association regardless of the exposure and a case-only GxE interaction test regardless of the genotype distribution in controls. By combining these two designs into a unified approach, we show here that it is possible to maintain a satisfactory power while fulfilling the two concerns of detecting a potential genetic factor and not missing it because it interacts with a particular environment. This is well exemplified by the fact that our approach is similar or only slightly less powerful than both tests of *G* effect only in the absence of GxE interaction and than the Case-Only-I test in the presence of pure GxE interaction (no main *G* effect), situations where these two tests are respectively the most powerful.

Interestingly, over all the GxE interaction models explored, the Multinomial-GI test outperformed the Binomial-GI test proposed by Kraft *et al* (9) that makes use of the environmental statuses of cases and controls. This unexpected result suggests that as long as independence between *G* and *E* holds, the exposure information of controls is not mandatory to explore the effect of *G* and its interaction with *E*. However, exposure information in controls becomes more important when there is gene-environment correlation in the underlying population. In this situation, this information becomes crucial to avoid false positive detection of GxE interactions as shown by the inflated type I errors of the Multinomial-GI and the Case-Only-I tests. The non robustness of the case-only design to the presence of gene-environment correlation is well-known in the literature and has sometimes limited its use to detect GxE interaction. However, being aware of this problem, we think the Multinomial-GI test is worth considering when exposure information is lacking in controls as

it is often the case in large scale association studies where a reference control panel is used. One could then either rely on previous studies in the same population to exclude an underlying genetic correlation with the environmental factor studied or study the association between both factors in a second step in a more specific control group where exposure information is available. This strategy is also discussed in the case-only design literature(19-22).

With a better variance and a similar precision, the estimators of the genetic effect and of the GxE interaction of the multinomial model have better coverage probabilities than estimators obtained with methods that account for the exposure status of controls, under the assumption of independence between *G* and *E*.

Throughout this study, we only explored positively correlated risk factors (θ_{GE}>1). Scenarios where the *G* and *E* factors are negatively correlated but are positively interacting on the disease seem very unlikely. We also concentrated on dominant genetic models but similar trends with overall lower power values would be expected if assuming a recessive genetic model. We also explored flip-flop GxE interactions where the risk genotype becomes protective when the exposure status changes. In this situation, the Multinomial-GI test proves to be very interesting and to outperform all other methods. Flip-flop interactions are rarely described in real datasets, probably because it is not a common situation, but also maybe because traditional logistic approaches have not enough power to detect them.

A very similar approach to the Multinomial-GI was proposed by Umbach and Weinberg using log-linear models.(23) Their model takes advantage of the gene-environment independence assumption to parameterize, additionally to the gene-environment interaction, an environmental (or a genetic) main effect with a case-control design and unavailable genotypes (or exposures) in controls. According to their argumentation, the logistic framework would be unable to explicitly impose the independence assumption. However, our method has the same

sampling design and a similar parameterization. The main difference between the log-linear and the logistic modeling resides in the choice of the risk factors estimator, either relative risk or odds-ratio, respectively.(24)

Easy to implement in most existing statistical packages (Stata, SAS, R), the multinomial model is flexible and can also handle diseases categorized into sub-phenotypes,(16) as well a multinomial environmental exposures and combination of the two. As in the standard logistic regression, adjusting on specific covariates such as age, sex or recruitment category is possible. We believe it could help in improving our understanding and appraisal of GxE interactions in genetic association studies particularly at the genome-wide level provided a few “interesting” exposures could be selected to be tested for.(25)

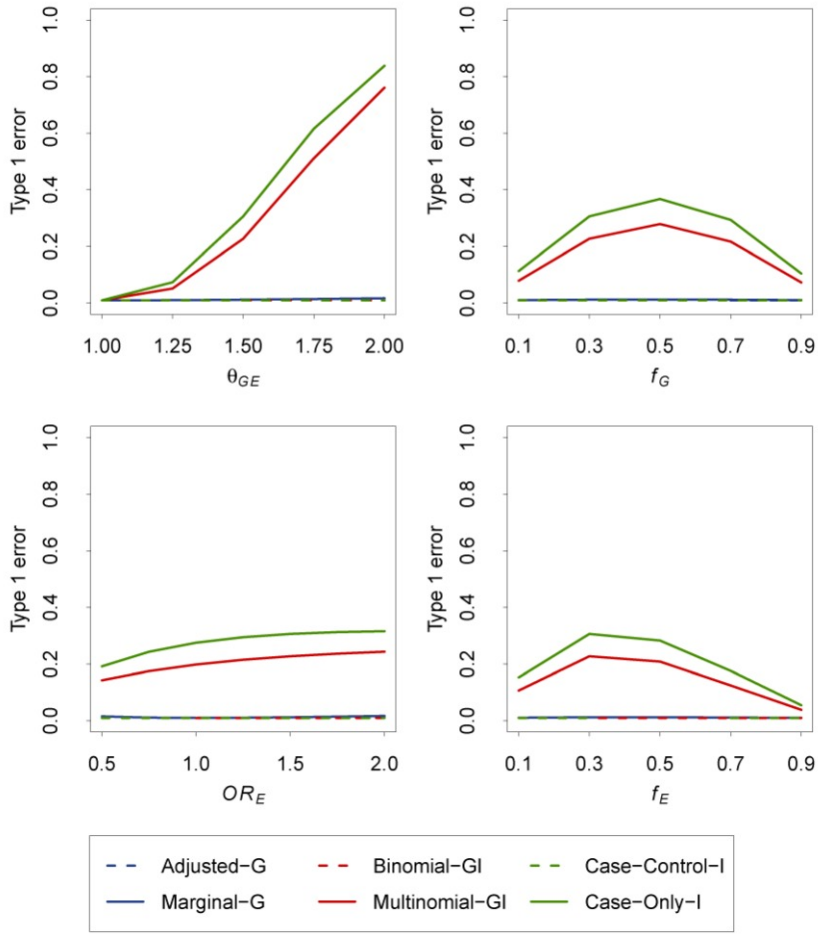
REFERENCES

1. Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994;265:2037-48.
2. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. J Clin Invest 2008;118:1590-605.
3. Luca D, Ringquist S, Klei L, et al. On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet 2008;82:453-63.
4. Nelson MR, Bryc K, King KS, et al. The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. Am J Hum Genet 2008;83:347-58.
5. The GAIN Collaborative Research Group. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. Nat Genet 2007;39:1045-51.
6. The 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.
7. Wichmann HE, Gieger C, Illig T. KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen 2005;67 Suppl 1:S26-30.
8. Selinger-Leneman H, Genin E, Norris JM, et al. Does accounting for gene-environment (GxE) interaction increase the power to detect the effect of a gene in a multifactorial disease? Genet Epidemiol 2003;24:200-7.
9. Kraft P, Yen YC, Stram DO, et al. Exploiting gene-environment interaction to detect genetic associations. Hum Hered 2007;63:111-9.
10. Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. Am J Epidemiol 2009;169:219-26.

11. Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! Am J Epidemiol 1996;144:207-13.
12. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 1994;13:153-62.
13. Lindstrom S, Yen YC, Spiegelman D, et al. The impact of gene-environment dependence and misclassification in genetic association studies incorporating gene-environment interactions. Hum Hered 2009;68:171-81.
14. Dobson AJ. Nominal and Ordinal Logistic Regression. An introduction to generalized linear models. Boca Raton: Chapman & Hall/CRC, 2002:135-50.
15. Kleinbaum DG, Klein M. Polytomous logistic regression. Logistic regression. New-York: Springer-Verlag, 2002:267-92.
16. Morris AP, Lindgren CM, Zeggini E, et al. A powerful approach to sub-phenotype analysis in population-based genetic association studies. Genet Epidemiol 2009.
17. R version 2.9.1. R Development Core Team. Vienna, Austria.
18. StataSE version 10. StataCorp. College Station, Texas, USA.
19. Albert PS, Ratnasinghe D, Tangrea J, et al. Limitations of the case-only design for identifying gene-environment interactions. Am J Epidemiol 2001;154:687-93.
20. Gatto NM, Campbell UB, Rundle AG, et al. Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias. International journal of epidemiology 2004;33:1014-24.
21. Goldstein AM, Andrieu N. Detection of interaction involving identified genes: available study designs. Journal of the National Cancer Institute 1999:49-54.

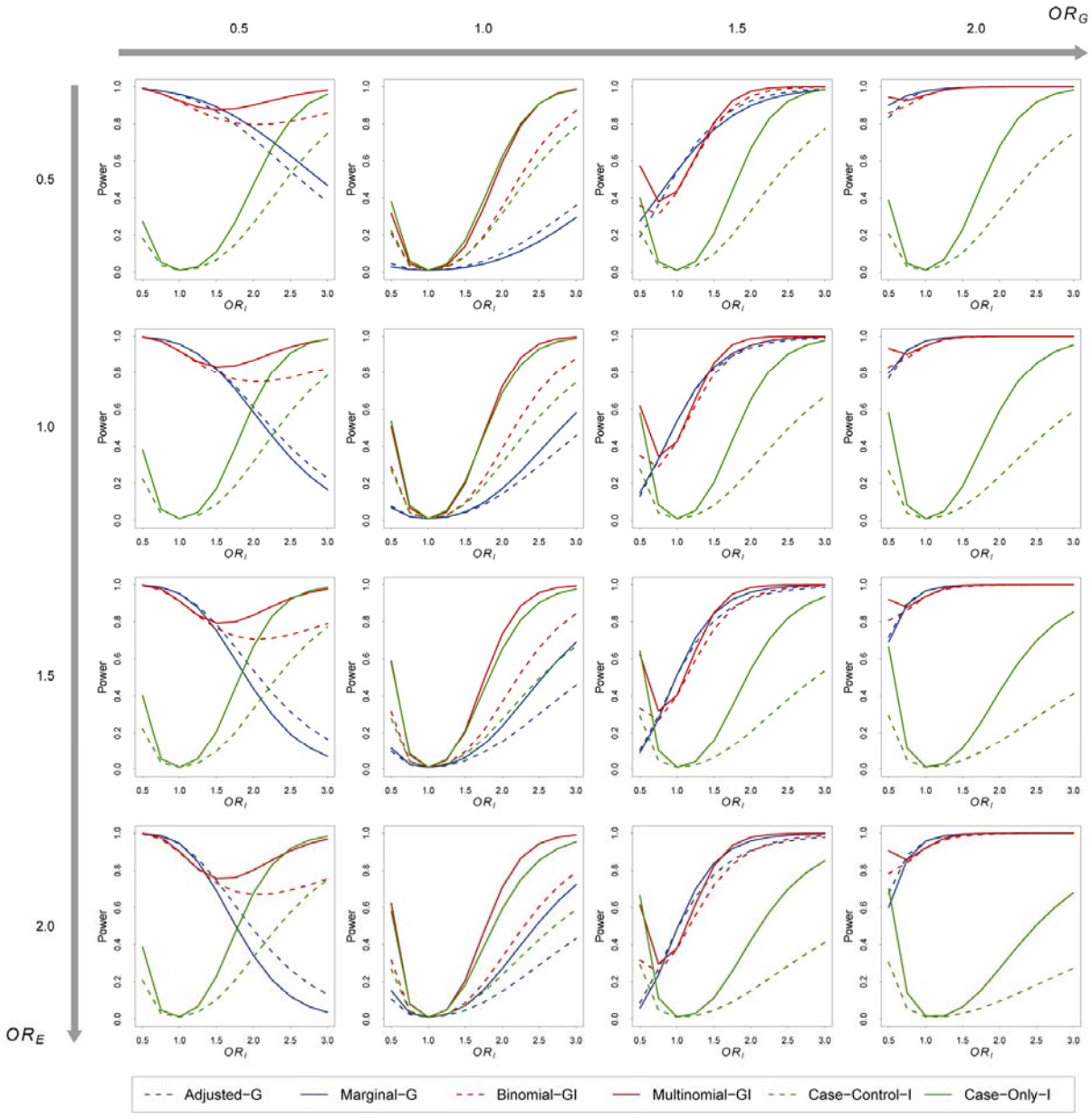
22. Schmidt S, Schaid DJ. Potential misinterpretation of the case-only study to assess gene-environment interaction. Am J Epidemiol 1999;150:878-85.
23. Umbach DM, Weinberg CR. Designing and analysing case-control studies to exploit independence of genotype and exposure. Stat Med 1997;16:1731-43.
24. Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. Stat Med 2002;21:35-50.
25. Kazma R, Bonaiti-Pellie C, Norris JM, et al. On the use of sibling recurrence risks to select environmental factors liable to interact with genetic risk factors. Eur J Hum Genet 2010;18:88–94.

FIGURE 1. Type I error-rates as a function of the gene-environment correlation (θ_{GE}), the risk genotypes frequency (f_G), the environmental main effect (OR_E) and the environmental exposure frequency (f_E).



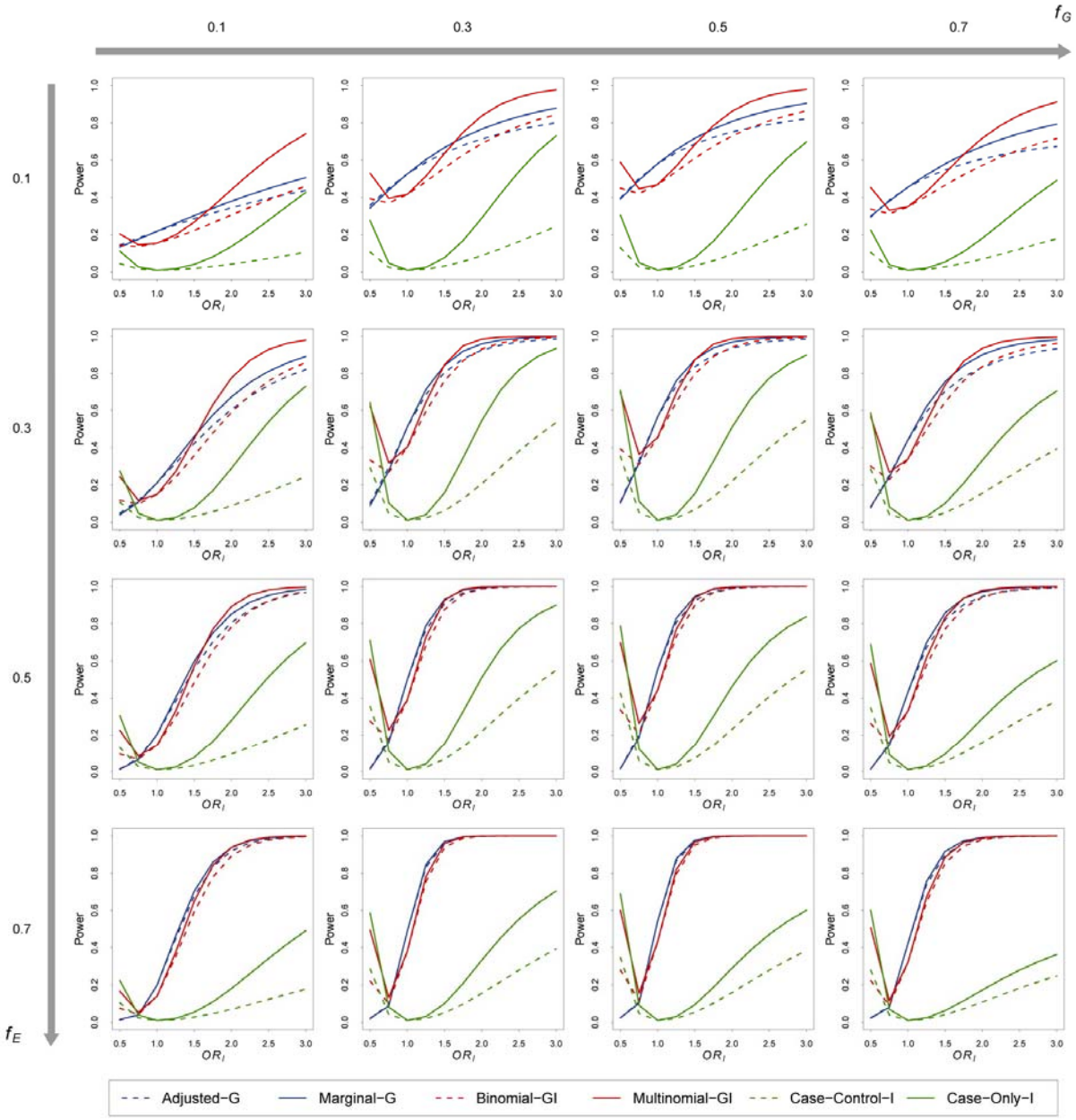
The base model from which these parameters vary is the following: genetic main effect, OR_G=1; GxE interaction effect, OR_I=1; disease prevalence, f_D=0.1; gene-environment correlation, θ_{GE}=1.5; environmental main effect, OR_E=1.5; environmental exposure frequency, f_E=0.3; risk genotypes frequency, f_G=0.3; bilateral nominal type I error-rate of 0.01 for a sample of 500 cases and 500 controls

FIGURE 2. Asymptotic power as a function of the GxE interaction (OR_I) for a range of genetic main effects (OR_G) and environmental main effects (OR_E)



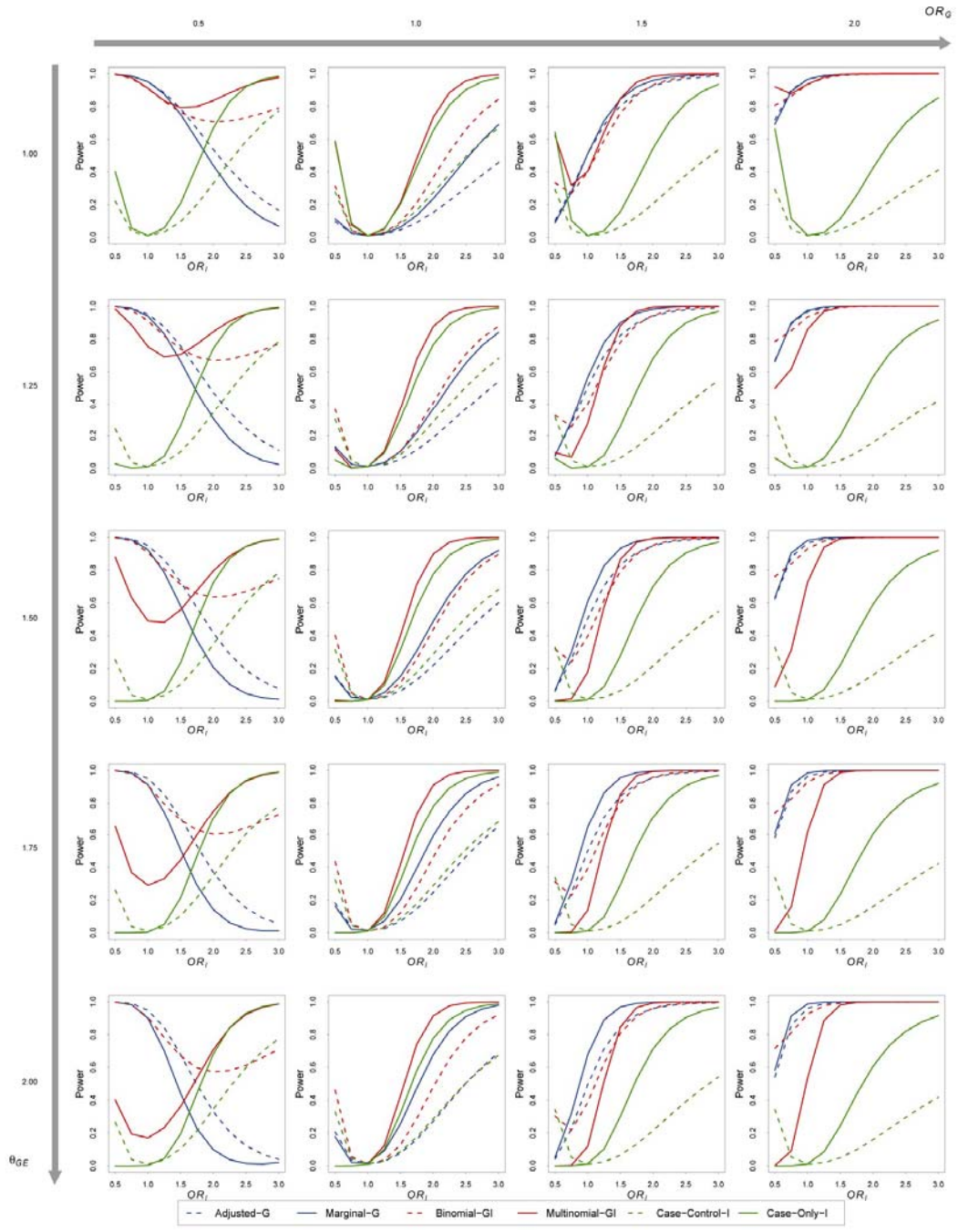
Fixed parameters: environmental exposure frequency, f_E=0.3; dominant genetic model with risk genotypes frequency, f_G=0.3; disease prevalence, f_D=0.1; gene-environment correlation, θ_{GE}=1.0; bilateral nominal type I error-rate of 0.01 for a sample of 500 cases and 500 controls

FIGURE 3. Asymptotic power as a function of the GxE interaction (OR_I) for a range of risk genotypes frequency (f_G) and environmental exposure frequencies (f_E)



Fixed parameters: environmental main effect, OR_E=1.5; dominant genetic main effect OR_G=1.5; disease prevalence, f_D=0.1; gene-environment correlation, θ_{GE}=1.0; bilateral nominal type I error-rate of 0.01 for a sample of 500 cases and 500 controls

FIGURE 4. Asymptotic power as a function of the GxE interaction (OR_I) for a range of genetic main effects (OR_G) and gene-environment correlations (θ_{GE})



Fixed parameters: environmental main effect, OR_E=1.5; environmental exposure frequency, f_E=0.3; dominant genetic model with risk genotypes frequency f_G=0.3; disease prevalence, f_D=0.1; bilateral nominal type I error-rate of 0.01 for a sample of 500 cases and 500 controls

TABLE 1. Summary of available methods for different patterns of exposure and genotype data availability

Tested effect	Data availability in cases		Data availability in controls		Logistic model	Null hypothesis	Degrees of freedom	Notations in Figures and Tables
	Genotypes	Exposures	Genotypes	Exposures				
Genetic only	Yes	No	Yes	No	Binomial	$\beta_{GM} = 0$	1	Marginal-G
	Yes	Yes	Yes	Yes	Binomial	$\beta_{GA} = 0$	1	Adjusted-G
GxE interaction only	Yes	Yes	Yes	Yes	Binomial	$\beta_{ICC} = 0$	1	Case-Control-I
	Yes	Yes	No	No	Binomial	$\beta_{ICO} = 0$	1	Case-Only-I
Genetic and GxE interaction	Yes	Yes	Yes	Yes	Binomial	$\beta_{GCC} = 0$ & $\beta_{ICC} = 0$	2	Binomial-GI
	Yes	No	Yes	Yes	Multinomial	$\beta_{G1} = 0$ & $\beta_{G2} = 0$	2	Multinomial-GI

β_{GM} is the regression coefficient of the genetic effect in a model where only the genetic factor is modeled: logit P(D=1|G)=$\beta_0+\beta_{GM}G$

β_{GA} is the regression coefficient of the genetic effect in a model with adjustment on the exposure: logit P(D=1|G,E)=$\beta_0+\beta_E E+\beta_{GA}G$

β_{GCC} and β_{ICC} are the regression coefficients of the genetic and of the GxE interaction effects in a full model: logit P(D=1|G,E)=$\beta_0+\beta_E E+\beta_{GCC}G$ $+\beta_{ICC}GE$

β_{ICO} is the regression coefficient of the GxE interaction effect in a case-only design: logit P(E=1|G)=$\beta_0+\beta_{ICO}G$

β_{G1} and β_{G2} are the regression coefficients of the multinomial model: logit P(D_M=j|G)=$\beta_{0j}+\beta_{Gj}G$

TABLE 2. Variation of asymptotic power in the presence of gene-environment correlation

OR_G	θ_{GE}	Adjusted-G		Marginal-G		Binomial-GI		Multinomial-GI		Case-Control-I		Case-Only-I	
		OR_I		OR_I		OR_I		OR_I		OR_I		OR_I	
		0.5	2.0	0.5	2.0	0.5	2.0	0.5	2.0	0.5	2.0	0.5	2.0
1.0	1.00	**9.68**	**15.20**	**11.35**	**23.62**	**31.43**	**36.67**	**58.86**	**73.30**	**27.77**	**27.39**	**57.84**	**65.25**
	1.25	+ 2.89	+ 3.51	+ 1.23	+ 12.05	+ 4.86	+ 4.05	– 47.43	+ 14.28	+ 2.18	+ 0.95	– 52.78	+ 11.24
	1.50	+ 5.71	+ 6.62	+ 1.68	+ 22.10	+ 8.80	+ 7.11	– 58.37	+ 16.58	+ 3.60	+ 1.34	– 57.81	+ 12.35
	1.75	+ 8.40	+ 9.39	+ 1.68	+ 30.05	+ 12.03	+ 9.49	– 58.84	+ 17.58	+ 4.51	+ 1.37	– 57.83	+ 12.64
	2.00	+ 10.96	+ 11.86	+ 1.46	+ 36.30	+ 14.71	+ 11.37	– 58.85	+ 18.07	+ 5.05	+ 1.19	– 57.83	+ 12.44
1.5	1.00	**10.30**	**92.60**	**8.88**	**96.04**	**33.34**	**93.22**	**62.39**	**98.57**	**29.23**	**20.60**	**64.26**	**54.49**
	1.25	– 2.39	+ 1.39	– 1.63	+ 2.12	– 0.76	+ 1.29	– 52.79	+ 0.83	+ 2.20	+ 0.71	– 58.19	+ 13.65
	1.50	– 4.08	+ 2.27	– 3.25	+ 2.98	– 1.54	+ 2.08	– 62.21	+ 0.84	+ 3.63	+ 1.01	– 64.22	+ 15.69
	1.75	– 5.31	+ 2.88	– 4.60	+ 3.37	– 2.32	+ 2.59	– 62.38	+ 0.82	+ 4.55	+ 1.05	– 64.25	+ 16.60
	2.00	– 6.22	+ 3.31	– 5.65	+ 3.58	– 3.08	+ 2.95	– 62.38	+ 0.79	+ 5.10	+ 0.93	– 64.25	+ 16.84

Asymptotic powers in absence of gene-environment correlation are shown in bold (θ_{GE}=1). Other values represent variations in percent from this reference value: either increase, when the value is positive or decrease when the value is negative. All values are multiplied by 10^2.

Fixed parameters: environmental main effect, OR_E=1.5; environmental exposure frequency, f_E=0.3; dominant genetic model with risk genotype frequency f_G=0.3; disease prevalence, f_D=0.1; for a sample of 500 cases and 500 controls

OR_G and OR_I are the genetic and the GxE interaction odds-ratio terms of the true penetrance model, respectively. θ_{GE} is the gene-environment correlation.

TABLE 3. Squared bias, variance and coverage probability of the different estimators of the genetic parameter

OR_G	OR_I	OR_E	Squared bias × 10^2				Variance × 10^2				Coverage probability × 10^2			
			β_{GM}	β_{GA}	β_{GCC}	β_{G1}	β_{GM}	β_{GA}	β_{GCC}	β_{G1}	β_{GM}	β_{GA}	β_{GCC}	β_{G1}
0.5	1.0	1.0	0.18	0.17	0.19	0.18	2.30	2.31	3.30	2.88	95.7	95.6	94.6	95.1
		2.0	0.41	0.38	0.20	0.21	2.29	2.35	3.80	3.39	91.3	91.5	93.9	93.9
	2.0	1.0	8.20	7.51	0.22	0.19	2.12	2.14	3.40	2.98	50.0	52.9	94.9	93.9
		2.0	14.44	10.17	0.14	0.13	2.06	2.16	3.97	3.55	23.4	40.9	95.2	96.2
1.0	1.0	1.0	0.00	0.00	0.00	0.00	1.91	1.91	2.74	2.32	95.1	94.9	95.6	94.9
		2.0	0.00	0.00	0.00	0.00	1.91	1.96	3.08	2.67	94.5	94.4	94.9	94.3
	2.0	1.0	5.04	4.39	0.00	0.00	1.84	1.86	2.84	2.43	62.2	67.3	94.8	95.6
		2.0	7.53	4.81	0.01	0.00	1.83	1.91	3.25	2.84	46.9	65.7	94.8	95.0
1.5	1.0	1.0	0.19	0.18	0.21	0.18	1.81	1.81	2.58	2.17	94.0	94.2	92.3	94.2
		2.0	0.26	0.23	0.15	0.13	1.81	1.85	2.88	2.47	93.2	93.6	94.5	95.2
	2.0	1.0	2.42	1.99	0.27	0.26	1.77	1.79	2.69	2.28	79.0	81.5	93.2	93.2
		2.0	3.00	1.57	0.26	0.29	1.77	1.84	3.04	2.63	73.9	84.7	92.7	93.9
2.0	1.0	1.0	0.94	0.92	0.94	0.97	1.77	1.77	2.53	2.12	87.9	88.1	91.0	88.8
		2.0	1.31	1.18	0.58	0.66	1.77	1.81	2.81	2.39	84.8	86.4	92.1	92.2
	2.0	1.0	1.19	0.91	0.67	0.72	1.76	1.77	2.65	2.23	88.7	91.1	92.1	91.8
		2.0	0.92	0.34	0.77	0.78	1.76	1.82	2.95	2.54	88.0	92.5	91.5	90.4

Fixed parameters: environmental exposure frequency, f_E=0.3; dominant genetic model with risk genotype frequency f_G=0.3; disease prevalence, f_D=0.1; gene-environment correlation, θ_{GE}=1; for a sample of 500 cases and 500 controls

OR_G, OR_E and OR_I are the genetic, the environmental and the interaction odds-ratio terms of the true penetrance model, respectively. β_{GM} is the genetic parameter estimator of the marginal model; β_{GA} is the genetic parameter estimator of the adjusted model; β_{GCC} is the genetic parameter estimator in the full model; β_{G1} is the genetic parameter estimator of the multinomial model.

TABLE 4. Squared bias, variance and coverage probability of the different estimators of the GxE interaction parameter

OR_I	OR_G	OR_E	Squared bias × 10^2			Variance × 10^2			Coverage probability × 10^2		
			β_{ICC}	β_{ICO}	β_{G2}/β_{G1}	β_{ICC}	β_{ICO}	β_{G2}/β_{G1}	β_{ICC}	β_{ICO}	β_{G2}/β_{G1}
0.5	1.0	1.0	0.22	0.21	0.21	11.00	6.41	6.41	94.7	94.7	94.7
		2.0	0.71	0.53	0.53	9.53	4.91	4.91	93.0	93.1	93.1
	2.0	1.0	0.83	0.67	0.67	9.51	4.90	4.90	92.6	93.1	93.1
		2.0	0.84	0.58	0.58	8.45	3.84	3.84	93.9	93.1	93.1
1.0	1.0	1.0	0.04	0.00	0.00	9.17	4.60	4.60	95.5	94.2	94.2
		2.0	0.01	0.00	0.00	8.52	3.91	3.91	95.3	93.8	93.8
	2.0	1.0	0.00	0.00	0.00	8.50	3.91	3.91	95.6	95.4	95.4
		2.0	0.78	0.62	0.62	7.94	3.35	3.35	94.3	91.7	91.7
1.5	1.0	1.0	0.17	0.16	0.16	8.65	4.05	4.05	93.8	95.6	95.6
		2.0	0.29	0.48	0.48	8.26	3.65	3.65	92.7	92.8	92.8
	2.0	1.0	0.53	0.63	0.63	8.25	3.66	3.66	93.4	92.2	92.2
		2.0	3.49	3.97	3.97	7.87	3.26	3.26	88.4	81.0	81.0
2.0	1.0	1.0	0.63	0.80	0.80	8.43	3.82	3.82	94.0	93.6	93.6
		2.0	1.63	2.22	2.22	8.18	3.57	3.57	92.6	87.6	87.6
	2.0	1.0	2.44	2.33	2.33	8.16	3.57	3.57	91.7	86.1	86.1
		2.0	9.64	9.86	9.86	7.84	3.25	3.25	79.8	59.9	59.9

Fixed parameters: environmental exposure frequency, f_E=0.3; dominant genetic model with risk genotype frequency f_G=0.3; disease prevalence, f_D=0.1; gene-environment correlation, θ_{GE}=1; for a sample of 500 cases and 500 controls

OR_G, OR_E and OR_I are the genetic, the environmental and the interaction odds-ratio terms of the true penetrance model, respectively. β_{ICC} is the GxE interaction parameter estimator in the full model of the case-control design; β_{ICO} is the GxE interaction parameter estimator in the case-only design; β_{G2}/β_{G1} is the GxE interaction parameter estimator of the multinomial model.

Annexe 3

Kazma R, Dizier MH, Guilloud-Bataille M, Bonaïti-Pellié C, Génin E.
Power Comparison of different methods to detect genetic effects and gene-environment interactions.
BMC Proc 2007, 1(Suppl 1):S74

Power Comparison of Different Methods to Detect Genetic Effects and Gene-Environment Interactions

Rémi KAZMA[1,2§], Marie-Hélène DIZIER[2,1], Michel GUILLOUD-BATAILLE[2,1], Catherine BONAITI-PELLIE[2,1], Emmanuelle GENIN[2,1]

[1]Univ Paris-Sud, UMR-S 535, Villejuif, F-94817, France
[2]INSERM UMR-S 535, Villejuif, F-94817, France

[§]Corresponding author

Email addresses:
RK: kazma@vjf.inserm.fr
MHD: dizier@vjf.inserm.fr
MGB: guilloud@vjf.inserm.fr
CBP: bonaiti@vjf.inserm.fr
EG: genin@vjf.inserm.fr

Fax: (33) 1 45 59 53 31

Abstract

Identifying gene-environment (GxE) interactions has become a crucial issue in the past decades. Different methods have been proposed to test for GxE interactions in the framework of linkage or association testing. Their respective performances have however rarely been compared. Using GAW15 simulated data, we compare here the power of four methods: one based on affected sib-pairs that tests for linkage and interaction (the mean interaction test) and three methods that test for association and/or interaction: a case-control test, a case-only test and a log-linear approach based on case-parent trios. Results show that for the particular model of interaction between tobacco use and locus B simulated here, the mean interaction test has poor power to detect either the genetic effect or the interaction. The association studies, i.e. the log-linear-modeling approach and the case-control method, are more powerful to detect the genetic effect (power of 78% and 95% respectively) and taking into account interaction moderately increases the power (increase of 9% and 3% respectively). The case-only design exhibits a 95% power to detect gene-environment interaction but the type I error rate is increased.

Background

Gene-environment (GxE) interactions are likely to play an important role in multifactorial diseases. The detection of GxE interaction can be of major interest in epidemiological studies to help identify subgroups of the population which are at high risk of disease and at which prevention and screening programs should be targeted. The presence of interaction can conceal environmental and/or genetic effects if not considered in the analysis [1]. On the other hand, taking it into account may either enhance or reduce the power to detect genetic susceptibility factors depending on the parameters inherent to the model underlying disease susceptibility [2,3]. Towards that end, many statistical methods have been developed, in the past decades, either to directly investigate GxE interaction or to enhance detection of genetic factors by taking into account exposure status. They can be classified according to the design followed, the kind of data used and the hypothesis tested [1,4].

The purpose of our work is to compare the power of different methods to detect the effect of locus B and its interaction with history of tobacco use. We used the simulated data (problem 3) of the 15th Genetic Analysis Workshop (GAW15) with knowledge of the "answers" and compared four methods to test for genetic effect and/or GxE interaction. The first method referred to as the mean interaction test (MIT) method [5] tests linkage and GxE interaction among sib-pairs. It is compared to three association testing methods: a log-linear-modeling approach [6] that uses case-parent triads and a case-control design [4], both of which test for the effect of the gene and GxE interaction and finally a case-only design [7] that tests for interaction only.

Methods

One hundred replicates were studied at the disease susceptibility locus B that controls the effect of smoking on rheumatoid arthritis risk. In each replicate, 1500 affected sib-pairs were considered for the MIT, 1500 case-parent trios for the log-linear method, 1500 cases and controls for the case-control design and only the 1500 cases for the case-only test. We also studied smaller sample sizes (500 trios and 750 cases and controls) in order to compare the three association methods for the same number of genotyped individuals. Cases were obtained by considering the first affected case in each sib-pair and controls were the first 1500 control subjects among the 2000 available for each replicate.

Since none of the SNPs close to locus B were in linkage disequilibrium with this locus, we used genotypes of all the individuals at that locus for association tests and the exact identity-by-descent (IBD) provided in the problem 3 "answers" for the linkage test. For the exposure

status, we considered the lifetime smoking status and did not account for the indirectly increased risk through smoke effect on IgM.

Four following methods were compared.

Mean interaction test

The MIT developed by Gauderman and Siegmund [5] is an extension of the mean sharing test [8] to account for GxE interaction. It compares the proportion of alleles shared IBD, π, (which is expected to be equal to 0.5 under the null hypothesis of no linkage) across the three groups of affected sib-pairs differing for the number of exposed sibs (2, 1 or 0). The following regression model is used: $\pi_i = \pi + \beta\ (X_i - X) + \varepsilon_i$ where π is the intercept and β the regression coefficient for the exposure, with X_i the covariate of exposure centred on its mean X. We conducted analysis using the coding scheme consisting of two variables (X_{EE} and X_{EU}) contrasting sib-pairs with 2, 1 or 0 exposed sibs. The null hypothesis of no linkage is tested by the likelihood ratio test (LRT): $T_{\pi\beta} = -2\ [\ln\ \{L(\pi = 0.5, \beta = 0)\} - \ln\ \{L(\pi, \beta)\}]$, which follows a 50:50 mixture distribution of 2 and 3 degrees of freedom (df) χ^2. The alternative hypothesis corresponds to linkage with or without GxE interaction.

In its original presentation, the mean interaction test method allows accounting for GxE interaction in the search for linkage but does not test for GxE interaction. We therefore developed a LRT for GxE interaction: $T_\beta = -2\ [\ln\ \{L(\pi, \beta = 0)\} - \ln\ \{L(\pi, \beta)\}]$. This test follows a 2 df χ^2 distribution.

Log-linear-modeling approach for case-parent triads

Proposed by Umbach and Weinberg (2000) [6], this method consists in comparing the conditional genotype distribution of exposed cases, given parental genotypes, versus that of unexposed cases. Briefly, case-parent triads are divided into 20 categories based on the parental genotypes, the genotype of the case and the exposure status of the case. The expected number of triads can be expressed according to a log-linear model [3,6]. LRT are performed to test for (1) a gene effect ignoring GxE interaction (which follows a 2 df χ^2), (2) a gene effect accounting for GxE interaction (which follows a 4 df χ^2) and (3) a GxE interaction (which follows a 2 df χ^2). Fit of the data with a dominant model is also tested as the true model was dominant.

Case-control design

Case-control designs have been widely used to compare risks of developing a disease according to their genotype and exposure status [4]. Odds-ratios (OR) associated with the exposure, the genotypes and their interaction factors are estimated and tested for significance. Three likelihood ratio tests are performed: a 2df χ^2 test for genetic effect alone, a 4df χ^2 test

for genetic effect accounting for GxE interaction, and 2df χ^2 test of GxE interaction. Fit of the data with a dominant model is tested using a 2df LRT.

Case-only design

Case-only studies [4,7] test the interaction between an exposure and a genotype among case subjects only. This type of design assesses the departure from a multiplicative scale, assuming independence between both factors. To test for the interaction, a 2 df LRT of homogeneity between the genotype distribution in exposed and unexposed cases is performed.

Powers of the different tests were estimated by determining the number of replicates among the 100 replicates that were significant at a nominal 0.05 type I error rate. Type I error rates to test for GxE interaction are estimated on the 7 loci (A, C-H) that are not supposed to interact with lifetime smoking status.

Results

Table 1 gives the mean proportion of alleles shared IBD in the whole sample, and in each of the three sib-pair categories of exposure. Table 2 shows the power of the different tests. We found that MIT has almost no power to detect linkage even when accounting for GxE interaction. This could have been expected given the proportion of alleles shared IBD in the whole sample and in each of the 3 categories based on exposure. Indeed, these proportions are very close to the null expectation of 0.5 (Table 1).

With the log-linear model, the power to detect the gene effect is 78 % and is increased to 87 % when accounting for GxE interaction. There is thus a gain in power to detect the gene effect when accounting for GxE interaction under the simulated model. For the case-control design, the power to detect the gene effect is 95 % and improves to 98 % when accounting for interaction. As shown in Figure 1, the p-values of test accounting for GxE are smaller than those of the test not accounting for GxE for most of the replicates and similar trends (gain or loss of power) are observed between the two methods in 74 % of the replicates.

Concerning the detection of the GxE interaction, we found that the case-only design is by far the most powerful test. It reaches a 95 % power where the case-control design only reaches 69 %, the log-linear approach 53 % and the linkage test (MIT) 12 %. When constraining the number of genotyped individuals to be the same for the three association methods, the differences in power are even more increased. Figure 2 shows, for the first 25 replicates, the p-values of the GxE interaction test for the log-linear-modeling approach, the case-control and the case-only designs. We observe that it is generally in the same replicates that the different

methods give the most significant results with the highest significance achieved for the case-only method.

Estimates of interaction factors presented in Table 2 do not seem to comply with a dominant model and indeed a dominant model is rejected in 60 % of the replicates with the case-control and in 46 % of the replicates with the log-linear model.

Average type I error rates for the interaction test over the 7 loci were respectively 13 % for the case-only design (ranging from 5 % for locus H to 26 % for locus C), 10 % for the case-control (from 4 % for locus H to 30 % for locus C) and 8 % (ranging from 3 % for loci A and F to 23 % for locus C) for the log-linear model.

Discussion

Under the GxE simulated model presented here, it is more powerful to test for association than to test for linkage. Indeed, the MIT method has extremely poor power to detect the genetic factor either with or without taking GxE interaction into account. This could be explained by the low value of the interaction coefficient used in the simulations. Gauderman and Siegmund [5] actually showed that for an interaction coefficient lesser than 3 (or greater than 1/3), the MIT will not be efficient.

For the association-based approaches, accounting for the environmental factor increases the power to detect the genetic susceptibility factor from 78 % to 87 % for the log-linear method and from 95 % to 98 % for the case-control method. This gain in power is rather limited even though under the simulated model, the gene has an effect only in exposed subjects. This could be linked to the fact that the exposure is relatively frequent in the population as shown by Selinger-Leneman et al. [3].

If one is interested in detecting the interaction, the case-only design is shown to be the most efficient. However, its validity depends on some assumptions, in particular, the independence between both genetic and environmental factors. Type I error rates were actually higher than expected (13 % instead of 5 %). It should be noted however that type I errors estimated for the two other methods were also inflated. This was essentially due to locus C that interacts with sex and might thus indirectly be associated to tobacco exposure. When locus C is excluded, type I error rates were close to expectation with the log-linear model (5 %) and with the case-control (6 %), but were still increased for the case-only design (10 %).

Another point of concern was the model issue. In fact, the true model was dominant but dominance is rejected in the majority of the replicates but less often for the log-linear method than for the case-control. A plausible explanation for this distortion could be the fact that sib-

pairs are ascertained, leading to a modification in expected parental genotype distributions. This is partially corrected for in the log-linear model by the conditioning on the parent genotypes.

All association approaches considered here do not take full advantage of the data since only one of the two sibs is used in each sib-pair. It would be interesting to extend the methods to use the whole sibship while correcting for the dependence between the sibs.

Conclusions

Although this study argues in favour of the use of the case-only design to detect a GxE interaction, it shows that if one is interested in detecting gene effect, accounting for the exposure is not necessary. This of course depends strongly on the underlying model and could probably be linked to the high exposure frequency. It will be of interest to compare again the different methods presented here over a much larger range of models.

References

1. Ottman R: Gene-environment interaction: definition and study designs. Prev Med 1996, 25: 764-770

2. Dizier MH, Selinger-Leneman H, Genin E: Testing linkage and gene x environment interaction: comparison of different affected sib-pair methods. Genet Epidemiol 2003, 25: 73-79

3. Selinger-Leneman H, Genin E, Norris JM, Khlat M: Does accounting for gene-environment (GxE) interaction increase the power to detect the effect of a gene in a multifactorial disease? Genet Epidemiol 2003, 24: 200-207

4. Andrieu N, Goldstein AM: Epidemiologic and genetic approaches in the study of gene-environment interaction: an overview of available methods. Epidemiol Rev 1998, 20: 137-147

5. Gauderman WJ, Siegmund KD: Gene-environment interaction and affected sib pair linkage analysis. Hum Hered 2001, 52: 34-46

6. Umbach DM, Weinberg CR: The use of case-parent triads to study joint effects of genotype and exposure. Am J Hum Genet 2000, 66: 251-261

7. Khoury MJ, Flanders WD: Non-traditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! Am J Epidemiol 1996, 144: 207-213

8. Blackwelder W, Elston R: A comparison of sib-pair linkage tests for disease susceptibility loci. Genet Epidemiol 1985, 2: 85-97

Tables

Table 1 - Proportion of alleles shared IBD in the sib-pairs

Average, standard deviation (SD), minimum (min) and maximum (max) values of the proportion of shared alleles between the 1500 sib-pairs over the 100 replicates are presented. π is the total proportion and π_{UU}, π_{EU}, π_{EE} the proportion in sib-pairs with 0, 1 or 2 exposed sibs, respectively

	π	π_{UU}	π_{EU}	π_{EE}
average	0.502	0.500	0.501	0.503
SD	0.008	0.018	0.018	0.013
min	0.485	0.464	0.455	0.480
max	0.525	0.543	0.543	0.539

Table 2 - Power and estimates of interaction coefficients of the four tests

Power in percent (%) to detect (1) the genetic effect accounting for interaction (G+I), (2) the genetic effect not accounting for interaction (G) and (3) the GxE interaction (I) were obtained over the 100 replicates. Average values of the interaction coefficients estimates for exposure x Bb genotypes (IBb) and exposure x BB genotypes (IBB) are represented with their empirical 95% confidence interval (CI)

	G+I	G	I	I_{Bb}	I_{BB}
mean interaction test	6	8	12	-	-
log-linear-modeling [(a)]	87	78	53	1.33[1.03-1.71]	1.72[1.13-1.83]
case-control [(a)]	98	95	69	1.39 [0.97-1.96]	1.88 [1.08-3.10]
case-only [(a)]	-	-	95	1.39 [1.05-1.72]	1.86 [1.39-2.96]
log-linear-modeling [(b)]	33	23	20	1.35[0.79-2.15]	1.79[0.82-3.68]
case-control [(c)]	79	68	42	1.41 [0.85-2.09]	1.96 [0.99-3.37]

[(a)] Samples of 1500 families were used corresponding to 4500 (1500 triads), 3000 (1500 cases and 1500 controls) and 1500 genotyped individuals for the log-linear-modeling, the case-control and the case-only design, respectively.

[(b)] Samples of 500 triads are considered corresponding to 1500 genotyped individuals.

[(c)] Samples of 750 cases and 750 controls are considered here to limit the number of genotyped individuals to 1500.

Figures

Figure 1 - Difference in p-values of G+I and G tests

Difference is represented for the case-control (red plot) and the log-linear-modeling (blue plot) by ln(pG)-ln(pG+I) reported over the first 25 replicates

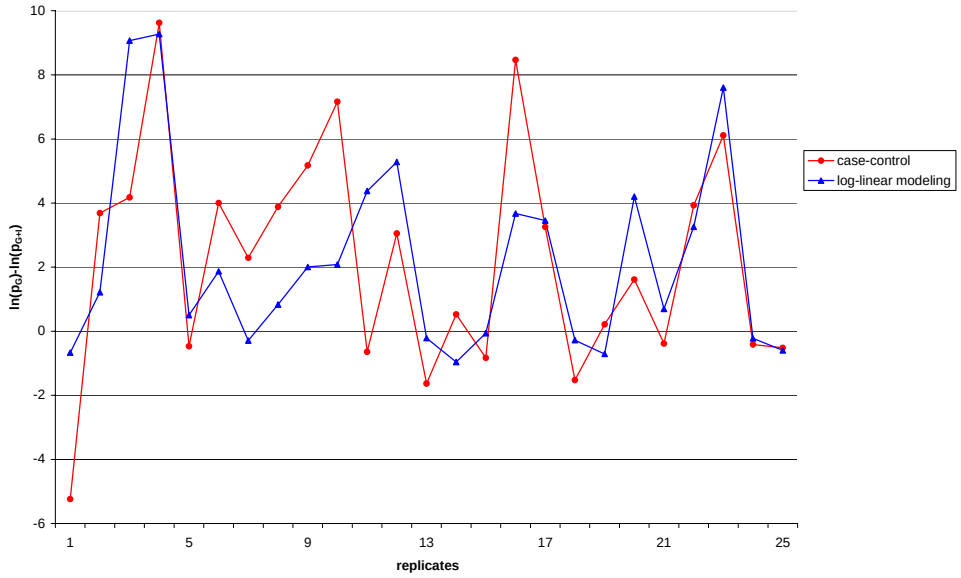


Figure 2 - Comparison of the p-values of the interaction tests

-ln(p) are reported for the case-only design (green plot), the case-control design (red plot) and the log-linear-modeling method (blue plot) over the first 25 replicates