Case-Only Studies of Gene-Environment Interaction: Role of Linkage Disequilibrium and Population Stratification

Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Christian-Albrechts Universität zu Kiel

vorgelegt von

Pankaj Yadav

Kiel, 2015

First referee:	Prof. Dr. Michael Krawczak
Second referee:	Prof. Dr. Tal Dagan
Date of oral examination:	23.11.2015
Approved for publication:	23.11.2015
Dean:	Prof. Dr. Wolfgang J. Duschl

Summary

Studies of gene-environment interactions (G×E) have been considered important owing to their scientific and public health implications. Indeed, many common complex diseases including inflammatory bowel disease (IBD) are presumed to rely on both genetic (G) and environmental (E) risk factors. One major challenge to G×E studies is the insufficient power of traditional epidemiological study designs. A nontraditional approach, the case-only (CO) design, has been proposed as a potentially efficient strategy to assess G×E. Previously, the CO approach was shown to provide better per-sample power compared to other epidemiological study designs including case-control or cohort designs. This approach relies upon two key assumptions, namely that (i) the disease is sufficiently rare in the general population and that (ii) G and E are uncorrelated in the general population. When these assumptions are valid, departures from a multiplicative relative risk model, colloquially known as a 'multiplicative interaction', can be evaluated by testing the association between G and E in cases only. Therefore, in contrast to case-control studies, CO studies require genotype and exposure information from a set of affected individuals alone ('no controls') to track down the underlying G×E. In the past, CO studies of G×E usually followed a candidate (or single-) gene approach, but their utility on genome-wide level remained unexplored.

This thesis addresses two important issues that might potentially affect the genome-wide CO studies of G×E. First, linkage disequilibrium (LD) has been exploited in genome-wide association studies (GWAS) to indirectly infer the causal variants. Using an analogous approach, this thesis explores the role of LD information for the detection of G×E following a CO study. It is shown that single nucleotide polymorphisms (SNPs) in LD with a truly interacting SNP can be used as proxies to indirectly infer the respective G×E signal. Second, population stratification (PS) is a well-known confounder in GWAS, but its impact on the validity of CO studies of G×E has not been studied in detail. Therefore, another major focus of this thesis was to examine the PS scenarios which render CO studies of G×E invalid and to explore alternative means to correct for PS. It was found that the CO approach provided seriously inflated type I error rates under joint stratification by G and E. Moreover, it is shown that genomic control-based methods can be successfully employed to correct for PS in CO studies. Finally, the improved CO approach was applied to a real IBD data set to search for potential gene-smoking interactions in IBD at the genome-wide level. This study identifies 42 SNPs with strong evidence of an interaction with smoking in IBD.

In summary, the issues addressed in this thesis will add to an improved understanding of the genome-wide utility of the CO approach. It was shown that the future research of G×E may safely adopt a CO approach to exploit existing GWAS data sets. At the same time, this work suggests that relevant data resources should generally aim at comprising large numbers of cases for whom genetic and environmental exposure data may be easier to obtain than for controls or patient relatives.

Zusammenfassung

Die Untersuchung von Gen-Umwelt-Interaktionen (G×E) wird aufgrund ihrer Bedeutung für die Grundlagen- und klinische Forschung als wichtig angesehen. Bei Volkskrankheiten einschließlich der entzündlichen Darmerkrankungen (engl. inflammatory bowel disease, IBD) wird vermutet, dass sowohl genetische (G) als auch umweltbedingte (E) Risikofaktoren eine Rolle spielen. Eine große Herausforderung für G×E-Studien ist die unzureichende statistische Power bei klassischen epidemiologischen Studiendesigns. Eine mögliche Strategie G×E zu untersuchen, bietet das nicht-traditionelle Case-only (CO) Design. Es konnte gezeigt werden, dass der CO-Ansatz im Vergleich zu anderen epidemiologischen Studiendesigns, wie Fall-Kontroll- oder Kohortenstudien, eine höhere statistische Power erreichen kann. Dieser Ansatz beruht auf zwei wesentlichen Annahmen: (i) die Krankheit ist in der Allgemeinbevölkerung selten, und (ii) G und E sind in der Allgemeinbevölkerung voneinander unabhängig.

Wenn diese Annahmen gültig sind, können Abweichungen von einem Modell mit multiplikativen relativen Risiken, gemeinhin als 'multiplikative Interaktion' bezeichnet, durch das Testen der Assoziation zwischen G und E ausschließlich in einer Studie an Patienten überprüft werden. Im Gegensatz zu Fall-Kontroll-Studien erfordern CO-Studien die Genotyp- und Expositionsinformationen lediglich von Betroffenen (keine Kontrollen) um die zugrunde liegenden G×E zu detektieren. In der Vergangenheit wurden CO-Studien für G×E in der Regel im Rahmen von Kandidaten- (oder Einzel-) Gen Studien durchgeführt, aber ihr Nutzen auf genomweiter Ebene ist weitestgehend unerforscht.

Die vorliegende Dissertation befasst sich mit zwei wichtigen Fragen, die den Einsatz genomweiter CO-Studien für G×E betreffen. Einerseits wird Kopplungsungleichgewicht (engl. linkage disequilibrium, LD) in genomweiten Assoziationstudien (engl. genome-wide association studies, GWAS) genutzt, um indirekt die kausalen Varianten zu ermitteln. Unter Verwendung eines analogen Ansatzes untersucht diese Arbeit die Rolle der LD-Informationen für den Nachweis von G×E in einer CO-Studie. Es wird gezeigt, dass single nucleotide polymorphisms (SNPs), die mit einem wirklich interagierenden SNP im LD sind, als Proxy dienen können, um das jeweilige G×E Signal zu identifizieren. Andererseits ist Populationsstratifikation (PS) ein bekannter Störfaktor in GWAS, doch ist ihre Auswirkung auf die Gültigkeit von CO-Studien für G×E nicht hinreichend untersucht. Daher ist ein weiterer Schwerpunkt dieser Arbeit, CO-Studien zu G×E bezüglich ihrer Gültigkeit in PS-Szenarien zu prüfen und Methoden zu finden, um für PS zu korrigieren. Es wird gezeigt, dass der CO-Ansatz bei Stratifikation sowohl bezüglich G als auch bezüglich E eine Erhöhung der Typ-I-Fehlerrate mit sich bringt. Darüber hinaus wird gezeigt, dass Verfahren die auf genomischer Kontrolle (engl. genomic control, GC) basieren, erfolgreich eingesetzt werden können um für PS in CO-Studien zu korrigieren. Schließlich wird der verbesserte CO-Ansatz auf Daten aus einer Studie zu IBD aus einer Studie zu angewandt, um potenzielle Interaktionen genetischer Faktoren mit Rauchen auf genomweiter Ebene zu untersuchen. Diese Studie identifiziert 42 SNPs mit starker Evidenz auf eine Interaktion mit Rauchen bei Patienten mit IBD.

Zusammenfassend trägt diese Arbeit dazu bei, die Anwendung des CO-Designs für genomweite Studien zu verbessen. Es wird gezeigt, dass bei zukünftiger Erforschung von G×E CO-Ansätze für bestehende GWAS Datensätze genutzt werden können. Gleichzeitig wird empfohlen, bei relevanten Fragestellungen eine möglichst große Anzahl an Fällen zu rekrutieren, da für diese genetische Daten und Umweltfaktoren leichter zu erheben sind als für Kontrollen oder Verwandte von Patienten.

Contents

Summary **v**

Zusammenfassung **vi**

List of Publications **xi**

List of Figures **xii**

Abbreviations **xiii**

Preface **xv**

1 **Introduction** **1**

1.1 Motivation 1

1.2 Genome-Wide Association Studies, GWAS 2

1.3 Gene-Environment Interactions, G×E 3

1.4 Challenges to Genome-Wide G×E Studies 5

1.4.1 Study Design 6

1.4.2 Linkage Disequilibrium 7

1.4.3 Confounding by Population Stratification 8

2 **Results** **10**

2.1 The Role of Linkage Disequilibrium in Case-Only Studies of Gene-Environment Interactions 10

2.1.1 Summary 10

2.1.2 Publication (i) 12

2.2 Allowing for Population Stratification in Case-Only Studies of Gene-Environment Interaction, using Genomic Control 21

2.2.1 Summary 21

2.2.2 Publication (ii) 22

2.3 Gene-smoking Interaction in Inflammatory Bowel Disease: Meta-analysis of 10 Case-only Studies Comprising over 12,750 Patients 32

2.3.1 Summary 32

2.3.2 Publication (iii) 33

3 **Discussion** **51**

3.1 The Improved CO Approach 51

3.1.1 LD in CO Studies of G×E 51

3.1.2 PS in CO Studies of G×E 52

3.1.3 IBD Data Example 53

3.2 Strengths and Limitations 54

3.3 Conclusion and Outlook 55

Bibliography **57**

Acknowledgements **63**

Declaration **64**

Curriculum Vitae **65**

List of Publications

Published work by the author incorporated into this thesis

Yadav P; Freitag-Wolf S; Lieb W; Krawczak M (2015). The role of linkage disequilibrium in case-only studies of gene-environment interactions. Hum Genet 134(1):89-96

Yadav P; Freitag-Wolf S; Lieb W; Dempfle A; Krawczak M (2015). Allowing for population stratification in case-only studies of gene-environment interaction, using genomic control. Hum Genet 134(10):1117-1125

Unpublished work by the author incorporated into this thesis

Yadav P et al. (in preparation). Gene-smoking interaction in inflammatory bowel disease: Meta-analysis of 10 case-only studies comprising over 12,750 patients.

List of Figures

1.1 Gene-environment interaction 4

1.2 Number of gene-environment studies available in PubMed 6

Abbreviations

CD	Crohn Disease
CO	Case-Only
CRC	Colorectal Cancer
D	Disease status
E	Environmental factor
G	Genetic factor
GC	Genomic Control
G×E	Gene-Environment Interactions
GWAS	Genome-Wide Association Studies
IBD	Inflammatory Bowel Disease
LD	Linkage Disequilibrium
$\log_{10}$	Decadic Logarithm
MAF	Minor Allele Frequency
NAT2	N-acetyltransferase 2
OMIM	Online Mendelian Inheritance in Man
OR	Odds Ratio
PS	Population Stratification
PS_E	Environmental Population Stratification
PS_G	Genetic Population Stratification
PS_{GE}	Genetic and Environmental Population Stratification
SNP	Single Nucleotide Polymorphism
TDT	Transmission Disequilibrium Test
UC	Ulcerative Colitis
UK	United Kingdom
USA	United States of America

Preface

The work described in this thesis focuses on the case-only (CO) design and its utility to assess gene-environment interactions (G×E) on a genome-wide level. Specifically, the following two issues are addressed in this thesis:

i. How does linkage disequilibrium (LD) affect the power to detect an underlying G×E following a CO study on genome-wide level?
ii. How does population stratification (PS) impact upon the validity of CO analysis of genome-wide G×E?

Chapter 1 introduces the key concepts and issues addressed in this thesis. The motivation behind this work is provided at the beginning (section 1.1). The genome-wide association studies are introduced (section 1.2). Further, the concept of G×E is illustrated and the CO design is introduced (section 1.3). The major challenges encountered by genome-wide G×E studies particularly when using a CO approach are described (section 1.4).

The core of this thesis is formed of three publications listed in Chapter 2. There the original publications are incorporated as three different sections. Each publication is preceded with a brief summary of the main findings. The bibliographic data and short summary of the three publications, hereafter denoted as “Publication (i)”, “Publication (ii)” and “Publication (iii)”, respectively, are outlined below.

(i) Yadav, Freitag-Wolf, Lieb and Krawczak (2015)

The Role of Linkage Disequilibrium in Case-Only Studies of Gene-Environment Interactions.

Human Genetics 134(1): 89-96.

Abstract

This paper illustrates the role of LD in genome-wide G×E studies by following a CO approach. The way in which LD impacts upon the chance to detect G×E through the analysis of proxy markers was not studied in much detail before. Therefore, this paper systematically assessed the power to indirectly detect a given G×E through exploiting LD in a CO design. The simulations revealed a strong relationship between LD and detection power that was subsequently validated in a real colorectal cancer data set (section 2.1.2).

(ii) Yadav, Freitag-Wolf, Lieb, Dempfle and Krawczak (2015)

Allowing for Population Stratification in Case-Only Studies of Gene-Environment Interaction, using Genomic Control.

Human Genetics 134(10): 1117-1125.

Abstract

This paper used comprehensive simulation to systematically assess the type I error rate, power and effect size bias of CO studies of G×E in the presence of PS. Three types of PS were considered, namely genetic-only (PS_G), environment-only (PS_E), and joint genetic and environmental stratification (PS_{GE}). The results reveal that the type I error rate of an unadjusted Wald test, appropriate for the CO design, would be close to its nominal level (0.05 in this study) as long as PS involves only one interaction partner (i.e. either PS_G or PS_E). In contrast, if the study population is stratified with respect to both G and E (i.e. if there is PS_{GE}), then the type I error rate is seriously inflated and estimates of the underlying G×E are biased. Further, this paper explored alternative means to allow for PS_{GE}. The results confirm that genomic control-based methods are capable of successfully and efficiently correcting the PS-induced inflation of the type I error rate in CO studies of G×E (section 2.2.2).

(iii) Yadav et al. (in preparation)

Gene-smoking Interaction in Inflammatory Bowel Disease: Meta-analysis of 10 Case-only Studies comprising over 12,750 Patients.

Abstract

This paper investigated the possible interaction between single nucleotide polymorphisms (SNPs) and smoking in relation to Crohn disease (CD) and ulcerative colitis (UC) risk. The analysis covered 10 Immunochip data sets collated by the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC), comprising 12,776 cases (7076 CD, 5700 UC) of known smoking status. A total of 156,499 SNPs were tested for gene-smoking interaction, using the CO design. Three meta-analyses each were performed for CD and UC, considering one of the following smoking-status contrasts: ‘never vs. ever’, ‘never vs. current’, or ‘never vs. former’. This study identified 23 and 19 SNPs with suggestive evidence of gene-smoking interaction in CD and UC, respectively. The majority of these markers (16 SNPs) were located on chromosome 6, thus indicating a potential role of the HLA region in IBD. Noteworthy, several interaction effect differences were observed between CD and UC thereby suggesting a differential role of tobacco smoking in the etiology of IBD (section 2.3.2).

In Chapter 3, the overall findings of this work are discussed. In particular, the LD and PS aspects are discussed (section 3.1) and the strengths and limitations of this work are provided (section 3.2). At the end, this work is concluded and an outlook for conducting future research in genome-wide analysis of G×E is provided (section 3.3). This is followed by a list of references that are used in this thesis and the acknowledgements.

Chapter 1

Introduction

1.1 Motivation

This work is largely motivated by our current enthusiasm for studying the joint effect of genetic (G) and environmental (E) factors to disease risk. Human diseases are broadly classified as ‘simple’ or ‘complex’ contingent upon the number of factors involved in the causal mechanism. Simple genetic diseases by convention follow a simple ‘Mendelian inheritance’ pattern i.e. dominant or recessive. Under a dominant inheritance, only one risk allele (gene variant) is sufficient to cause the disease whilst under recessive inheritance both alleles will be needed. The term ‘penetrance’ is used to describe phenotypes as function of genotypes. Formally, it is the conditional probability of being affected with a disease given a specific genotype. Examples of diseases that follow a simple inheritance pattern include Huntington’s disease (OMIM #143100) and cystic fibrosis (OMIM #219700).

Most diseases, however, do not follow simple inheritance patterns. Examples of such ‘complex’ diseases include cancer, diabetes, heart disease, and inflammatory bowel disease (IBD). These diseases are complex in the sense that their occurrences depend on simultaneous presence of multiple genetic or environmental factors. For instance, not all women who have inherited a risk allele for breast cancer will develop breast cancer. One possible reason for this ‘incomplete penetrance’ (i.e. penetrance <100%) can be an interaction with other G or E risk factors.

In the past, G and E factors have been studied largely as two independent components contributing to diseases. Indeed, several lifestyle habits (e.g. smoking and alcohol abuse) have been reported as risk factors for many common complex diseases [1–3]. However, often it is of interest to investigate the joint effect of both G and E factors on disease risk. Moreover, it has been increasingly accepted that the etiology of most common diseases involves not only discrete G and E causes, but also interactions between the two [4–7]. One well-established example of this is the interaction between smoking and N-acetyltransferase 2 (*NAT2*) for bladder cancer [4, 8]. Hunter (2005) provides a list of rationales for the study of gene-environment interactions (G×E) in the context of medical genetics and epidemiology [9]. Most importantly, accounting for G×E would help to identify individuals at high risk for developing a disease based on both their genetic and exposure profiles. Further, knowledge of

G×E might improve our ability to develop personalized medicines based on an individual’s genetic and life-long exposure information.

In the past few years, large investments have been made for obtaining and storing reliable data on both G and E factors and for elucidating the role of G×E. The National Institutes of Health (NIH) announced ‘Genes, Environment and Health Initiative’ in 2006 which aimed at leveraging the latest genomic technologies and encouraging the development of new environmental measurement methods to study G×E [10]. More recently, the German National Cohort (GNC), aims to integrate the G and E data of disease risk for a prospective cohort study of 200,000 individuals during the next 25-30 years [11]. Moreover, the German Research Foundation has funded several projects on the theme ‘Gene, Environment and Inflammation’ recognizing the importance of G×E studies [12]. These considerations motivate the researchers to find and further develop the available tools and methods that can be applied to investigate the complex G×E in the genome-wide context.

1.2 Genome-Wide Association Studies, GWAS

Lander and Schork (1994) grouped the methods for genetic dissection of complex traits into four categories: linkage analysis, allele-sharing methods, association studies in human populations, and genetic analysis of large samples of crosses in model organisms such as the mouse and rat [13]. In the recent past, genome-wide association studies (GWAS) has gained increasing popularity with the development of economical and advanced genotyping platforms. GWAS serve as an important advancement compared to candidate (or single-) gene studies where only a limited number of selected variants are assayed generally using a smaller sample size. Some researchers consider GWAS as an important landmark beyond family-based linkage studies which suffered from low power for variants with modest effect in complex diseases [14–17].

Basically, GWAS intend to identify alleles that are more frequent in a group of affected individuals than in the unaffected individuals. A given allele with a higher frequency among diseased individuals compared to unaffected individuals is regarded as the variant that increases disease risk. The most common form of genetic variation between individuals is due to single nucleotide polymorphisms (SNPs) [18, 19]. A SNP is a position in the genome where one base is substituted with one of three possible alternative bases. Therefore, in theory four alleles are possible for a single SNP; however most of the SNPs have only two alleles i.e. they are biallelic [20]. A stricter definition requires frequency of the rarer allele (or minor allele frequency, MAF) to be at least 1% in the population to be called a SNP. SNPs are frequently used in GWAS because they are known to be stable and also have reasonable genotyping costs. GWAS, therefore, follow ‘agnostic’ approach with no prior hypotheses and tests for an association between disease status and several thousands of SNPs.

GWAS successfully identified several genetic risk factors in the initial phase and provided optimistic hopes to manifest valuable insights into the genetic architecture of complex diseases. The first successful GWAS was published in 2005 [21]. Since then, GWAS have identified thousands of loci not hitherto detected by earlier methods and harbor susceptibility variants associated with many common human diseases and traits. However, most of the variants identified through GWAS occur outside the coding regions of genes and their causality for the disease is therefore controversial [22, 23]. Moreover, despite meta-analyses of many GWAS, the overall contribution of identified loci to disease variation in the population is frequently <10% [24]. Several explanations for this ‘missing heritability’ have been suggested [22, 25], but a conclusion has not been drawn on the responsible factors. Manolio et al. (2009) quotes:

> “*A consensus is lacking on approaches and priorities for research to examine what has been termed ‘dark matter’ of GWAS—dark matter in the sense that one is sure it exists, can detect its influence, but simply cannot ‘see’ it (yet).*”

Thus, we must address the key question of what factors constitute the remaining heritability of common complex diseases. Some of this ‘dark matter’ could be due to G×E, which remained understudied primarily due to lack of data on environmental exposures. Therefore, attempts are made now to assess the joint effect of G and E risk factors and focus has turned to reliable exposure data collection in conjunction with the genetic data [10, 11].

1.3 Gene-Environment Interactions, G×E

Traditionally, G and E factors have been viewed and studied largely as two independent entities contributing to disease. In reality, however, the interaction between G and E factors can significantly contribute towards the development of disease. Therefore, gene-environment interactions (G×E) have been invoked to partially explain the ‘missing heritability’ of complex diseases that remained unaccounted in GWAS. Loosely, G×E refers to the interplay of G and E factors in causing some phenotype effect [26].

Phenotypes are the observable characteristics and can be binary (e.g. disease status) or quantitative (e.g. body mass index). The way in which genes and environment jointly affect phenotypes can be conceptualized by means of a genetic model. Figure 1.1 schematically illustrates one example of G×E with the help of a genetic model. Here, both G and E components jointly contribute to cause the disease. The risk factor G can have only a small effect on its own to the disease susceptibility and may have been identified previously through

GWAS. In the presence of an interaction between G and E, the risk inferred by G will be modified by the presence or absence of E.

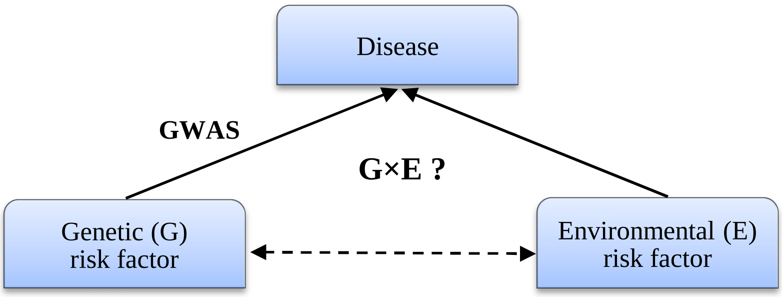


Figure 1.1: Schematic representation of a gene-environment interaction model, showing how genetic (G) and environmental (E) risk factors can relate to a disease. The genetic risk factor G may refer to single nucleotide polymorphism (SNP) identified previously through genome-wide association studies (GWAS).

The term ‘interaction’ is interpreted with different meanings across the scientific disciplines and several tailored, sometimes conflicting, definitions are provided. Two major categories of interaction frequently cited in the literature are, namely, ‘biological interaction’ and ‘statistical interaction’. In the biological interaction, one or more genetic or environmental factors are simultaneously involved in the same causal mechanism in a given individual [27, 28]. In contrast, statistical interaction is equivalent to the ‘moderation’ in regression analysis which means that the statistical association between an outcome of interest and a particular influence factor depends upon the value taken by another factor. Contingent upon the regression scale used to measure the association (mostly log or logit for binary outcomes, such as a disease status), statistical interaction is tantamount to the lack of additivity on that particular scale [29–32]. Statistical interactions can be further classified as ‘quantitative’ or ‘qualitative’. In quantitative interaction, the effects of one factor go in the same direction at different levels of the other factor, but differ in magnitude. On the other hand, in qualitative interaction: the effects go in opposite directions; there is an increased effect only in the presence of both the factors; the effect of one factor is present at only one level of the other factor. Additional details on these definitions can be found in the literature [33, 34]. Most importantly, the presence or absence of statistical interaction between two factors depends to a large extent on the scale chosen to measure the effects. This work follows the convention by which a multiplicative model is considered where the relative risk of disease in individuals carrying both the risk factors is the product of the relative risk of each factor separately.

Conceptually, any departure from this multiplicative model can be interpreted as a measure of interaction; colloquially known as a ‘multiplicative interaction’ (see section 2.1.2 for details).

Previously, the case-only (CO) design has been proposed as a potentially more efficient strategy to detect G×E [35]. Here, G and E information from cases alone (‘no controls’) is used to assess the level of underlying interaction. G×E can be estimated in a CO study if two key assumptions are met, namely that (i) the disease is sufficiently rare in the general population and that (ii) G and E are uncorrelated in the general population. This implies that the CO approach is mathematically valid as long as G and E can be assumed to be statistically independent in controls, which would be the case for any rare disease if G and E are independent at the population level. In this case, G×E approximately equals the odds ratio (OR) of the association between G and E in cases. Preference of CO over other designs has been recommended by many scholars if the primary goal of a study is G×E detection [36–38]. Moreover, a recent meta-analysis suggests that CO studies of G×E are hardly biased because they have consistently yielded the same results as traditional case-control studies [39]. Finally, CO studies trade on the fact that cases are usually much easier to recruit and characterize in terms of their phenotype and environmental exposure than controls, particularly in retrospective studies.

1.4 Challenges to Genome-Wide G×E Studies

Genome-wide G×E are worth studying in order to deeply understand the genetic architecture of complex human diseases. In the past, G×E studies followed a candidate (single-) gene approach which often begins with a well-known association with an E factor and eventually explore genes in pathways that are known to metabolize this factor. These studies have provided only limited information so far on the number and size of G×E effects expected to truly exist in human populations. In addition, the candidate G×E studies suffers from many of the same problems that plagued the candidate genes studies for marginal effects, including small sample sizes and multiple testing problem. Moreover, G×E effects are difficult to be replicated mainly due to variation in exposure measurement protocols across studies, differences in the scale of reported G×E effects, and differential distribution of exposures across studies. The number of genome-wide G×E studies reported in literature is far less compared with the total number of G×E studies published (Figure 1.2). This low number is partly explained by the fact that genome-wide G×E studies face additional challenges, including insufficient power of available study designs, confounding by population stratification, and ambiguous role of linkage disequilibrium in G×E studies. The major challenges currently encountered in genome-wide G×E studies are described in the next sections.

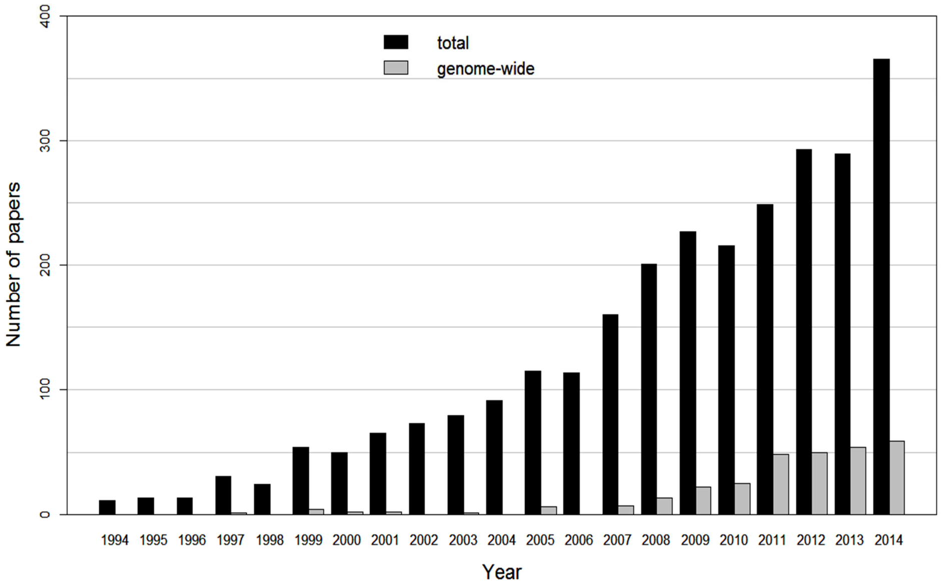


Figure 1.2: Number of papers in PubMed with ('gene–environment' or 'gene-by-environment' or 'gene × environment' or 'G × E') and 'interaction' in the title or abstract (in blue). Furthermore, the number of papers is shown which additionally to the previous search term also contain ('genome-wide' or 'genomewide') in the title or abstract (in grey). It should be noted that this search only retrieves 'potential' G×E studies and that the real numbers of G×E studies are probably even lower than the reported counts (adapted from [26]).

1.4.1 Study Design

Despite the fact that data allow for the investigation of complex G×E, challenge is to find whether the available study designs and methods can be applied in the genome-wide context. Selection of an optimal study design is crucial for conducting successful G×E studies like any other epidemiologic study. Thomas (2010) compiled a list of traditional as well as non-traditional study designs used in the context of G×E studies [34]. Conventionally, epidemiological study designs are either family-based or population-based. The typical family-based designs include case-parent triad (i.e. parents and their affected offspring) and sib-pair designs (i.e. case and siblings). The popular population-based designs include prospective cohorts and case-control studies.

In principle, most of the standard epidemiologic study designs can be utilized to search for G×E, but their performance is driven by several factors including disease prevalence, inheritance mode, and underlying interaction effect sizes [26, 40]. Many researchers preferred family-based designs for G×E studies because they require weaker assumptions on distribution of G and E factors and can be more efficient for rare diseases [41]. However, family-based designs might not be the best choice particularly for late-onset diseases with E component - rendering it difficult to collect data from biological relatives.

Among the population-based designs, cohort studies have long been recommended for G×E studies [42, 43]. However, cohort studies are often expensive and time consuming. Cohort studies of rare diseases would require unrealizably large sample sizes or long follow-up (especially for diseases with late onset). Because of these potential drawbacks, several authors adopted the classical case-control design to study G×E and G×G [5, 44]. Even though case-control studies offer a better compromise between cost and efficiency, they also suffer from a requirement for large samples and biased selection of controls.

At present, innovative and powerful epidemiological designs are needed in order to completely understand the role of genes and environment in causing complex diseases. A non-traditional, the CO design, has been proposed to overcome these issues and for efficiently assessing interactions (section 1.3). In the past, the CO approach has been utilized mainly following candidate (single-) gene studies. Challenges still remain as how best to present and analyze genome-wide G×E using the CO design.

1.4.2 Linkage Disequilibrium

In genetics, linkage disequilibrium (LD) is defined as the non-random association between the alleles of two or more genetic loci in a given population [45]. Most often, LD is a consequence of (but not equivalent to) physical linkage between genetic loci. The theory of LD is well developed in population genetics and it has been widely exploited to provide insight into past evolutionary and demographic events and as the basis for mapping genes in humans and in other species as well [46]. Slatkin (2008) stated that:

> *“LD is of importance in evolutionary biology and human genetics because so many factors affect it and are affected by it.”*

The persistence of strong LD between a mutant allele and the loci closely linked to it has many practical implications. One successful example is the LD mapping of a disease-associated allele based on the slow decay of LD with closely linked markers [47]. The same idea underlies association mapping of complex diseases. In the past, fine-scale pattern of LD in humans confirmed that the human genome is comprised of haplotype blocks within which most SNPs are in high LD [48]. While some variants may be functional, others may simply be markers for undiscovered genetic variants. These high levels of LD among SNPs are presumed to be true for alleles that increase the diseases risk. GWAS is based on the premise that a causal genetic variant is located on a haplotype, and therefore a marker allele in LD with the causal variant should show (by proxy) an association with a trait of interest. Missing genotype information restricts the identification of the actual causal variant that contributes to disease manifestation [14]. Therefore, GWAS take advantage of LD to be able to identify indirect associations.

However, unlike GWAS, the role of LD in G×E studies has not been determined yet. In G×E studies, where a second, non-genetic main effect (E) plays an important role, it is by no means clear how the peculiarities of the genetic effect impact upon the power to detect the interaction. In other words, in G×E studies there is one link more to the chain that relates the proxy SNP to the epidemiological effect of interest - which is interaction, not disease association. It is difficult to draw any conclusion on whether proxies can efficiently detect a true G×E in the absence of truly interacting variant. Therefore, a systematic analysis is required how the level of LD between an interacting marker and a proxy marker determines the chance to detect G×E interaction of the former through studying the latter. Publication (i) illustrates the role of LD in G×E studies by following CO approach (section 2.1.2).

1.4.3 Confounding by Population Stratification

Confounding is caused by an extraneous variable that simultaneously correlate with a risk factor and the outcome of interest, thus creating a spurious association between risk factor and the outcome of interest. In the presence of such variables it becomes difficult to ensure that any observed association between the risk factor and outcome of interest reflects a causal effect. Confounding by ethnicity or population stratification (PS) is extensively studied in the context of GWAS and several approaches are available and commonly in use to tackle this issue [49–52].

As with the genotype-phenotype association studies themselves, hidden PS can impact the validity of G×E studies using a CO design. The aspects of this problem have been subject to little research to date. In principle, three types of PS can possibly occur in G×E studies, namely genetic-only (PS_G), environment-only (PS_E), and joint genetic and environmental stratification (PS_{GE}). Concerns about the widespread of PS or the bias it may induce in epidemiologic studies of G×E or G×G have been raised before. Wang (2006) studied, through computer simulations, the impact of PS bias on G×E in a case-control study [53]. They suggested that the interaction bias is small to nonexistent. Wang and Lee 2008, in contrast, observed a huge interaction bias due to PS in CO studies [54]. The likely reason for this contrast can be that the exposed and the unexposed subgroups of a case-control study suffer from a similar PS bias. Therefore, the bias in a G×E interaction and the ratio of the main genetic effects between the exposed and the unexposed cancel each other out. Conversely, CO situation does not have two subgroups to cancel each other out.

In the context of CO, PS can induce a dependency between G and E and thus need to be controlled to ensure that any G and E association among cases reflects their interaction in causing disease. As yet, the role of PS in G×E studies using a CO design has not been explored in much detail. Most importantly, a systematic analysis is still lacking of how the various possible types of PS (i.e. PS_G, PS_E, and PS_{GE}) may affect the validity and power of such studies. This issue is addressed in Publication (ii) using extensive computer simulations (section 2.2.2).

Chapter 2

Results

2.1 The Role of Linkage Disequilibrium in Case-Only Studies of Gene-Environment Interactions

2.1.1 Summary

This paper systematically investigated, through computer simulations and a real data example, the effect of LD on the power to detect underlying G×E effect γ following CO approach. LD is often noticed in physically linked genetic variants. Events such as natural selection, non-random mating and population structure can also influence the level of LD. It is well-known that most SNPs act as proxies for causal mutations in genotype-phenotype association studies. Indeed, most strongly associated variant at a locus identified through GWAS is presumed to be in LD with the causal variant and is often used for follow-up studies. Therefore, GWAS benefits from strong amount of LD distributed throughout the genome. This paper determined the impact of LD between proxy and interacting SNPs on the utility and validity of the CO design.

The preliminary analyses for this task relied on simulated datasets. SNPs in LD with a causal SNP were simulated and utilized to estimate the true G×E. Simulations were performed in a two-tiered fashion. In the first step, only pairs of SNPs (i.e. the interacting and a proxy SNP) with systematically varying LD were simulated. In the second step, a simulated population sample of haplotypes was obtained from HapMap so as to mimic a realistic LD pattern. In both steps, the two haplotypes of an individual were drawn at random with replacement from the respective haplotype pool and their disease status was assigned using a logistic risk model. This study employed either an additive model (AM) or a dominance model (DM) of the genotype–phenotype relationship to define genotype G and to assign a disease status to an individual. Only affected individuals were retained for further analyses.

The main finding of this paper is that the level of LD significantly affects the power of SNPs to detect the interaction. Specifically, the power to detect G×E is proportional to the amount of LD surrounding that locus. High amount of LD around causative loci results in SNPs that effectively tag the functional loci and allow the signal to be detected. Further, it was found that power under an AM was greater that under a DM irrespective of whether the interaction effect was antagonistic (i.e.

$\gamma<0$) or synergistic (i.e. $\gamma>0$). An analysis of a real colorectal cancer data set in relation to smoking suggested that even SNPs in low LD ($0.3 < r^2 < 0.4$) with a known interacting SNP (rs9877596) may interact at a nominally significant level themselves.

In summary, results from both the simulated and a colorectal cancer datasets show that the screened SNPs can be used as proxies to indirectly infer the underlying G×E.

2.1.2 Publication (i)

Yadav, Freitag-Wolf, Lieb, Dempfle and Krawczak (2015)

Human Genetics 134(1): 89-96.

Hum Genet (2015) 134:89–96
DOI 10.1007/s00439-014-1497-2

ORIGINAL INVESTIGATION

The role of linkage disequilibrium in case-only studies of gene–environment interactions

Pankaj Yadav · Sandra Freitag-Wolf · Wolfgang Lieb · Michael Krawczak

Received: 28 July 2014 / Accepted: 3 October 2014 / Published online: 11 October 2014

Abstract Gene–environment ($G \times E$) interactions have been invoked to account, at least in part, for the gap between the known heritability of common human diseases and the phenotypic variation hitherto explained by genetic variants. Noteworthy in this context, a case-only (CO) design has been proposed in the past as a means to detect $G \times E$ interactions possibly more efficiently than by using classical case–control and cohort designs. So far, however, most CO studies have followed a candidate (or single) gene approach, and the genome-wide utility of the CO design is still more or less unknown. In particular, the way in which linkage disequilibrium (LD) impacts upon the chance to detect $G \times E$ interaction through the analysis of proxy markers has not been studied in much detail before. Therefore, we systematically assessed the power to indirectly detect a given $G \times E$ interaction through exploiting LD in a CO design. Our simulations revealed a strong relationship between LD and detection power that was subsequently validated in a real colorectal cancer data set.

Introduction

Most common, non-traumatic human diseases are thought to be caused by multiple factors (Hunter 2005; Manolio et al. 2009), rendering the presence of gene–gene and gene–environment interactions a likely scenario for these conditions. However, the term ‘interaction’ has different, sometimes conflicting, definitions in the scientific literature, and the existence of an inextricable link between biological and statistical interaction has even been refuted by some scholars (Siemiatycki and Thomas 1981; Phillips 1998; Cordell 2002; Wang et al. 2010; Thomas 2010). Here, we follow the convention by which biological interaction implies that one or more genetic (G) or environmental (E) factors are concurrently causal in a given individual (Rothman et al. 2008; Yang and Khoury 1997). Statistical interaction, in contrast, is equivalent to ‘moderation’ in regression analysis which means that the statistical association between an outcome of interest and a particular influence factor depends upon the value taken by another factor. Contingent upon the regression scale used to measure the association (mostly log or logit for binary outcomes, such as a disease status), statistical interaction is tantamount to the lack of additivity on that particular scale (Bhattacharjee et al. 2010; Greenland 2009; Siemiatycki and Thomas 1981; Thompson 1991). In our study, we will focus exclusively upon statistical interactions between G and E (henceforth referred to as $G \times E$) with regard to odds ratios (ORs). For ORs, interaction means that there is a lack of additivity on the logit scale, i.e. the association in question is not multiplicative.

So far, most $G \times E$ interaction studies have followed a candidate (or single)-gene approach (Franks 2011; Vercelli 2010; Wu et al. 2011). However, such studies can only detect some of the interactions present, not the least because candidate genes for disease causation are not necessarily good candidates for $G \times E$ interaction because they have often resurfaced in different populations with different environmental exposure profiles. On the other hand, genome-wide $G \times E$ studies have only just begun to appear in the scientific

P. Yadav · S. Freitag-Wolf · M. Krawczak (✉)
Institute of Medical Informatics and Statistics, Christian-Albrechts University Kiel, Brunswiker Straße 10,
24105 Kiel, Germany
e-mail: krawczak@medinfo.uni-kiel.de

W. Lieb
Institute of Epidemiology, Christian-Albrechts University Kiel,
Niemannsweg 11, 24105 Kiel, Germany

literature (Ege et al. 2011; Hamza et al. 2011). Irrespective of their actual scope (genetic or genomic), the most straightforward design for $G \times E$ analyses would be a case–control or a cohort setting where ORs or relative risks, respectively, can be estimated and compared between genetic strata. However, a potentially more efficient strategy to study $G \times E$ interactions is by the adoption of a case-only (CO) design. Here, G and E information from cases alone is used to assess the level of interaction present (Piegorsch et al. 1994). The CO approach is mathematically valid as long as G and E can be assumed to be statistically independent in controls, which would be the case for any rare disease if G and E are independent at the population level.

Preference of CO over other designs has been recommended by many scholars if the primary goal of a study is $G \times E$ detection (Yang et al. 1997; Gauderman 2002; Kraft et al. 2007). Moreover, a recent meta-analysis suggests that CO studies of $G \times E$ are hardly biased because they have consistently yielded the same results as traditional case–control studies (Dennis et al. 2011). Finally, CO studies trade on the fact that cases are usually much easier to recruit and characterize in terms of their phenotype and environmental exposure than controls, particularly in retrospective studies.

As yet, the genome-wide utility of the CO design has not been studied in much detail. In particular, a systematic analysis is still lacking of how the level of linkage disequilibrium (LD) between an interacting marker and a proxy marker determines the chance to detect $G \times E$ interaction of the former through studying the latter. In genetics, LD is defined as the non-random association in a given population between the alleles of two or more genetic loci (Lewontin and Kojima 1960). Most often, LD is a consequence of (but not equivalent to) physical linkage between genetic loci. In fact, genome-wide association studies (GWAS) systematically exploit LD to identify disease genes indirectly by relying upon the premise that some marker alleles in LD with a causal mutation should be associated with the disease of interest as well. Moreover, following a Bayesian argument, the most strongly disease-associated marker loci identified in a GWAS usually become the focus of follow-up studies. The success of past GWAS strongly suggests that the same strategy may also sensibly be followed to investigate $G \times E$ interactions at a genome-wide level. To assess the validity of this presumption, we studied the 'rate-limiting' role of LD in CO studies geared to uncover and quantify $G \times E$ interactions.

Materials and methods

Let SNP* denote a disease-related genetic variant that interacts with an environmental exposure E (Fig. 1). If nearby variants, say SNP1 and SNP2, are in LD with SNP* then these loci may also interact with E, thereby potentially allowing inference of the underlying $G \times E$ interaction even if SNP* is not typed itself. Our goal is to investigate how the actual level of LD between SNP* and SNP1 or SNP2 influences the power to detect the original $G \times E$ interaction.

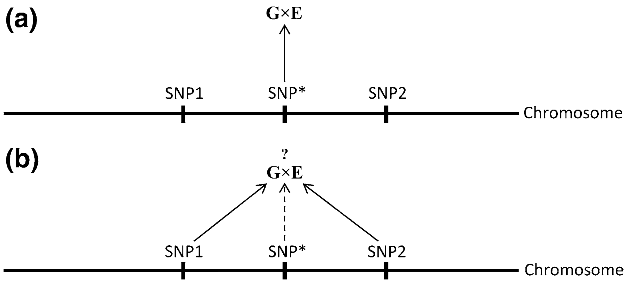


Fig. 1 Detection of $G \times E$ interaction by the analysis of proxy SNPs. **a** Disease-related SNP* is involved in a $G \times E$ interaction, SNP1 and SNP2 are in LD with SNP*. **b** Even if SNP* is not typed, analysis of the proxy markers may reveal the nearby $G \times E$ interaction

Estimating $G \times E$ interaction in CO studies

Let $G \times E_{\mathrm{OR}}$ denote a statistical gene–environment interaction on the logit scale, implying that genetic and environmental ORs do not multiply. Let D be the (binary) disease status. For the sake of simplicity, we assume that G and E are also binary. Conceptually, a lack of interaction between G and E would mean that the disease odds ratio $\mathrm{OR}_{\mathrm{GE}}$ for carriers of both the genetic and the environmental risk factor (i.e. $G = 1$, $E = 1$), relative to individuals carrying neither factor (i.e. $G = 0$, $E = 0$), equals

$$\mathrm{OR}_{\mathrm{GE}} = \mathrm{OR}_G \cdot \mathrm{OR}_E.$$

Here, OR_G denotes the disease OR of $(G = 1, E = 0)$ relative to $(G = 0, E = 0)$, and OR_E is defined analogously. With these conventions, the $G \times E$ interaction term $G \times E_{\mathrm{OR}}$ is defined as the extent to which the true joint effect of G and E ($\mathrm{OR}_{\mathrm{GE}}$) differs from the product of the two individual effects (OR_G and OR_E), i.e.

$$G \times E_{\mathrm{OR}} = \frac{\mathrm{OR}_{\mathrm{GE}}}{\mathrm{OR}_G \cdot \mathrm{OR}_E}. \qquad (1)$$

Under a logistic model of disease risk π_{D} as a function of G and E, i.e.

$$\mathrm{logit}(\pi_{\mathrm{D}}) = \beta_0 + \beta_G \cdot G + \beta_E \cdot E + \gamma \cdot G \cdot E\,, \qquad (2)$$

it follows that $\mathrm{OR}_G = \exp(\beta_G)$, $\mathrm{OR}_E = \exp(\beta_E)$ and $\mathrm{OR}_{\mathrm{GE}} = \exp(\beta_G + \beta_E + \gamma)$ so that

$$G \times E_{\mathrm{OR}} = \exp(\gamma). \qquad (3)$$

Following Piegorsch et al. (1994), $G \times E_{\mathrm{OR}}$ can be estimated in a CO study if two key assumptions are met, namely that (i) the disease is sufficiently rare in the general population and that (ii) G and E are uncorrelated in the general population. In this case, $G \times E_{\mathrm{OR}}$ approximately equals the OR of the association between G and E in cases, denoted by G–E $\mathrm{OR}_{\mathrm{cases}}$ (see "Appendix" for details), which can be estimated fitting

$$\mathrm{logit}(P(E = 1)) = \varphi_0 + \varphi \cdot G \quad (4)$$

to the CO data. Any model-based estimate of φ would also be an estimate of parameter γ in Eq. (2) if the above-mentioned assumptions underlying the CO design are correct.

Data simulation

We simulated data for subsequent analyses in two-tiered fashion. In the first step, only pairs of SNPs (i.e. the interacting and a proxy SNP) with systematically varying LD were considered. In the second step, a simulated population sample of haplotypes was obtained from HapMap so as to mimic a realistic LD pattern. In both steps, the two haplotypes of an individual were drawn at random with replacement from the respective haplotype pool and their disease status assigned using Eq. (2) with varying parameter combinations. Only affected individuals were retained for further analyses.

In the step 1 simulation, the four haplotype frequencies of the SNP pairs were chosen such that specific r^2 values were obtained. In step 2, a population sample of 1,000 haplotypes was first created using the HapSim tool implemented in R (Montana 2005). These simulations were based upon the 176 chromosome 1 haplotypes (comprising 116,415 SNPs) in HapMap (http://hapmap.ncbi.nlm.nih.gov/) that belong to the 88 Utah residents of northern and western European ancestry (labelled CEU in HapMap). SNPs of poor quality were removed and only SNPs with a minor allele frequency (MAF) $\geq$0.45 were retained to allow efficient simulation of cases. To further speed up our computations, case simulations were confined to the 1,001 most 5′ SNPs retained after filtering. MAFs in the HapSim-derived sample were comparable to those in the original CEU sample (Fig. 2), as were the levels of pairwise LD (data not shown).

For each individual, their two haplotypes were either simulated according to the four predefined haplotype frequencies of a SNP pair or were drawn randomly with replacement from the HapSim-derived population pool. In step 1, the first SNP was always presumed to be the interacting one whereas, in step 2, SNP* was the SNP with strongest 'LD emanation'. Here, LD emanation was defined as the number of SNPs that were in notable LD with SNP* (defined as $r^2 > 0.5$ with $p < 0.01$). The primary genotype of an individual was encoded by the allele dosage (i.e. 0, 1 or 2) of an arbitrarily chosen reference allele. We then employed either an additive model (AM) or a dominance model (DM) of the genotype–phenotype relationship to define genotype G and to assign a disease status to an individual using Eq. (2).

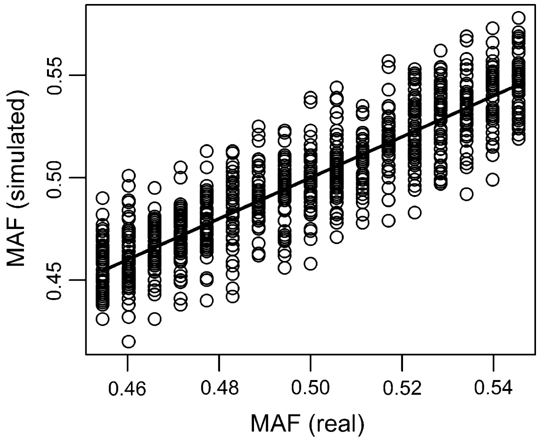


Fig. 2 Commensurability of minor allele frequencies (MAFs) in the original CEU data and the HapSim-derived haplotype pool

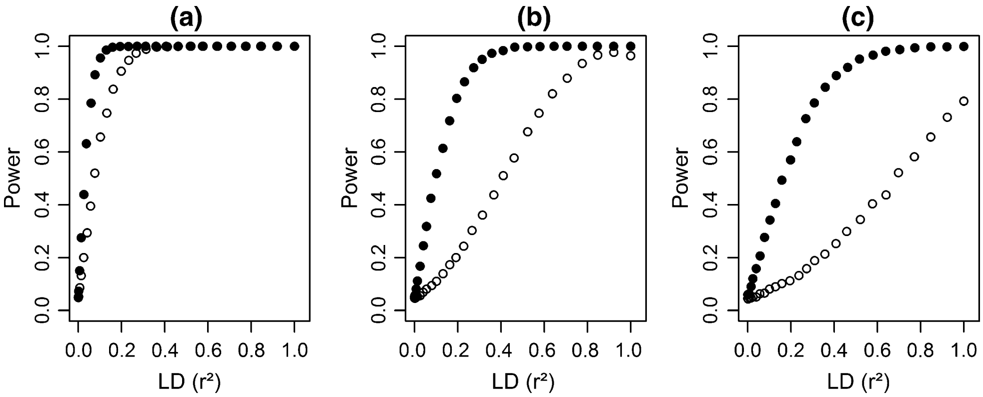
We simulated 5,000 replicates per parameter set of CO data comprising 1,000 patients. Case simulations were carried out with fixed $\beta_0 = -4.6$, $\beta_G = 0.41$, $\beta_E = 0.41$ and varying $\gamma \in \{-2, 1, 2\}$. The values of β_G and β_E were chosen so as to induce a main effect (OR = 1.5) that appears generic for a multifactorial disease. The $G \times E$ interaction was considered to be either moderate, with $\gamma = 1$ ($G \times E_{\mathrm{OR}} = 2.7$), or strong, with $\gamma = \pm 2$ ($G \times E_{\mathrm{OR}} = 7.4$ or 0.14). Baseline risk β_0 was fixed at -4.6 (~1 %) to comply with the rarity assumption (i.e. prevalence <5 %) implicit to the CO design (see "Materials and Methods"). The environmental exposure frequency was set to 10 % in all simulations.

Results

Simulations

Our simulation-based analysis of SNP pairs (step 1) revealed that the power to detect $\gamma \neq 0$ through the analysis of a proxy SNP is an increasing non-linear function of LD (measured by r^2) between the proxy and the interacting SNP* (Fig. 3). Similar results were obtained in these and all other simulations when D' was used instead of r^2 as a measure of LD. With increasing r^2, the detection power increased rapidly and approached unity at r^2 ~ 0.5 for all models and parameter sets considered. As was to be expected, the power attained its maximum for $r^2 = 1$, i.e.

Fig. 3 Power of a proxy SNP to detect $\gamma \neq 0$ as a function of LD (measured by r^2) with interacting SNP*. Cases were simulated at systematically varying LD, using either an AM (*filled circles*) or a DM (*open circles*) of the underlying genotype–phenotype relationship. Interaction parameter γ was set to −2 (**a**), 2 (**b**) or 1 (**c**)



when SNP* and the proxy SNP were perfectly correlated. Greater power emerged under an AM than under a DM irrespective of whether the interaction was antagonistic ($\gamma = -2$) or synergistic ($\gamma = 1$ and 2). Moreover, the power was positively correlated with the absolute value of the interaction effect under both models, and the antagonistic interaction was easier to detect than the synergistic interaction of the same magnitude. This difference is explicable in terms of a synergistic interaction, which exacerbates the main effects, being less 'visible' on the logit scale than an antagonistic interaction, which shifts the argument of the logit function towards the point of maximum slope.

The systematic results from the consideration of SNP pairs were corroborated by simulations using realistic LD patterns based upon HapMap (Fig. 4). Under a DM, the power was higher when the dominant allele of the proxy SNP was associated with the dominant allele of SNP* (indicated by circles in Fig. 4d–f) rather than with the recessive allele (crosses). No such phase effect became apparent under an AM. Moreover, because physical distance is closely related to LD, the power of a proxy SNP to detect $\gamma \neq 0$ at SNP* was also found in step 2 to depend upon inter-marker distance (Fig. 5). Power attained its maximum at SNP* itself and decreased with increasing distance between proxy SNP and SNP*. In some settings, however, particularly for an antagonistic interaction under an AM (Fig. 5a), even rather distal SNPs still provided >80 % power because these markers were in substantial LD with SNP* in the simulated population sample. The fact that the power occasionally fell below 5 % (Fig. 5e, f) implies that the test for $G \times E$ interaction would have been particularly conservative in these cases.

Colorectal cancer data

To further validate our simulation results, we used real data from a recent study by Siegert et al. (2013). Analysing the relationship between SNP genotypes and smoking in a GWAS of colorectal cancer, these authors observed a statistically significant $G \times E$ interaction for SNP rs9877596 (Table 1; for further details, see the original report). The study comprised 1,316 Germans (1,002 controls, 314 cases) originally recruited through the PopGen biobank (Krawczak et al. 2006). Values of r^2 with rs9877596 were calculated with PLINK (Purcell et al. 2007) to identify a total of 568 SNPs in significant LD ($p < 0.05$) with the interacting SNP. Analysis of the CO data yielded a nominally significant $G \times E$ interaction for seven of these SNPs (Fig. 6) thereby illustrating that, in praxis, even proxy SNPs in comparatively low LD with an interacting SNP (e.g. $r^2 \sim 0.4$ for chr3:20573772) may have sufficient power to detect a nearby $G \times E$ interaction.

Discussion

Using computer simulations and real data, we studied the impact of linkage disequilibrium (LD) and physical distance on the power to detect $G \times E$ interactions in case-only (CO) studies of proxy SNPs. To this end, we (i) systematically varied the level of LD between an interacting SNP (labelled SNP*) and a proxy SNP (step 1 of our simulations) and (ii) exploited HapMap to simulate haplotypes with a realistic LD pattern. Our main finding in both instances was that the level of LD present strongly affects the power of a proxy SNP to detect a given $G \times E$ interaction, and that LD and power are more or less proportional to one another as expected. High levels of LD allow a SNP to effectively 'tag' the interacting SNP, thereby extending the $G \times E$ signal to the proxy SNP itself. However, our simulations also revealed that even a distal SNP (>20 Mb away from the interacting SNP) may still have >25 % power to detect the interaction at the level considered here (Fig. 5a–d). Along the same vein, re-analysis of a real colorectal cancer data set in relation to smoking revealed that even SNPs in low LD ($0.3 < r^2 < 0.4$) with a known interacting SNP (rs9877596) may interact at a nominally significant level themselves.

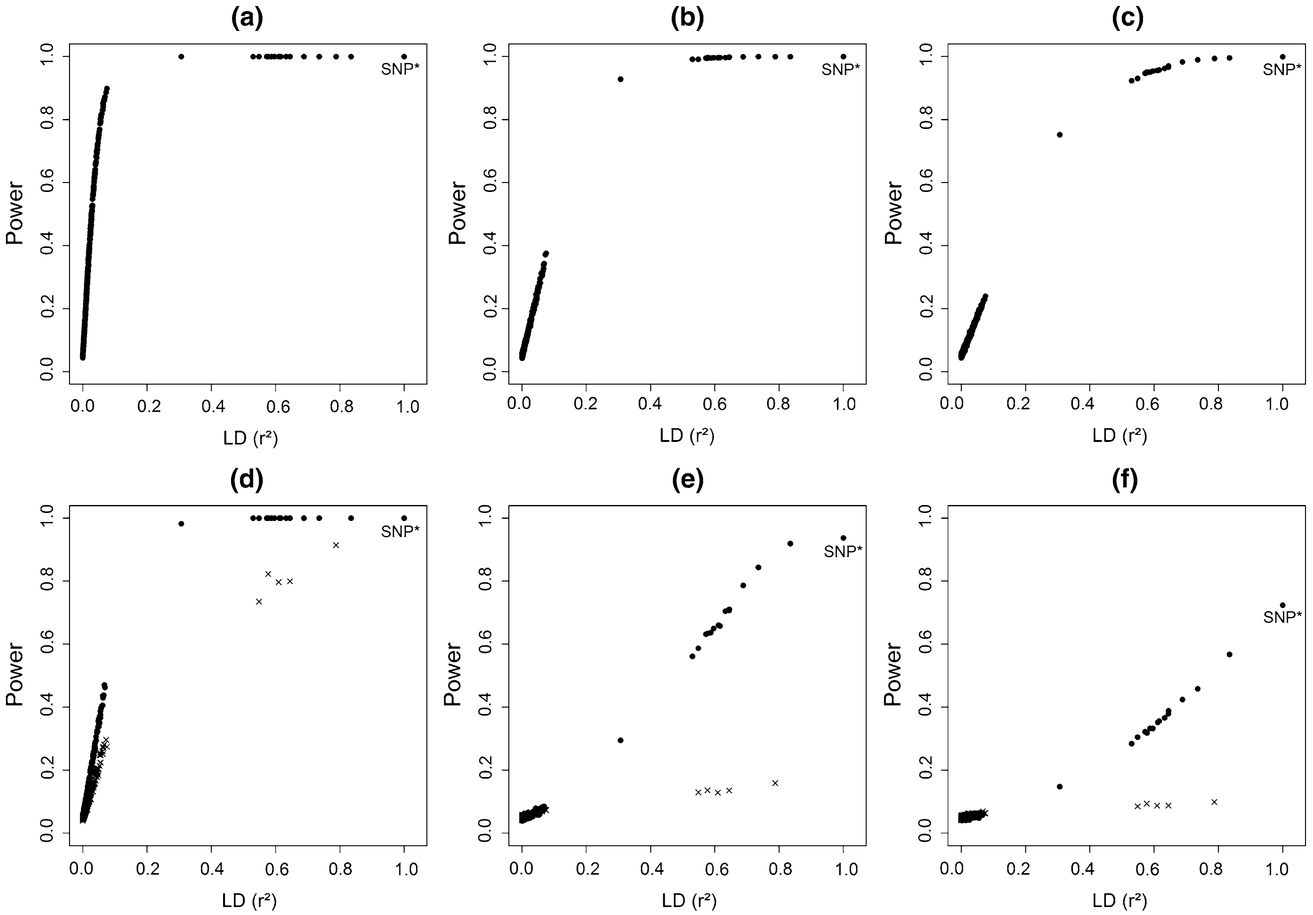


Fig. 4 Power of a proxy SNP to detect $\gamma \neq 0$ as a function of LD (measured by r^2) with interacting SNP*. Cases were simulated adopting a HapMap-derived LD pattern, using either an AM (**a**–**c**) or a DM (**d**–**f**) of the underlying genotype–phenotype relationship. Interaction parameter γ was set to −2 (**a**, **d**), 2 (**b**, **e**) or 1 (**c**, **f**). In *insets* **d**–**f**, proxy SNPs for which the dominant (recessive) allele was associated with the dominant allele of SNP* are indicated by *circles* (*crosses*)

Also meeting prior expectations, the power to detect $G \times E$ was found to be higher when the truly underlying interaction was strong (i.e. $|\gamma| = 2$ compared to $\gamma = 1$). As has been suggested previously, different models of a genotype–phenotype relationship result in different efficiency of association and interaction studies to unravel the corresponding effects (Lettre et al. 2007; Wang and Zhao 2003). In both our simulations, an additive model generally rendered the CO design more powerful than a dominance model (Figs. 3, 4).

Although potentially highly efficient, the advantageousness of CO studies nevertheless must be considered with some caution. For instance, the validity of our conclusions is necessarily limited by the extent to which the implicit assumption of G and E being independent is met. The more this assumption is violated, the more biased would be the resulting interaction estimates (Albert et al. 2001; Saunders et al. 2001). However, independence between G and E seems a sensible assumption in most real-life situations unless a large number of individuals adapt their environmental exposure to their genotype at an interacting SNP, which seems a possible albeit rather unlikely scenario. Another possible limitation is that the measures of interaction obtained from CO analysis reflect a lack of additivity only on the log or logit scale, thus highlighting non-multiplicative relative risks or ORs. In some situations, however, unravelling non-additivity of absolute risks may also be of great scientific interest (Rothman et al. 2008).

In summary, we have shown that SNPs in LD with a truly interacting SNP can be used as proxies to indirectly infer the respective $G \times E$ interaction. This implies that the association paradigm of GWAS may reasonably be extended to the study of $G \times E$ interactions. Furthermore, our results also suggest that future research may adopt the CO approach at genome-wide level so as to allow exploitation of existing GWAS data of cases whose environmental exposure data may be easier to obtain than those of controls.

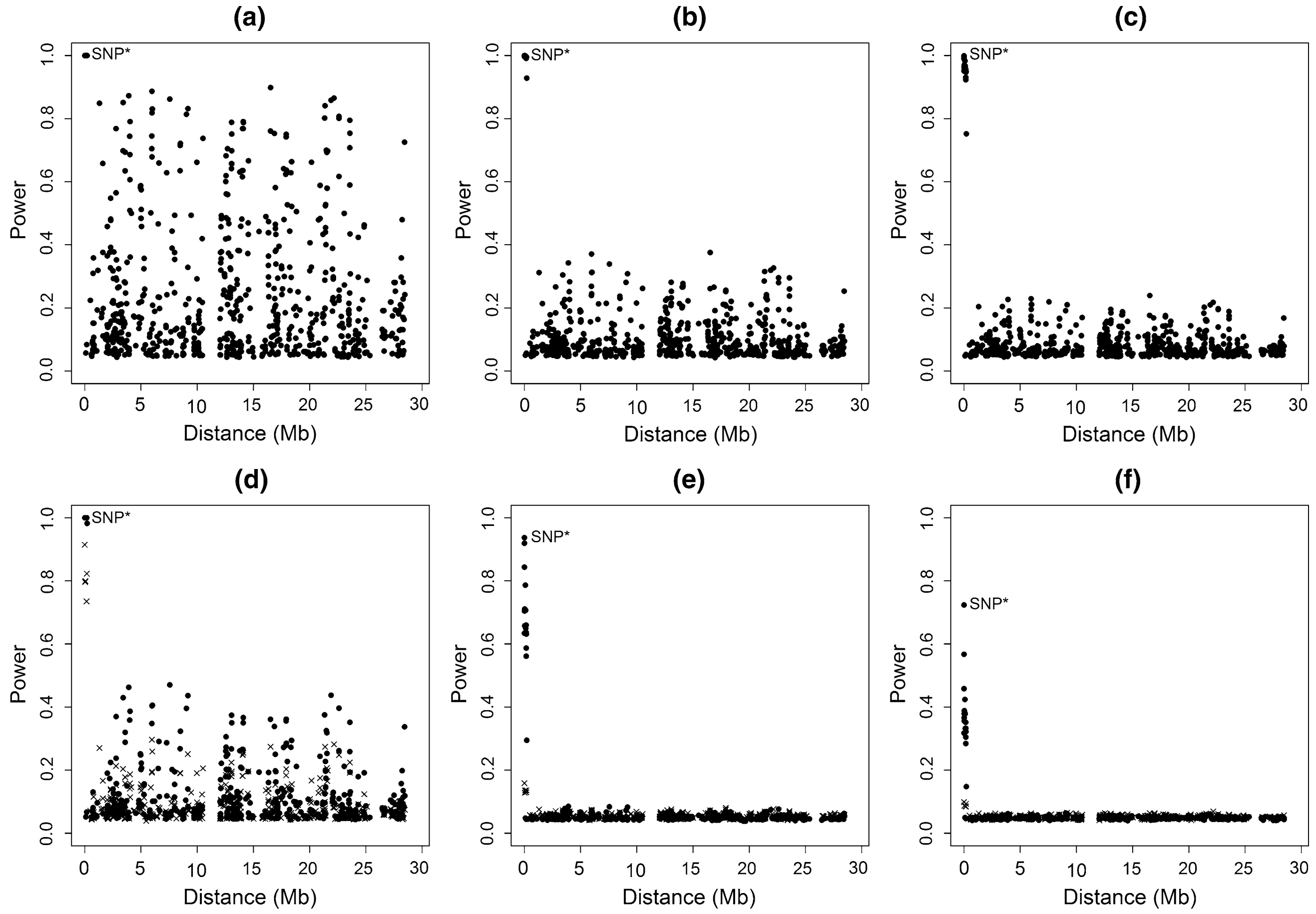


Fig. 5 Power of proxy SNPs to detect $\gamma \neq 0$ as a function of physical distance (in Mb) to interacting SNP*. Cases were simulated adopting a HapMap-derived LD pattern, using either an AM (**a**–**c**) or a DM (**d**–**f**) of the underlying genotype–phenotype relationship. Interaction parameter γ was set to −2 (**a**, **d**), 2 (**b**, **e**) or 1 (**c**, **f**).). In *insets* **d**–**f**, proxy SNPs for which the dominant (recessive) allele was associated with the dominant allele of SNP* are indicated by *circles* (*crosses*)

Table 1 A colorectal cancer-associated SNP interacting with smoking (Siegert et al. 2013)

SNP	Chromosomal region	γ	95 % CI	p value
rs9877596	chr3:20573619… 20576566	−0.74	[−1.09, −0.38]	5.4×10^{-5}

Acknowledgments This study was carried out as part of the Research Training Group 'Genes, Environment and Inflammation', supported by the Deutsche Forschungsgemeinschaft (GRK 1743). Additional funding was provided by Research Area XI ('Epidemiology') of the Excellence Cluster 'Inflammation at Interfaces'. The colorectal cancer dataset was obtained from the PopGen Biobank Kiel, which is supported by the State of Schleswig-Holstein and the German Federal Ministry of Education and Research (grant 01EY1103 to the PopGen 2.0 Network).

Conflict of interest The authors declare that there is no conflict of interest.

Appendix: Estimating the level of $G \times E$ interaction in case-only studies

ORs for the joint effect of G and E and for the marginal effects of G and E are calculated as follows (adapted from Gatto et al. 2004; see "Materials and methods" of the main text for definitions):

Joint effect of G and E			Marginal effect of G (given $E = 0$)			Marginal effect of E (given $G = 0$)		
	$D = 1$	$D = 0$		$D = 1$	$D = 0$		$D = 1$	$D = 0$
$G = 1$, $E = 1$	a_1	b_1	$G = 1$	c_1	d_1	$E = 1$	a_2	b_2
$G = 0$, $E = 0$	c_2	d_2	$G = 0$	c_2	d_2	$E = 0$	c_2	d_2
$\mathrm{OR}_{\mathrm{GE}} = \frac{a_1/b_1}{c_2/d_2}$			$\mathrm{OR}_G = \frac{c_1/d_1}{c_2/d_2}$			$\mathrm{OR}_E = \frac{a_2/b_2}{c_2/d_2}$		

Note that OR_G and OR_E entail the genotype and exposure odds given $E = 0$ and $G = 0$, respectively

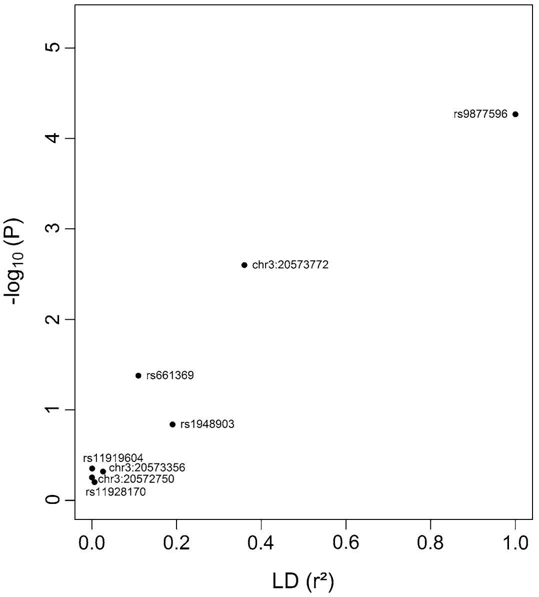


Fig. 6 $G \times E$ interaction in colorectal cancer. Depicted are the $-\log_{10}(p)$ values of a Wald test for the interaction between a given SNP and smoking. SNP rs9877596 was originally reported by Siegert et al. (2013) to show nominally significant $G \times E$

Rearranging the above tables and stratifying the G and E data by disease status allows calculation of the gene–environment OR in both cases and controls (denoted as Term I and Term II below)

Cases ($D = 1$)			Controls ($D = 0$)		
	$G = 1$	$G = 0$		$G = 1$	$G = 0$
$E = 1$	a_1	a_2	$E = 1$	b_1	b_2
$E = 0$	c_1	c_2	$E = 0$	d_1	d_2
$G - E\ \mathrm{OR}_{\mathrm{cases}} = \frac{a_1/a_2}{c_1/c_2}$			$G - E\ \mathrm{OR}_{\mathrm{controls}} = \frac{b_1/b_2}{d_1/d_2}$		
Term 1			Term 2		

Conceptually, multiplicative interaction between G and E ($G \times E_{\mathrm{OR}}$) refers to any deviation of the product of OR_G and OR_E from $\mathrm{OR}_{\mathrm{GE}}$. Therefore, substituting $\mathrm{OR}_{\mathrm{GE}}$, OR_G and OR_G by the respective terms yields

$$G \times E_{\mathrm{OR}} = \frac{\mathrm{OR}_{\mathrm{GE}}}{\mathrm{OR}_G \cdot \mathrm{OR}_E} = \frac{\frac{a_1/b_1}{c_2/d_2}}{\frac{c_1/d_1}{c_2/d_2} \cdot \frac{a_2/b_2}{c_2/d_2}} = \frac{\left.\frac{a_1 c_2}{a_2 c_1}\right\}\text{Term I}}{\left.\frac{b_1 d_2}{b_2 d_1}\right\}\text{Term II}}$$

Under the assumptions that G and E are independent in the general population and that the disease is rare (i.e. that controls are almost representative of the population as a whole), Term II will be equal to 1. Then, $G \times E_{\mathrm{OR}}$ can be estimated by the gene environment OR in cases (see Piegorsch et al. 1994 for further details).

References

Albert PS, Ratnasinghe D, Tangrea J, Wacholder S (2001) Limitations of the case-only design for identifying gene–environment interactions. Am J Epidemiol 154:687–693

Bhattacharjee S, Wang Z, Ciampa J, Kraft P, Chanock S, Chatterjee N (2010) Using principal components of genetic variation for robust and powerful detection of gene–gene interactions in case–control and case-only studies. Am J Hum Genet 86:331–342

Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11:2463–2468

Dennis J, Hawken S, Krewski D, Birkett N, Gheorghe M, Frei J, McKeown-Eyssen G, Little J (2011) Bias in the case-only design applied to studies of gene–environment and gene–gene interaction: a systematic review and meta-analysis. Int J Epidemiol 40:1329–1341

Ege MJ, Strachan DP, Cookson WO, Moffatt MF, Gut I, Lathrop M, Kabesch M, Genuneit J, Buchele G, Sozanska B, Boznanski A, Cullinan P, Horak E, Bieli C, Braun-Fahrlander C, Heederik D, von Mutius E (2011) Gene–environment interaction for childhood asthma and exposure to farming in Central Europe. J Allergy Clin Immunol 127:138–144

Franks PW (2011) Gene × environment interactions in type 2 diabetes. Curr Diab Rep 11:552–561

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 33:1014–1024

Gauderman WJ (2002) Sample size requirements for association studies of gene–gene interaction. Am J Epidemiol 155:478–484

Greenland S (2009) Interactions in epidemiology: relevance, identification, and estimation. Epidemiology 20:14–17

Hamza TH, Chen H, Hill-Burns EM, Rhodes SL, Montimurro J, Kay DM, Tenesa A, Kusel VI, Sheehan P, Eaaswarkhanth M, Yearout D, Samii A, Roberts JW, Agarwal P, Bordelon Y, Park Y, Wang L, Gao J, Vance JM, Kendler KS, Bacanu SA, Scott WK, Ritz B, Nutt J, Factor SA, Zabetian CP, Payami H (2011) Genome-wide gene–environment study identifies glutamate receptor gene GRIN2A as a Parkinson's disease modifier gene via interaction with coffee. PLoS Genet 7:e1002237

Hunter DJ (2005) Gene–environment interactions in human diseases. Nat Rev Genet 6:287–298

Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ (2007) Exploiting gene–environment interaction to detect genetic associations. Hum Hered 63:111–119

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

Lettre G, Lange C, Hirschhorn JN (2007) Genetic model testing and statistical power in population-based association studies of quantitative traits. Genet Epidemiol 31:358–362

Lewontin RC, Kojima K (1960) The evolutionary dynamics of complex polymorphisms. Evolution 14:458–472

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Montana G (2005) HapSim: a simulation tool for generating haplotype data with pre-specified allele frequencies and LD coefficients. Bioinformatics 21:4309–4311

Phillips PC (1998) The language of gene interaction. Genetics 149:1167–1171

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 13:153–162

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575

Rothman K, Greenland S, Lash T (2008) Modern Epidemiology, 3rd edn. Lippincott Williams & Wilkins, Philadephia

Saunders CL, Gooptu C, Bishop DT, Barrett JH (2001) The use of case-only studies for the detection of interactions, and the non-independence of genetic and environmental risk factors for disease (Abstract). Genet Epidemiol 21:174

Siegert S, Hampe J, Schafmayer C, von Schönfels W, Egberts JH, Försti A, Chen B, Lascorz J, Hemminki K, Franke A, Nothnagel M, Nöthlings U, Krawczak M (2013) Genome-wide investigation of gene–environment interactions in colorectal cancer. Hum Genet 132:219–231

Siemiatycki J, Thomas DC (1981) Biological models and statistical interactions: an example from multistage carcinogenesis. Int J Epidemiol 10:383–387

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

Thompson WD (1991) Effect modification and the limits of biological inference from epidemiologic data. J Clin Epidemiol 44:221–232

Vercelli D (2010) Gene–environment interactions in asthma and allergy: the end of the beginning? Curr Opin Allergy Clin Immunol 10:145–148

Wang S, Zhao H (2003) Sample size needed to detect gene–gene interactions using association designs. Am J Epidemiol 158:899–914

Wang X, Elston R, Zhu X (2010) The meaning of interaction. Hum Hered 70:269–277

Wu C, Hu Z, He Z, Jia W, Wang F, Zhou Y, Liu Z, Zhan Q, Liu Y, Yu D, Zhai K, Chang J, Qiao Y, Jin G, Liu Z, Shen Y, Guo C, Fu J, Miao X, Tan W, Shen H, Ke Y, Zeng Y, Wu T, Lin D (2011) Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. Nat Genet 43:679–684

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

Yang Q, Khoury MJ, Flanders WD (1997) Sample size requirements in case-only designs to detect gene–environment interaction. Am J Epidemiol 146:713–720

2.2 Allowing for Population Stratification in Case-Only Studies of Gene-Environment Interaction, using Genomic Control

2.2.1 Summary

This paper illustrates an important issue of the impact of PS upon the validity of CO studies of G×E. Genetic PS (PS_G) could occur due a systematic difference in the allele frequencies between populations. The role of PS_G in GWAS is well-known and several means including genomic control (GC) and principal component analysis (PCA) exist to tackle this issue in GWAS. However, the effect of PS in CO studies of G×E was known only partially, in particular a systematic analysis of this problem was lacking. PS can be a major problem in context of G×E studies because here environmental stratification (PS_E) could also occur in addition to PS_G. Consequently, three types of PS are possible in G×E studies, namely PS_G, PS_E, and PS_{GE}, where PS_{GE} denotes the joint environmental and genetic PS. The purposes of this paper were to determine the PS scenarios under which the validity of the Wald test suitable for detecting G×E in CO studies is affected and to explore alternative means to correct for PS.

The analysis work in this paper relied mainly upon the simulated data sets. The simulation study was carried out in a multi-tiered fashion. In short, CO genotype and exposure data were simulated in the presence of one of three different types of PS (i.e. PS_G, PS_E, and PS_{GE}). To investigate the impact of PS on unadjusted and adjusted Wald test statistics, CO data were simulated in two subpopulations, each with a different parameter setting, and were combined. The risk allele frequency (f_g) and the environmental exposure frequency (f_e) were varied in one subpopulation in order to induce various levels of PS (see *Data simulation* section of the paper).

One main finding in this study was that PS_{GE} can impact the validity of the Wald test suitable for CO studies of G×E and can provide highly inflated type I error rates. Further, this paper demonstrated that the GC paradigm from GWAS can be extended to G×E to rectify this problem. Moreover, it was shown that family-based methods such an extension of Transmission Disequilibrium Test (TDT) as proposed by Schaid (1999) can be robust to PS, but their applicability is contingent up on the availability of family-data.

In summary, the results from this paper imply that future research of G×E may safely adopt GC approach to correct for PS in CO studies of G×E. This indicates that relevant data resources should generally aim at comprising large numbers of cases for whom genetic and environmental exposure data may be easier to obtain than for controls or patient relatives.

2.2.2 Publication (ii)

Yadav, Freitag-Wolf, Lieb, Dempfle and Krawczak (2015)

Human Genetics 134(10): 1117-1125.

Hum Genet
DOI 10.1007/s00439-015-1593-y

ORIGINAL INVESTIGATION

Allowing for population stratification in case-only studies of gene–environment interaction, using genomic control

Pankaj Yadav[1] · Sandra Freitag-Wolf[1] · Wolfgang Lieb[2] · Astrid Dempfle[1] · Michael Krawczak[1]

Received: 26 May 2015 / Accepted: 16 August 2015

Abstract Gene–environment interactions (G × E) have attracted considerable research interest in the past owing to their scientific and public health implications, but powerful statistical methods are required to successfully track down G × E, particularly at a genome-wide level. Previously, a case-only (CO) design has been proposed as a means to identify G × E with greater efficiency than traditional case–control or cohort studies. However, as with genotype–phenotype association studies themselves, hidden population stratification (PS) can impact the validity of G × E studies using a CO design. Since this problem has been subject to little research to date, we used comprehensive simulation to systematically assess the type I error rate, power and effect size bias of CO studies of G × E in the presence of PS. Three types of PS were considered, namely genetic-only (PS_G), environment-only (PS_E), and joint genetic and environmental stratification (PS_{GE}). Our results reveal that the type I error rate of an unadjusted Wald test, appropriate for the CO design, would be close to its nominal level (0.05 in our study) as long as PS involves only one interaction partner (i.e., either PS_G or PS_E). In contrast, if the study population is stratified with respect to both G and E (i.e., if there is PS_{GE}), then the type I error rate is seriously inflated and estimates of the underlying G × E interaction are biased. Comparison of CO to a family-based case–parents design confirmed that the latter is more robust against PS_{GE}, as expected. However, case–parent trios may be particularly unsuitable for G × E studies in view of the fact that they require genotype data from parents and that many diseases with an environmental component are likely to be of late onset. An alternative approach to adjusting for PS is principal component analysis (PCA), which has been widely used for this very purpose in past genome-wide association studies (GWAS). However, resolving genetic PS properly by PCA requires genetic data at the population level, the availability of which would conflict with the basic idea of the CO design. Therefore, we explored three modified Wald test statistics, inspired by the genomic control (GC) approach to GWAS, as an alternative means to allow for PS_{GE}. The modified statistics were benchmarked against a stratified Wald test assuming known population affiliation, which should provide maximum power under PS. Our results confirm that GC is capable of successfully and efficiently correcting the PS-induced inflation of the type I error rate in CO studies of G × E.

Introduction

Nearly all non-traumatic human diseases involve some kind of interaction between genetic and environmental risk factors. In consequence, studying the joint health effects of genes (G) and the environment (E) is at least as important for understanding the etiology of a given disease as unraveling its genetic basis alone. In the past, G × E studies mostly followed a candidate (single-) gene approach (Begg and Zhang 1994; Hwang et al. 1995). For instance, Hwang et al. (1995) reported an interaction with regard to cleft palate between maternal smoking and infant genotype at the transforming growth factor alpha (*TGFA*) locus. Children carrying the rarer C2 allele exhibited a more than sevenfold

✉ Michael Krawczak
krawczak@medinfo.uni-kiel.de

[1] Institute of Medical Informatics and Statistics, Christian-Albrechts University of Kiel, Brunswiker Straße 10, 24105 Kiel, Germany

[2] Institute of Epidemiology, Christian-Albrechts University of Kiel, Niemannsweg 11, 24105 Kiel, Germany

Published online: 22 August 2015

higher risk for cleft palate only compared to non-carriers, but only if their mother had smoked during pregnancy. The epidemiologic literature has spawned a number of strategies to detect gene–environment interactions (G × E) using classical case–control and cohort designs (Thomas 2010). Despite the abundance of analytical methods available, however, only a few successful identifications of G × E have been reported in the scientific literature so far. This apparently low level of success is explicable mainly by the fact that G × E studies face additional challenges over and above those of mere genotype–phenotype association studies (Dempfle et al. 2008; Aschard et al. 2012). Most importantly, case–control and cohort designs have limited power to detect G × E when the marginal and interaction effects are moderate at best, as is probably the case for most complex diseases in humans.

A less popular albeit very efficient strategy to unravel G × E is the adoption of a case-only (CO) design (Piegorsch et al. 1994). The CO design has been deemed superior to other approaches by many scholars because of its greater per-sample power (Yang et al. 1997; Gauderman 2002; Kraft et al. 2007; Yadav et al. 2015). Moreover, a meta-analysis by Dennis et al. (2011) suggests that CO and case–control studies of G × E yield similar estimates of the underlying interaction effects. The CO design also has practical benefits in that it sidesteps the costs and difficulties of identifying and recruiting suitable controls. In addition, cases are usually easier to characterize in terms of both their phenotype and their environmental exposure than controls, particularly in retrospective studies.

The ability to detect G × E in a CO study critically depends upon the validity of the key assumption underlying the design, namely that G and E are stochastically independent in the general population. This assumption seems legitimate in most instances because humans hardly adapt their environment to their genotype (anymore). Moreover, it is sometimes even possible to dispense with the independence assumption if the causes of non-independence, or proxies thereof, can be measured (Albert et al. 2001; Gatto et al. 2004; Cheng 2006). However, by definition, such an assessment would be difficult for one particular type of curtailment of the G–E independence assumption, namely hidden population stratification (PS) (Wang and Lee 2008; Chen et al. 2009). In fact, PS has been recognized before as a potential confounder of genome-wide association studies (GWAS) and various strategies have been proposed to allow for PS at the level of the statistical data analysis (Pritchard and Rosenberg 1999; Devlin and Roeder 1999; Price et al. 2006; Wang 2009). One common approach has been the quantification of surrogate measures of population affiliation, for example, through principal component analysis (PCA), followed by their inclusion into the actual statistical evaluation of the genotype–phenotype relationship (Pritchard and Rosenberg 1999; Devlin and Roeder 1999; Price et al. 2006; Wang 2009). However, as will be expounded in more detail below (see "Discussion"), this strategy is inapt under the CO paradigm because PCA would require genetic data at the population level. Another popular method to allow for PS in GWAS has been the use of so-called 'genomic control' (GC) markers, i.e., of markers unlinked to the test marker, to adjust the test statistic of interest (Devlin and Roeder 1999; Bacanu et al. 2000; Devlin et al. 2001, 2004). More specifically, a correction factor quantifying the PS-induced extra variation of the test statistic under the null hypothesis is estimated from the GC markers, presuming that the vast majority of them are not disease-associated. Having stood the test of many case–control GWAS, the GC method appears to be a good candidate to allow for PS in CO studies of G × E as well. Alternatively, if and when possible, confounding by PS can be circumvented altogether by the adoption of a family-based design. The most popular version of this approach is the evaluation of case–parent trios by a so-called 'transmission-disequilibrium test (TDT)' that conditions the impact on disease risk of the case genotype on the respective parental genotypes (Schaid and Sommer 1994; Schaid 1999; Lake and Laird 2004; Kistner et al. 2009; Chen et al. 2009).

As yet, the role of PS in G × E studies using a CO design has not been explored in much detail. Most importantly, a systematic analysis is still lacking of how the various possible types of PS may affect the validity and power of such studies. Therefore, we systematically evaluated, by simulation, the effects of PS on the type I error rate, power and bias of the ensuing effect estimates in CO studies of G × E interaction. Three plausible scenarios of PS were considered, namely genetic-only (PS_G), environment-only (PS_E), and joint genetic

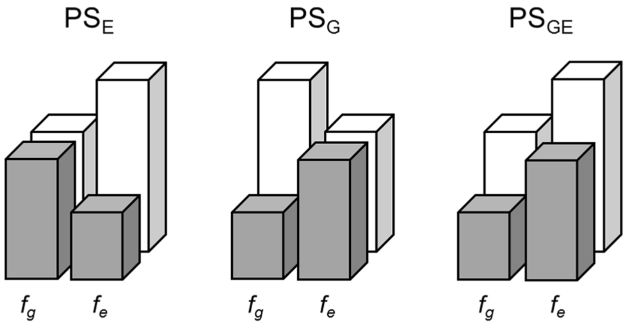


Fig. 1 Illustration of three types of population stratification (PS) potentially affecting G × E studies. The *gray* and *white bars* correspond to two distinct populations. The height of each *bar* is proportional to either the risk allele frequency f_g (*left* hand side) or the exposure frequency f_e (*right* hand side). Under PS_E, the two subpopulations differ only in terms of f_e whereas f_g is constant (and vice versa for PS_G). Under PS_{GE}, both f_g and f_e are different in the two subpopulations

and environmental stratification (PS_{GE}; Fig. 1). Within this framework, we explored three GC-modified test statistics as a means to allow for PS in CO studies of G × E interaction. We also benchmarked the proposed test statistics against a stratified analysis assuming known population affiliation, which should provide maximum statistical power under PS. The results were compared to those obtained with a TDT-like test of G × E interaction, applied to case–parent trios.

Materials and methods

Assessing G × E in CO studies

Let D denote a (binary) disease status. For the sake of simplicity, E is assumed here to be binary as well (e.g., 'exposed' vs 'non-exposed'). Let $G = 0$, 1 or 2 be the allele count of a biallelic marker to be tested for G × E interaction. We will henceforth assume a logistic model relating disease risk π_D to G and E, i.e.,

$$logit(\pi_D) = \beta_0 + \beta_G \cdot G + \beta_E \cdot E + \gamma \cdot G \cdot E \tag{1}$$

In Eq. (1), γ is the statistical interaction between G and E while β_G and β_E are the two main effects. Following Piegorsch et al. (1994), γ can be estimated in a CO study if two key assumptions are met, namely that (1) the disease of interest is sufficiently rare and (2) that G and E are stochastically independent in the general population. Under these conditions, γ approximately equals the log odds ratio (OR) of the association between G and E among cases, which can be estimated by fitting

$$logit\{P(E = 1)\} = \theta_0 + \theta \cdot G \tag{2}$$

to the CO data. In the absence of G × E interaction (i.e., if $\gamma = 0$), θ equals zero provided that the assumptions underlying the CO design are correct. In other words, testing the null hypothesis H_0:$\theta = 0$ (i.e., no association between G and E among cases) is equivalent to testing $H_0{}'$:$\gamma = 0$ (i.e., no G × E in the source population). The significance of rejecting H_0 on the basis of CO data can be assessed using the Wald test statistic

$$w = \frac{\hat{\theta}}{s_\theta} \tag{3}$$

where $\hat{\theta}$ is the maximum-likelihood estimate of θ and s_θ is the standard error of $\hat{\theta}$. Under H_0, the test statistic w asymptotically follows a standard normal distribution.

Accounting for PS in CO studies

In the presence of PS, the false positive rate of the Wald test may be inflated if the spurious population-level association between G and E is carried forward to the study sample. A straightforward albeit idealistic way to allow for PS in CO studies would be to stratify the Wald test by population affiliation which would render the test valid and should provide maximum power under PS. Therefore, we considered this stratified type of analysis as a benchmark for the efficiency of the proposed modifications of the Wald test statistic described below.

One way to allow for PS in CO studies would be an adoption of the GC approach originally devised for case–control disease association studies. Here, we propose three different GC-based test statistics tailored to the specifics of G × E studies using a CO design. The first statistic is reminiscent of the original idea by Devlin et al. (2001) to align the median of the relevant test statistic to its theoretical expectation under H_0. In CO studies, a robust estimate of the required correction factor is provided by the median of w^2 taken over the GC markers, divided by 0.456 (i.e., the median of a χ^2 distribution with one degree of freedom). Denoting this correction factor by $\hat{\lambda}_1$, the significance of an observed G × E interaction is assessed by comparing

$$w' = \frac{w^2}{\hat{\lambda}_1} \tag{4}$$

to a χ^2 distribution with one degree of freedom.

The second proposed statistic is similar to a suggestion by Devlin et al. (2004). This time, the correction factor $\hat{\lambda}_2$ equals the mean (rather than the standardized median) of w^2, taken over the GC markers. The significance of an observed G × E interaction is assessed by comparing

$$w'' = \frac{w^2}{\hat{\lambda}_2} \tag{5}$$

to a $F_{1,L}$ distribution, where L is the number of GC markers employed to estimate $\hat{\lambda}_2$. Use of $\hat{\lambda}_2$ rather than $\hat{\lambda}_1$ is motivated by the fact that the former allows better for the increased standard error of the correction factor when L is small.

The third modification of the Wald test statistic, inspired by Wang (2009), is based upon the $\hat{\theta}$ values of individual GC markers. Denoting the sample variance of these estimates by $\hat{\lambda}_3$, the modified test statistic equals

$$w''' = \frac{w}{\sqrt{1 + \frac{\hat{\lambda}_3}{s_\theta^2}}}. \tag{6}$$

For moderate to large sample sizes, w''' follows a standard normal distribution under H_0. Note that, contrary to w' and w'', the correction factor in w''' is not constant but depends upon the test marker under consideration. Since a non-negative term is added to the denominator, w''' will always be smaller than the original test statistic w.

Testing G × E in case–parent trios

An extension of the TDT to G × E studies was proposed by Schaid (1999). Let τ denote the probability that a heterozygous parent transmits a particular allele of the test marker to their affected child. The proposed test compares τ between trios with a non-exposed case (τ_0) and trios with an exposed case (τ_1). This comparison yields a valid test of $H_0{:}\theta = 0$ because H_0 logically implies $\tau_0 = \tau_1$.

Data simulation

Our simulation study was carried out in a multi-tiered fashion. First, CO genotype and exposure data were simulated in the presence of one of three different types of PS. We employed an additive model of the genotype–phenotype relationship to assign a disease status to an individual with genotype G and exposure status E, using Eq. (1). Only affected individuals were retained for further analysis. Simulations were carried out with fixed $\beta_0 = -4.6$, $\beta_G = 0.41$ and $\beta_E = 0.41$. The values of β_G and β_E were chosen so as to induce a main effect (OR = 1.5) that is generic for multifactorial diseases. Intercept β_0 was fixed at −4.6 (equivalent to a baseline risk of ~ 1 %) to comply with the rarity assumption underlying the CO design. We set $\gamma = 0$ to evaluate type I error rates and choose various positive values of γ to evaluate the power of the different tests under consideration.

To investigate the impact of PS on the original Wald test statistic w, we combined CO data that were simulated in two subpopulations (*pop1* and *pop2*), each with a different parameter setting. The risk allele frequency (f_g) and the environmental exposure frequency (f_e) were varied in *pop2* to induce various levels of PS. Three types of PS were generated. For purely genetic PS (PS_G), f_e was set equal to 0.1 in both subpopulations while f_g was fixed at 0.1 in *pop1* and varied between 0.1 and 0.9 in *pop2*. Similarly, for purely environmental PS (PS_E), f_g was set equal to 0.5 in both subpopulations while f_e was fixed at 0.1 in *pop1* and varied between 0.05 and 0.40 in *pop2*. Finally, for PS_{GE}, f_g and f_e were fixed at 0.5 and 0.1, respectively, in *pop1*. In *pop2*, the two parameters were varied between 0.1 and 0.5 (f_g) and between 0.05 and 0.20 (f_e), respectively. To evaluate the effect of population composition, the proportion of *pop1* data was varied between 0.1 and 0.9 while f_g and f_e were fixed at 0.5 and 0.1, respectively, in *pop1*, and at 0.1 and 0.2, respectively, in *pop2*.

To investigate the performance of the GC-modified test statistics under PS_{GE}, the genotypes of 500 additional GC markers were simulated for each case in addition to the test marker genotype. The minor allele frequency was drawn randomly from [0.05, 0.25] for every uneven marker number and from [0.20, 0.40] for every even marker number in *pop1*, and vice versa in *pop2*. This induced an average allele frequency difference of about 0.15 between the two subpopulations. Genotype frequencies of GC markers were set to Hardy–Weinberg expectations. At the test marker, f_g was set at 0.30 and 0.15, respectively, in *pop1* and *pop2* while f_e was fixed at 0.15 in *pop1* and varied from 0.03 to 0.15 in *pop2* to induce varying levels of PS_{GE}. Interaction term γ was set equal to 0.25 or 0.5 to evaluate the power of the different modified tests over a wide range of PS_{GE}. The two values of γ were chosen so as to induce either a small (OR = 1.3) or a moderate (OR = 1.6) interaction effect.

To confirm that family-based designs are robust against PS, trio data comprising cases and their parents were simulated using the *trioGxE* tool implemented in R (Shin et al. 2013). This tool supports simulation of both child exposure and G × E interaction. Child genotypes were first sampled according to the laws of Mendelian inheritance from given parental genotypes. For convenience, we modified the default risk model of *trioGxE* to comply with Eq. (1) and used the same β values as in the CO data simulation. We again employed an additive model of the genotype–phenotype relationship to assign a disease status to each child. Only trios with affected children were retained. Trio data from *pop1* and *pop2* were combined to induce PS. For all three types of PS, f_g and f_e were fixed at 0.3 and 0.15, respectively, in *pop1* and modified in *pop2* as follows. For PS_G, f_g was set to 0.15, for PS_E, f_e was set to 0.05 and, for PS_{GE}, f_g and f_e were set to 0.15 and 0.05, respectively, in *pop2*. To compare the type I error rates and power of the TDT to those of the GC-modified Wald tests statistics, trio data were simulated under varying levels of PS_{GE} using the same parameter combinations as employed in the simulation of singleton cases. Finally, to evaluate the power of the TDT under PS_{GE}, we varied interaction term γ in the interval [0, 1] while f_g and f_e were set at 0.3 and 0.15, respectively, in *pop1*, and at 0.15 and 0.05, respectively, in *pop2*.

For each parameter combination considered, we simulated 10,000 replicates comprising either a CO data set with 1000 cases or a set of 1000 case–parent trios. To ensure adequate coverage close to 0 and 1, the Agresti-Coull method was used to calculate 95 % confidence intervals for proportions (Agresti and Coull 1998). All simulations were implemented in R, using in-house protocols.

Results

Our simulation-based study of the CO design revealed that, under purely genetic PS (PS_G), the type I error rate of the unadjusted Wald test for G × E is close to its nominal

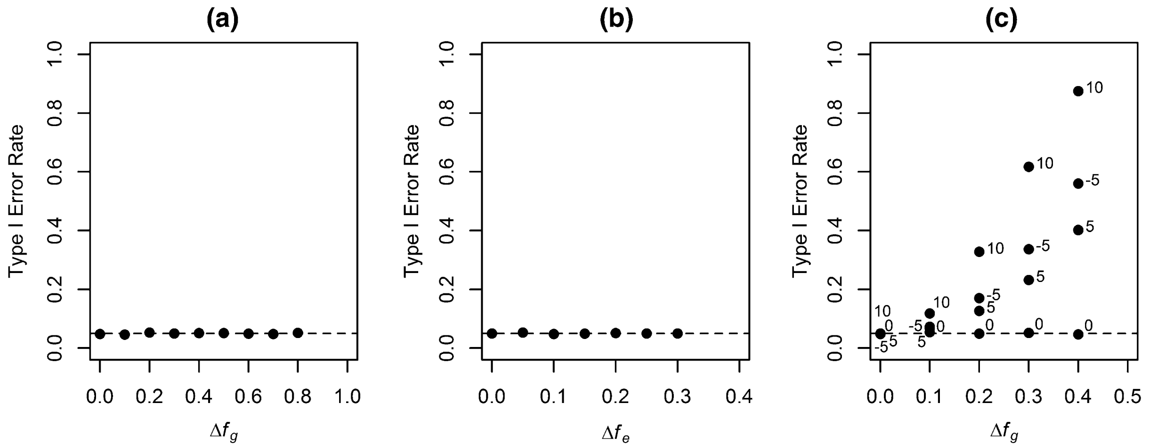


Fig. 2 PS-induced type I error rate of the unadjusted Wald test for G × E (H_0:$\theta = 0$) in a CO design. CO data were simulated under PS_G (**a**), PS_E (**b**) and PS_{GE} (**c**). Δf_g and Δf_e denote the risk allele and exposure frequency difference, respectively, between subpopulations. Under PS_G, only f_g was varied in *pop2* whereas f_e was kept constant. Similarly, under PS_E, only f_e was varied whereas f_g was kept constant. Under PS_{GE}, both f_g and f_e were varied. Labels in (**c**) correspond to Δf_e values, multiplied by 100; negative signs mark settings where f_e was lower in *pop2* than in *pop1*. *Dashed lines* depict the nominal significance level of 0.05

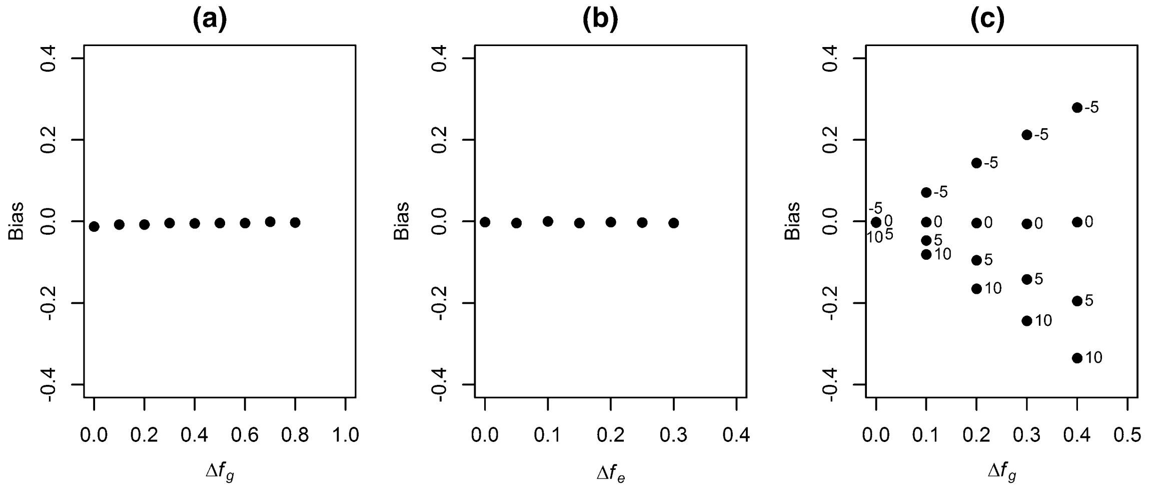


Fig. 3 PS-induced bias of estimates of interaction effect γ. For details: see legend to Fig. 2

level of 0.05, irrespective of the difference in risk allele frequency (Δf_g) between the two subpopulations (Fig. 2a). The same was found to be true in the presence of environment-only PS (PS_E). Again, the type I error rate was close to its nominal level of 0.05 irrespective of the subpopulation difference in exposure frequency (Δf_e; Fig. 2b). By contrast, in simulations carried out under joint PS with respect to both genetic and environmental factors (PS_{GE}), the type I error rate was found to be inflated proportional to the respective subpopulation differences (Fig. 2c). At the extreme end, the type I error rate was as high as 90 % for $\Delta f_g = 0.40$ and $\Delta f_e = 0.10$. Similarly, the bias of the γ estimate was found to be negligible in the presence of PS_G or PS_E, but turned out to be considerable under PS_{GE} (Fig. 3). Moreover, as was to be expected, both the type I error rate and the bias attained their maximum when the population-wise CO data were mixed in equal proportions (Fig. 4).

The type I error rate and power of the six test statistics considered in our study, namely w, w', w'', w''', w^* and TDT, were found to vary markedly with different degrees of PS_{GE} (Table 1). The original Wald test w showed an inflated type I error rate when there was PS_{GE} (i.e., when $\Delta f_e > 0$; Fig. 2c). By contrast, the stratified analysis with test statistic w^*, which idealistically assumed that the true subpopulation affiliation was known, was always valid and yielded a type I error rate close to its nominal level of 0.05, as expected. Median-adjusted GC statistic *w*' showed a type I error rate close to its nominal level only for $\Delta f_e \leq 0.03$, but was found to be rather conservative for $\Delta f_e \geq 0.06$.

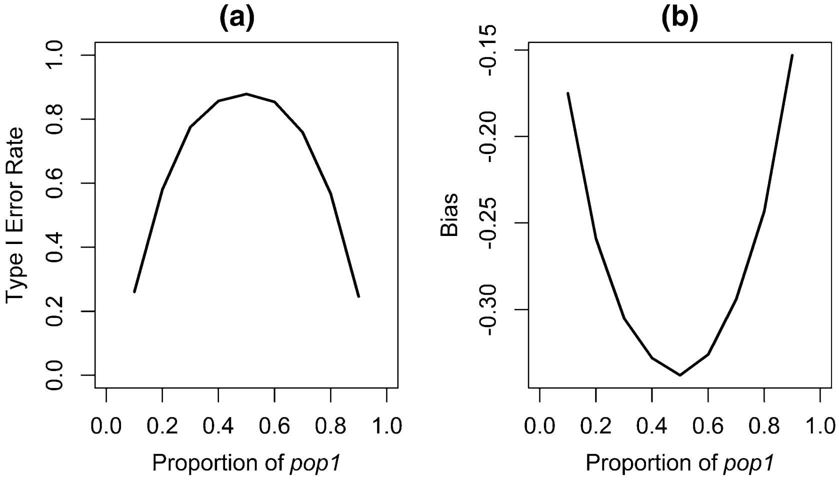


Fig. 4 Type I error rate of the unadjusted Wald test (**a**) and bias of the estimates of γ (**b**) as a function of the proportion of *pop1* data. CO data were simulated under PS_{GE} with f_g and f_e fixed at 0.5 and 0.1, respectively, in *pop1*, and at 0.1 and 0.2, respectively, in *pop2*

Table 1 Validity and power of the original Wald test (*w*) and five other tests under different levels of PS_{GE}

γ	Δf_e (×100)	*W*	*w*′	*w*″	*w*‴	w*	TDT
0	0	0.050 (0.046; 0.054)	0.049 (0.045; 0.053)	0.050 (0.046; 0.054)	0.003 (0.002; 0.004)	0.052 (0.048; 0.056)	0.046 (0.042; 0.050)
	3	0.075 (0.070; 0.080)	0.047 (0.043; 0.051)	0.049 (0.045; 0.053)	0.003 (0.002; 0.004)	0.046 (0.042; 0.050)	0.049 (0.045; 0.053)
	6	0.152 (0.145; 0.159)	0.027 (0.024; 0.030)	0.039 (0.035; 0.043)	0.002 (0.001; 0.003)	0.048 (0.044; 0.052)	0.053 (0.049; 0.058)
	9	0.302 (0.293; 0.311)	0.005 (0.004; 0.007)	0.020 (0.017; 0.023)	0.0007 (0.0003; 0.0015)	0.051 (0.047; 0.055)	0.049 (0.045; 0.053)
	12	0.568 (0.558; 0.578)	0.0002 (0; 0.0008)	0.006 (0.005; 0.008)	0.0001 (0; 0.0006)	0.046 (0.042; 0.050)	0.053 (0.049; 0.058)
0.25	0	0.580 (0.570; 0.590)	0.568 (0.558; 0.578)	**0.572** (0.562; 0.582)	0.198 (0.190; 0.206)	0.551 (0.541; 0.561)	0.402 (0.392; 0.412)
	3	0.732 (0.723; 0.741)	0.568 (0.558; 0.578)	**0.607** (0.597; 0.616)	0.218 (0.210; 0.226)	0.520 (0.510; 0.530)	0.381 (0.371; 0.391)
	6	0.855 (0.848; 0.862)	0.396 (0.386; 0.406)	**0.522** (0.512; 0.532)	0.156 (0.149; 0.163)	0.495 (0.485; 0.505)	0.356 (0.347; 0.365)
	9	0.939 (0.934; 0.943)	0.142 (0.135; 0.149)	0.360 (0.351; 0.369)	0.059 (0.054; 0.064)	**0.448** (0.438; 0.458)	0.334 (0.325; 0.343)
	12	0.983 (0.980; 0.985)	0.014 (0.012; 0.016)	0.154 (0.147; 0.161)	0.006 (0.005; 0.008)	**0.414** (0.404; 0.424)	0.313 (0.304; 0.322)
0.5	0	0.995 (0.993; 0.996)	0.993 (0.991; 0.994)	**0.994** (0.992; 0.995)	0.927 (0.922; 0.932)	0.992 (0.990; 0.994)	0.942 (0.937; 0.946)
	3	0.998 (0.997; 0.999)	0.987 (0.985; 0.989)	**0.992** (0.990; 0.994)	0.923 (0.918; 0.928)	0.988 (0.986; 0.990)	0.937 (0.932; 0.942)
	6	0.999 (0.998; 0.999)	0.933 (0.928; 0.938)	0.981 (0.978; 0.983)	0.840 (0.833; 0.847)	**0.985** (0.982; 0.987)	0.925 (0.920; 0.930)
	9	1.000 (0.999; 1.000)	0.684 (0.675; 0.693)	0.928 (0.923; 0.933)	0.586 (0.576; 0.596)	**0.974** (0.971; 0.977)	0.910 (0.904; 0.915)
	12	1.000 (0.999; 1.000)	0.206 (0.198; 0.214)	0.753 (0.744; 0.761)	0.182 (0.175; 0.190)	**0.955** (0.951; 0.959)	0.892 (0.886; 0.898)

The table contains the proportion (95 % CI) of nominally significant results, corresponding to the type I error when $\gamma = 0$ and corresponding to the power when $\gamma \neq 0$. The test statistic with the largest power for a given parameter combination is highlighted by bold print. Parameter f_g was fixed at 0.30 and 0.15 at the test marker, respectively, in *pop1* and *pop2*; f_e was fixed at 0.15 in *pop1* and varied from 0.03 to 0.15 in *pop2*, to induce PS_{GE}

Mean-adjusted GC statistic w'' performed slightly better in the sense that it was less conservative than w' for large values of Δf_e. Finally, the GC statistic w''' had a lower type I error rate than the original w at all Δf_e values considered here, but was very conservative compared to w' and w''. This finding is explicable by the fact that w''' has a non-negative term added to its denominator and therefore always deflates the original test statistic w. For all test statistics analyzed, the power to detect $\gamma \neq 0$ was positively correlated with the size of the true interaction effect as expected, i.e., greater power was attained at $\gamma = 0.5$ than at $\gamma = 0.25$. Furthermore, the power of all five valid test statistics was inversely related to Δf_e which implies that the power to detect a given interaction effect decreases with increasing PS_{GE} (note that the power of w was irrelevant because it is invalid under PS_{GE}). Noteworthy, at low levels of PS_{GE}, the power of all GC statistics except w''' was larger than that ensuing from the same number of cases (plus parents) with a TDT. Additionally, at low levels of PS_{GE}, the power of w' and w'' was found to be similar to that of w^*. At high levels of PS_{GE}, TDT provided maximum power of the non-stratified tests (i.e., of all tests other than w^*), followed by GC-modified statistics w'' and w'.

Comparison between CO and trio design

Both the TDT and the unmodified Wald test are valid when only a single factor (i.e., either G or E) is subject to PS (Table 2). However, while the unadjusted CO analysis of G × E suffered from an inflated type I error rate under PS_{GE}, this was not the case for the TDT. Instead, the extension of the TDT consistently exhibited type I error rates close to their nominal level. The power of the TDT was also positively correlated with the true interaction effect (Fig. 5) and approached unity for strong interaction (i.e., $\gamma = 1$). Interestingly, the GC-modified statistic w'' provided power similar to the TDT (with the same number of trios as cases) except for very small interaction effects. Overall, the power of the five different valid test statistics could be ranked as $w^* > \text{TDT} \geq w'' > w' > w'''$.

The resulting 95 % CIs (Tables 1 and 2) indicate that both the type I error rates and the power figures were estimated sufficiently accurately with the chosen number of simulation replicates (i.e., 10,000) to justify the above-mentioned qualitative conclusions about the relative merits of the different test statistics.

Table 2 Type I error rates of the Wald test (w) and the TDT in a CO or case–parents design, respectively, under different types of PS

pop1		*pop2*		Type I error rate (95 % CI)	
f_g	f_e	f_g	f_e	w	TDT
0.3	0.15	0.3	0.15	0.050 (0.046; 0.054)	0.050 (0.046; 0.054)
0.3	0.15	0.15	0.15	0.049 (0.045; 0.053)	0.047 (0.043; 0.051)
0.3	0.15	0.3	0.05	0.051 (0.047; 0.055)	0.052 (0.048; 0.056)
0.3	0.15	0.15	0.05	0.104 (0.098; 0.110)	0.049 (0.045; 0.053)

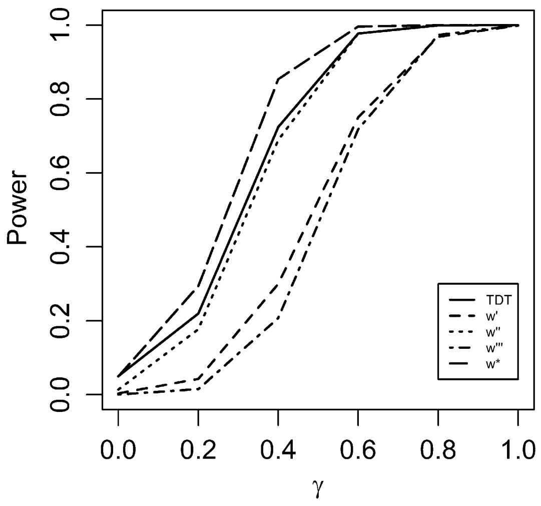


Fig. 5 Power of four modified Wald tests and the TDT, respectively. Data were simulated under PS_{GE} with f_g and f_e fixed at 0.3 and 0.15, respectively, in *pop1*, and at 0.15 and 0.05, respectively, in *pop2*. Interaction parameter γ was varied in the interval [0, 1]

Discussion

We studied the impact of population stratification (PS) on the validity of the case-only (CO) design for studies of gene–environment interactions (G × E). To this end, we simulated CO data under three possible PS scenarios, namely genetic-only (PS_G), environment-only (PS_E), and joint environmental and genetic stratification (PS_{GE}). Further, we described different ways in which to allow analytically for PS in CO studies of G × E. For the sake of simplicity, we considered only two subpopulations here but, qualitatively, our conclusions should also apply to more complex types of PS.

We systematically evaluated the type I error rate and bias arising under a wide spectrum of stratification scenarios. One important finding was that CO studies of G × E are not impeded by PS as long as the frequency of only one factor (i.e., either G or E) differs between subpopulations. However, without further adjustment, the CO approach was found to be invalid under PS_{GE}, with seriously inflated type I error rates as expected. This limitation is due to the fact that PS_{GE} creates an association at the population level between G and E that violates the key assumption of the CO design. Moreover, our simulations revealed a strong bias to affect the estimate of interaction term γ under PS_{GE}.

Wang and Lee (2008) made a similar observation and proposed a "boundary formula" to limit the amount of PS-induced bias. For their formula to be applicable, however, the genotype and exposure frequencies within subpopulations need to be known, which is difficult or even impossible in practice where the nature of PS is likely to be hidden.

The main goal of our work was to develop means to take PS properly into account in CO studies of G × E. To this end, we explored three different modifications of the Wald test statistic (w) that were inspired by the genomic control (GC) method. To the best of our knowledge, our study is the first to deploy the GC idea in the context of CO studies of G × E. We benchmarked the GC-modified test statistics against a stratified Wald test idealistically assuming that the population affiliation of each individual case was known. Our simulations confirmed that this test would be valid and provide greater power than any of the PS-naïve tests. Finally, we investigated how family-based approaches such as an extension of the TDT proposed by Schaid (1999) tend to avoid excessive false positive results under $\mathrm{PS}_{\mathrm{GE}}$. For two major reasons, we did not consider approaches involving the quantification of surrogates of population affiliation, including principal component analysis (PCA). First and foremost, for such adjustments to be valid, the nature of the underlying PS (i.e., the relative proportion of each subpopulation) would have to be determined at the population level which, in turn, requires genetic data that are population-representative as well. However, as we have demonstrated above, PS affects the validity of CO studies of G × E only if there is both PS_{G} and PS_{E}. In this case, the genetic PS present in a CO sample is not necessarily the same as the sought after PS_{G} at the population level because the composition of the CO sample would depend critically upon the (variable) subpopulation-specific exposure frequencies. We may therefore conclude that surrogate quantification of subpopulation affiliation, such as PCA, is unsuitable in principle to correct for PS in CO studies of G × E. Second, PCA and other techniques only partly resolve the PS problem because the surrogate measures (e.g., PCA dimensions) usually included in subsequent statistical analyses are likely to explain only a minor fraction of the true genetic variation. Whether or not the selected measures render the statistical tests sufficiently valid is most often unclear. By contrast, the GC method considered here corrects for PS directly at the level of the test statistic.

Our simulations revealed that the proposed GC modifications of w compensate well the impact of $\mathrm{PS}_{\mathrm{GE}}$ upon the validity of CO studies of G × E. Interestingly, at low levels of $\mathrm{PS}_{\mathrm{GE}}$, the power of all test statistics except w''' was found to be larger than that provided by the TDT using the same number of trios as cases (here: 1000). Moreover, even under extreme levels of $\mathrm{PS}_{\mathrm{GE}}$, statistic w'' provided reasonable power (0.75) to detect moderate interaction (i.e., $\gamma = 0.5$). The first of our GC-modified test statistics, w', uses a correction factor $\hat{\lambda}_1$ that is the same for all GC markers. It performs well at low levels of $\mathrm{PS}_{\mathrm{GE}}$, but is conservative under more extreme scenarios. This selective loss of power may be due to the considerable standard error of $\hat{\lambda}_1$ which in turn depends upon many factors, including the number of unlinked markers used for estimation and the truly underlying level of PS (Devlin et al. 2004). Our second proposed test statistic, w″, is similar to the GCF method suggested by Devlin et al. (2004) and takes the variance inflation of $\hat{\lambda}_2$ for small numbers of GC markers properly into account. Interestingly, this method was found to perform slightly better than w' at low levels of $\mathrm{PS}_{\mathrm{GE}}$ but was also conservative under more extreme settings. Finally, test statistic w''' uses a correction factor that varies from one test marker to the other. This type of modification served to reduce the type I error rate in all $\mathrm{PS}_{\mathrm{GE}}$ settings, but turned out to be very conservative probably because the modification always deflates the original test statistic w. Overall, the proposed GC-based modifications allow for the effect of PS at the level of the test statistic and, as was shown here, can be assumed conservative particularly at high levels of $\mathrm{PS}_{\mathrm{GE}}$.

Our simulations also confirmed the validity of the TDT and revealed that, at high levels of $\mathrm{PS}_{\mathrm{GE}}$, it provides more power per trio than the GC-adjusted statistics provide per case. However, although family-based designs may therefore seem to resolve the PS problem in G × E studies, they also have major drawbacks: The TDT requires genotype data from additional family members (e.g., parents) which are difficult to obtain in the first place. Moreover, recruiting older family members such as parents can be expensive, time-consuming and difficult, particularly for diseases with late onset. By contrast, the GC-based methods proposed here are easy to implement and require the genotyping of only a moderate number of unlinked markers in the same set of cases. In the context of a genome-wide G × E interaction study, a vast number of GC markers can be expected to be readily available anyway. Moreover, family-based tests of G × E recently have been shown to be potentially biased in the presence of $\mathrm{PS}_{\mathrm{GE}}$ if the genetic variant under study is not itself causal but only a proxy for a truly causal variant (Shi et al. 2011; Weinberg et al. 2011). This is because, in cases where the subpopulation-specific exposure prevalence is correlated with the subpopulation-specific linkage disequilibrium between the proxy and the causal variant, the exposure and the transmission at the proxy may be correlated even when there is no interaction between the exposure and the causal genetic variant (Shi et al. 2011).

In summary, we have shown that joint environmental and genetic PS can impact the validity of CO studies of G × E. We also demonstrated that the GC paradigm from gene-disease association studies can be extended to G × E to rectify this problem. Our results imply that future research of G × E may safely adopt a CO approach to exploit existing GWAS data sets. At the same time, we urge that relevant data resources should generally aim at comprising large numbers of cases for whom genetic and environmental exposure data may be easier to obtain than for controls or patient relatives.

Acknowledgments This study was carried out as part of the Research Training Group 'Genes, Environment and Inflammation', supported by the Deutsche Forschungsgemeinschaft (GRK 1743).

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interest.

References

Agresti A, Coull BA (1998) Approximate is better than "exact" for interval estimation of binomial proportions. Am Stat 52:119–126. doi:10.2307/2685469

Albert PS, Ratnasinghe D, Tangrea J, Wacholder S (2001) Limitations of the case-only design for identifying gene-environment interactions. Am J Epidemiol 154:687–693

Aschard H, Lutz S, Maus B, Duell EJ, Fingerlin TE, Chatterjee N, Kraft P, Van Steen K (2012) Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. Hum Genet 131:1591–1613. doi:10.1007/s00439-012-1192-0

Bacanu SA, Devlin B, Roeder K (2000) The power of genomic control. Am J Hum Genet 66:1933–1944. doi:10.1086/302929

Begg CB, Zhang ZF (1994) Statistical analysis of molecular epidemiology studies employing case- series. Cancer Epidemiol Biomark Prev 3:173–175

Chen YH, Lin HW, Liu H (2009) Two-stage analysis for gene-environment interaction utilizing both case-only and family-based analysis. Genet Epidemiol 33:95–104. doi:10.1002/gepi.20357

Cheng KF (2006) A maximum likelihood method for studying gene-environment interactions under conditional independence of genotype and exposure. Stat Med 25:3093–3109. doi:10.1002/sim.2506

Dempfle A, Scherag A, Hein R, Beckmann L, Chang-Claude J, Schäfer H (2008) Gene-environment interactions for complex traits: definitions, methodological requirements and challenges. Eur J Hum Genet 16:1164–1172. doi:10.1038/ejhg.2008.106

Dennis J, Hawken S, Krewski D, Birkett N, Gheorghe M, Frei J, McKeown-Eyssen G, Little J (2011) Bias in the case-only design applied to studies of gene-environment and gene-gene interaction: a systematic review and meta-analysis. Int J Epidemiol 40:1329–1341. doi:10.1093/ije/dyr088

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 60:155–166. doi:10.1006/tpbi.2001.1542

Devlin B, Bacanu SA, Roeder K (2004) Genomic control to the extreme. Nat Genet 36:1129–1130. doi:10.1038/ng1104-1131 **author reply 1131**

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 33:1014–1024. doi:10.1093/ije/dyh306

Gauderman WJ (2002) Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol 155:478–484

Hwang SJ, Beaty TH, Panny SR, Street NA, Joseph JM, Gordon S, McIntosh I, Francomano CA (1995) Association study of transforming growth factor alpha (TGF alpha) TaqI polymorphism and oral clefts: indication of gene-environment interaction in a population-based sample of infants with birth defects. Am J Epidemiol 141:629–636

Kistner EO, Shi M, Weinberg CR (2009) Using cases and parents to study multiplicative gene-by-environment interaction. Am J Epidemiol 170:393–400. doi:10.1093/aje/kwp118

Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ (2007) Exploiting gene-environment interaction to detect genetic associations. Hum Hered 63:111–119. doi:10.1159/000099183

Lake SL, Laird NM (2004) Tests of gene-environment interaction for case-parent triads with general environmental exposures. Ann Hum Genet 68:55–64. doi:10.1046/j.1529-8817.2003.00073.x

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 13:153–162

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 38:904–909. doi:10.1038/ng1847

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228. doi:10.1086/302449

Schaid DJ (1999) Case-parents design for gene-environment interaction. Genet Epidemiol 16:261–273

Schaid DJ, Sommer SS (1994) Comparison of statistics for candidate-gene association studies using cases and parents. Am J Hum Genet 55:402–409

Shi M, Umbach DM, Weinberg CR (2011) Family-based gene-by-environment interaction studies: revelations and remedies. Epidemiology 22:400–407. doi:10.1097/EDE.0b013e318212fec6

Shin JH, McNeney B, Graham J (2013) A data smoothing approach to explore and test gene-environment interaction in case-parent trio data. R package: version 0.1–1

Thomas D (2010) Gene-environment-wide association studies: emerging approaches. Nat Rev Genet 11:259–272. doi:10.1038/nrg2764

Wang K (2009) Testing for genetic association in the presence of population stratification in genome-wide association studies. Genet Epidemiol 33:637–645. doi:10.1002/gepi.20415

Wang LY, Lee WC (2008) Population stratification bias in the case-only study for gene-environment interactions. Am J Epidemiol 168:197–201. doi:10.1093/aje/kwn130

Weinberg CR, Shi M, Umbach DM (2011) A sibling-augmented case-only approach for assessing multiplicative gene-environment interactions. Am J Epidemiol 174:1183–1189. doi:10.1093/aje/kwr231

Yadav P, Freitag-Wolf S, Lieb W, Krawczak M (2015) The role of linkage disequilibrium in case-only studies of gene–environment interactions. Hum Genet 134:89–96. doi:10.1007/s00439-014-1497-2

Yang Q, Khoury MJ, Flanders WD (1997) Sample size requirements in case-only designs to detect gene-environment interaction. Am J Epidemiol 146:713–720

2.3 Gene-smoking Interaction in Inflammatory Bowel Disease: Meta-analysis of 10 Case-only Studies Comprising over 12,750 Patients

2.3.1 Summary

This paper uses a real inflammatory bowel disease (IBD) data set to illustrate the genome-wide utility of the CO design to detect gene-smoking interaction in IBD. The etiology of IBD, including Crohn disease (CD) and ulcerative colitis (UC), involves both G and E factors, but the underlying biological mechanisms are only poorly understood. In particular, little is known about the possible role of G×E. In consequence, despite the many genotype-phenotype associations identified through GWAS, more than 70% of the heritability of IBD remained unexplained. In addition to other environmental stimuli, such as hygiene, antibiotics and diet, smoking is known to be a major environmental risk factor for IBD.

This study investigated the possible interaction between SNPs and smoking in relation to CD and UC risk. The analysis covered 10 Immunochip data sets collated by the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC), comprising 12,776 cases (7076 CD and 5700 UC) of known smoking status. In addition, 3730 control samples with information on smoking history were available from four of the participating centers. A total of 156,499 SNPs were tested for gene-smoking interaction, using the CO approach.

Three meta-analyses each were performed for CD and UC, considering one of the following smoking-status contrasts: ‘never vs. ever’, ‘never vs. current’, or ‘never vs. former’. Interactions with a meta-analysis p value $<10^{-4}$ (Wald test) and a heterogeneity p value >0.05 (Cochrane Q test) were subsequently checked for the validity of the underlying gene-smoking independence assumption in controls. This study identified 23 and 19 SNPs with suggestive evidence of gene-smoking interaction in CD and UC, respectively. The majority of these markers (16 SNPs) were located on chromosome 6, thus indicating a potential involvement of the HLA region in IBD.

In summary, using CO approach, this study identified 42 SNPs with strong evidence of an interaction with smoking in IBD. None of these polymorphisms had been reported as being involved in a gene-smoking interaction in IBD before. The lack of LD between the 42 strongly suggestive SNPs and known IBD-associated markers further highlights the potential of an agnostic, hypothesis free genome-wide search for G×E. Undoubtedly, further functional and experimental studies are required to fully clarify the role of the identified gene-smoking interactions in IBD etiology.

2.3.2 Publication (iii)

Yadav et al. (to be submitted)

Gene-smoking interaction in inflammatory bowel disease: Meta-analysis of 10 case-only studies comprising over 12,750 patients

Yadav et al. (to be submitted)

Abstract

Both genetic and environmental factors are thought to play a role in the etiology of inflammatory bowel disease (IBD), including Crohn disease (CD) and ulcerative colitis (UC). In fact, smoking is presumed to be a major risk factor for IBD and at least some of the biological mechanisms of IBD development are likely to be modified by smoking. Therefore, we investigated the possible interaction between single nucleotide polymorphisms (SNPs) and smoking in relation to CD and UC risk. The analysis covered 10 Immunochip data sets collated by the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC), comprising 12,776 cases (7076 CD, 5700 UC) of known smoking status. In addition, 3730 control samples with information on smoking history were available from four of the participating centers. A total of 156,499 SNPs were tested for gene-smoking interaction, using a case-only design. Three meta-analyses each were performed for CD and UC, considering one of the following smoking-status contrasts: 'never vs. ever', 'never vs. current', or 'never vs. former'. Interactions with a meta-analysis p value $<10^{-4}$ (Wald test) and a heterogeneity p value >0.05 (Cochrane Q test) were subsequently checked for the validity of the underlying gene-smoking independence assumption in controls. Our analysis identified 23 and 19 SNPs with suggestive evidence of gene-smoking interaction in CD and UC, respectively. The majority of these markers (16 SNPs) were located on chromosome 6, thus indicating a potential role of the HLA region in IBD. In fact, a focused analysis of classical HLA alleles revealed a suggestive gene-smoking interaction for four alleles in CD and for one allele in UC. Moreover, approximately 30% of the SNPs identified as interacting with smoking in our meta-analyses lie in close vicinity (<1Mb) to SNPs identified as disease-associated in previous genome-wide association studies (GWAS). Noteworthy, several interaction effect differences were observed between CD and UC thereby suggesting a differential role of tobacco smoking in the etiology of IBD.

Keywords: meta-analysis; case-only design; inflammatory bowel diseases; gene-smoking interaction

Introduction

Inflammatory bowel disease (IBD), including Crohn disease (MIM 266600) and ulcerative colitis (MIM 191390), is a chronic, lifelong illnesses of early onset that seriously impedes the quality of life of affected families. The aetiology of IBD involves both genetic and environmental factors, but the biological mechanisms of disease development are only poorly understood. In particular, little is known about the possible role of gene-environment interaction. In consequence, despite the many genotype-phenotype associations identified in past genome-wide association studies (GWAS), more than 70% of the heritability of IBD is yet unaccounted for [1-4].

In addition to other environmental stimuli such as hygiene, oral contraceptives, antibiotics and diet [5], smoking is known to be a major environmental risk factor for IBD [6-11]. Early case-control studies revealed an increased risk for both CD and UC in former smokers whereas current smoking was found to be predisposing to CD, but protective against UC [6, 8-11]. This differential effect on risk was recently confirmed in a large prospective study of 229,111 women from the US Nurses' Health Studies (NHS). Compared to never smokers, the CD hazard ratio was 1.35 for former and 1.90 for current smokers whilst the UC hazard ratio was 1.56 for former, but 0.86 for current smokers. However, with a 95% confidence interval ranging from 0.61 to 1.20, the apparently protective effect against UC of current smoking failed to attain statistical significance.

The aetiological role of smoking in IBD is still not fully understood [12] mainly because of the complex chemical composition of tobacco smoke. Many candidate mechanisms are on the table. For example, smoking has been shown to cause epigenetic changes that promote altered gene expression potentially affecting the innate immune response [9, 12-15]. Smoking also induces changes of the intestinal microbiota, which represents another plausible link to disease aetiology [16-18]. Further possible mechanisms involve the post-translational modification of key proteins by constituents of tobacco smoke, activating the immune response and inducing inflammation. For example, tobacco smoking has been found to induce citrullination of various proteins [19-21]. Citrullination affects the 3-dimensional structure of proteins such that it may cause unfolding and the exposition of interior domains that can subsequently act as antigens. In rheumatoid arthritis, for example, smoking has been identified as an environmental trigger of anti-citrulline immunity in individuals with particular HLA-DRB1 SE alleles [19, 22], a mechanism that may likewise explain the sustaining of a high risk of UC decades after smoking cessation [7].

Gene-environment interaction studies are one way to shed light on the biological mechanisms of disease development [23-25]. As yet, however, only a few studies of gene-smoking interaction have been conducted in the context of IBD. One of these reported a statistically significant interaction between NOD2 gene variant 1007fs (rs2066847), predisposing to CD, and both ever and current smoking [26]. Two other small studies observed a significantly higher smoking-related CD risk among GG homozygotes for SNP rs2241880 of the ATG16L1 gene [27] and among CC (wild-type) homozygotes for SNP rs1343151 in the IL23R gene [28]. No study so far has investigated gene-smoking interactions in IBD at a genome-wide level. Using the genotype data available to the International IBD Genetics Consortium (IIBDGC), we therefore set out to investigate whether the smoking-associated risk for IBD is modified by any of the genetic variants that are either included on the Illumina Immunochip itself, or that are imputable from Immunochip data using publicly available databases. In so doing, we adopted a two-tiered approach adopting a case-only (CO) design to search for gene-smoking interactions (stage I) followed by the verification of the gene-smoking independence assumption implicit to the CO design in controls (stage II).

Materials and Methods

IBD dataset

All samples used in the present study were collected through the IIBDGC and originated from 15 countries in Europe, North America and Australia [4]. Genotyping with the Immunochip custom genotyping array (Illumina) was performed in 34 batches in 11 different centers, as described in detail elsewhere [35]. After quality control, data on a total of 156,499 SNPs became available to us through the IIBDGC and were tested for an interaction with smoking in our study. For SNPs identified as potential interaction candidates, we performed additional quality control through visual inspection of the respective cluster plots.

Only samples with known smoking status were included in our study. The 55 centers forming the IIBDGC were ranked by the number of cases available in each center. We then confined our meta-analyses to the top 10 centers which contributed a total of 12,776 cases (7076 CD, 5700 UC). Four of the ten centers also had controls with known smoking status available (N=3730).

In a previous study by the IIBDGC, focused upon the specific role of the HLA region, Immunochip high-density genotyping data from the major histocompatibility complex (MHC) region on chromosome 6 were used to impute classical HLA alleles [35]. This imputed data set comprised 11,248 variants including SNPs, HLA alleles at 4-digit resolution, HLA alleles at 2-digit resolution, and single-amino acid exchanges. For the purpose of the present study, we extracted imputed genotypes only for the 12,776 cases of interest and repeated all analyses on this data set. Further details on the imputation study, including the variant nomenclature used, can be found in the original report [35].

Statistical analysis

All statistical analyses were performed either with PLINK [36] or with the R software (v. 3.2.1), as appropriate. The statistical significance of pairwise gene-smoking interactions was assessed by logistic regression analysis as implemented in PLINK. We employed an additive allelic model of the genotype-phenotype relationship, thus encoding individual SNP genotypes (G) by allele numbers. Genotypes were treated as predictor variables whereas the binary smoking status (E) was treated as the response variable, i.e.

$$(1) \qquad logit\{P(E = 1)\} = \theta_0 + \theta \cdot G.$$

Any significant association between G and E as observed in cases points towards a gene-environment interaction at the population level, provided that the two key assumptions underlying the CO design are met, namely that (i) the disease is sufficiently rare and (ii) G and E are uncorrelated in the general population.

A two-tiered analysis was performed for CD and UC separately. In stage I, a CO analysis of gene-smoking interaction was carried out for all 156,499 SNPs in each of the top 10 center (Table 1). A separate analysis was carried out for each of the three smoking status contrasts 'never vs. ever', 'never vs. current' and 'never vs. former'. For meta-analysis, fixed- and random-effect models were fitted with PLINK. A SNP was considered worth further consideration if the meta-analysis gene-smoking interaction (Wald test) p value was smaller than 10^{-4} and the heterogeneity (Cochrane Q test) p value exceeded 0.05. It should be emphasized that these criteria were not meant to control the family-wise error rate, i.e. define a threshold for genome-wide statistical significance, but rather to serve as a sensible filter for reporting nominally significant gene-smoking interactions identified in our study.

Table 1: Overview of data used for gene-smoking interaction meta-analyses

Study center	Cases	Controls	*Smokers (%)		Male (%)
			Current	Former	
CD					
US/Los Angeles	1451	NA	11.3	8.3	52.2
Italy/Florence	1068	30	37.3	13.3	53.8
Belgium/Leuven	908	340	37.4	7.5	43.1
Germany/Kiel	714	2496	31.4	16.1	31.8
UK/Newcastle	655	NA	22.9	25.8	40.6
US/Pittsburgh	620	312	28.4	8.1	45.5
Australia/Brisbane	435	582	44.4	7.8	42.1
New Zealand/Christchurch	435	NA	25.5	23.4	35.2
UK/Exeter	428	NA	38.3	19.6	42.5
UK/London	362	NA	31.5	26.0	44.5
UC					
US/Los Angeles	791	NA	7.4	16.6	48.9
Italy/Florence	765	30	12.4	26.3	55.7
Germany/Kiel	692	2496	12.9	28.0	40.9
UK/Exeter	663	NA	15.7	38.5	52.9
UK/Newcastle	553	NA	6.5	30.2	55.1
Belgium/Leuven	516	340	21.7	29.8	56.4
US/Pittsburgh	449	312	7.8	19.1	57.5
Australia/Brisbane	447	582	19.9	28.6	47.9
New Zealand/Christchurch	425	NA	13.2	37.4	48.7
UK/Edinburgh	399	NA	9.8	43.6	46.6

NA: No control data were available from the respective center. *: Percentage refers to cases from the respective study center

To assess the robustness of our results, the stage I analyses were also carried out with adjustments made for sex and age at diagnosis (treated as a four-class ordinal variable: 0 to 20 years, 21 to 35 years, 36 to 55 years, and >55 years).

All logistic regression and meta-analyses were repeated including the SNP with the most significant gene-smoking interaction within a given 1 Mb region as a mandatory predictor variable in the analysis of all other SNPs from the same region. This was done to verify whether a given region harbored a single or multiple independent gene-smoking interactions. SNPs with a nominally significant Wald test result ($p<0.05$) in the conditional analysis were deemed independent gene-smoking interaction partners. In stage II, the validity of the G-E independence assumption underlying the CO design was checked for all SNPs identified in stage I. To this end, the logistic regression model of equation 1 was fitted to the available control data.

To allow for possible population stratification in individual study centers, we followed a recently proposed genomic control-based approach [37]. For each center, the genomic inflation factor λ was computed from a subset of 2842 'null' SNPs (chosen on the basis of GWAS of schizophrenia, psychosis, and reading and mathematics ability). Then, the standard errors of the interaction estimates were multiplied by $\sqrt{\lambda}$ before meta-analysis [38]. Population stratification was however only allowed for when λ>1 for a given combination of IBD type, study center and smoking contrast. To assess whether the gene-smoking interactions identified in our study overlapped or coincided with previously reported IBD-associated variants [34], pair-wise linkage disequilibrium (LD) was estimated from the available control samples (N=5582), irrespective of whether smoking information was available or not. To this end, r^2 was computed in each center between SNPs less than 1 Mb apart, followed by the calculation of a sample size-weighted average of the center-wise r^2. In order to identify SNPs that potentially interact differently with smoking in CD and UD, we searched for SNPs with a meta-analysis gene-smoking interaction (Wald test) p<0.01 and a heterogeneity (Cochrane Q test) p>0.05 for which the smoking odds ratio (OR) of one and the same risk allele was reversed between CD and UC (i.e. OR<1 in CD and OR>1 in UC, or vice-versa).

Results

All CO meta-analyses were performed separately for CD and UC. With the 'never vs. ever smoker' contrast, 49 SNPs for CD and 37 SNPs for UC fulfilled the two significance criteria in a fixed effects meta-analysis (Table 2). When conditioning upon the region-specific top SNP genotype, however, only two (CD) and one (UC) SNPs other than the top SNP showed a residual

Table 2: Gene-smoking interactions identified using a 'never vs. ever smoker' contrast

IBD type	Chromosome*	Top SNP	Minor allele	OR [95% CI]	p	p_h
CD	1 (1)	rs4503315	C	0.86 [0.81, 0.93]	3.8×10^{-5}	0.45
	3 (2)	rs7626254	G	1.27 [1.13, 1.42]	6.1×10^{-5}	0.22
	6 (4)	rs9262492	G	1.15 [1.08, 1.24]	7.7×10^{-5}	0.19
	6 (36)	rs3817966	G	1.21 [1.12, 1.31]	7.5×10^{-7}	0.59
	9 (4)	rs4979621	A	1.16 [1.09, 1.25]	1.7×10^{-5}	0.62
	10 (1)	rs12778349	A	0.85 [0.79, 0.92]	1.1×10^{-4}	0.71
	14 (1)	rs11625064	G	1.16 [1.08, 1.25]	8.4×10^{-5}	0.08
UC	6 (11)	rs2844776	G	0.82 [0.74, 0.90]	2.1×10^{-5}	0.24
	6 (3)	rs3129891	A	0.81 [0.73, 0.89]	1.1×10^{-5}	0.30
	7 (6)	rs10275045	A	0.86 [0.79, 0.93]	1.0×10^{-4}	0.22
	8 (11)	rs7831613	A	0.83 [0.76, 0.90]	8.1×10^{-6}	0.70
	12 (1)	rs1447876	C	0.82 [0.75, 0.91]	6.3×10^{-5}	0.82
	12 (1)	rs7958802	G	1.35 [1.17, 1.55]	4.0×10^{-5}	0.72
	12 (2)	rs6538534	G	0.85 [0.78, 0.92]	5.3×10^{-5}	0.27
	16 (1)	rs76391629	C	1.63 [1.31, 2.02]	1.0×10^{-5}	0.64
	16 (1)	rs1895539	C	0.83 [0.76, 0.91]	5.0×10^{-5}	0.72

*: Given in brackets is the number of SNPs located in a 1Mb region around the top SNP ('locus') that passed the applied significance criteria. OR, p: Exposure odds ratio and p value from a fixed-effects inverse-variance meta-analysis, based upon center-specific Wald tests; p_h: heterogeneity p value from a Cochrane Q test

gene-smoking interaction of nominal significance (Table 3). Thus, a total of nine and 10 SNPs were eventually identified as interacting with smoking for CD and UC, respectively. Random-effect analyses yielded similar results as the fixed-effect analyses (data not shown). When the G-E independence assumption underlying the CO design was checked using the available control data, three of the SNPs with a significant gene-smoking interaction (one for CD, two for UC) failed to comply with this assumption (Table 4). Meta-analyses with the 'never vs. current smoker' and 'never vs. former smoker' contrasts identified an additional 15 SNPs interacting with smoking for CD, and 11 additional SNPs were identified for UC (Table S1-S5). Taken over all three contrasts, we thus identified a total of 23 SNPs for CD and 19 SNPs for UC that showed evidence of gene-smoking interaction according to the significance criteria defined above (Table 5). Population stratification was allowed for when λ>1 for a given combination of IBD type, study center and smoking contrast. For 'never vs. ever smoker', for example, this was the case in three centers for CD, namely US/Los Angeles, Australia/Brisbane and UK/Exeter, where λ=1.10, 1.02 and 1.02, respectively, and in two centers for UC, namely UK/Newcastle (λ=1.08) and UK/Edinburgh (λ=1.11). Similar results were obtained with the 'never vs. current' and 'never vs. former' contrasts (see Table S6 for a summary of λ values). Since all λ values were small to moderate, with a maximum of 1.11 obtained in UK/Edinburgh (see above), population stratification-adjusted ORs and p values (Table 5) were found to be similar to their unadjusted counterparts.

Table 3: SNPs with a nominally significant gene-smoking interaction upon conditional analysis using a 'never vs. ever smoker' contrast

IBD type	Chromosome	SNP	Top SNP	Minor allele	OR_c [95% CI]	p_c
CD	6	rs537160	rs3817966	A	0.89 [0.82, 0.97]	6.3×10^{-3}
	6	rs9267798	rs3817966	C	0.80 [0.71, 0.91]	9.0×10^{-4}
UC	6	rs204991	rs3129891	G	0.88 [0.78, 0.99]	0.038

OR_c, p_c: Exposure odds ratio and p value as obtained after conditioning upon the top SNP in the respective region (i.e. when including the top SNP as a mandatory predictor in each logistic regression analysis)

Table 4: SNPs that showed a significant gene-smoking interaction ('never vs ever') in CO analyses but violated the G-E independence assumption in controls

Chromosome	SNP	Minor allele	OR [95% CI]	p
9	rs4979621	A	0.83 [0.74, 0.95]	2.0×10^{-3}
6	rs2844776	G	0.86 [0.74, 0.99]	0.034
6	rs204991	G	0.74 [0.64, 0.86]	1.1×10^{-4}

For details, see legend to Table 2

Table 5: SNPs with a suggestive gene-smoking interaction for at least one smoking status contrast

Chromosome	SNP	never vs. ever			never vs. current			never vs. former		
		OR [95% CI]	p	p_h	OR [95% CI]	p	p_h	OR [95% CI]	p	p_h
a) CD										
1	rs4503315	0.86 [0.81, 0.93]	3.8×10^{-5}	0.45	0.85 [0.79, 0.92]	5.7×10^{-5}	0.03	0.87 [0.79, 0.97]	0.011	0.67
1	rs12022663	0.81 [0.71, 0.92]	1.1×10^{-3}	0.71	0.73 [0.63, 0.85]	4.9×10^{-5}	0.33	1.00 [0.83, 1.21]	0.988	0.20
2	rs114129353	1.28 [0.97, 1.69]	0.086	0.04	1.11 [0.79, 1.57]	0.537	0.07	2.04 [1.42, 2.93]	1.1×10^{-4}	0.38
3	rs34166957	1.41 [1.08, 1.86]	0.013	0.03	1.27 [0.91, 1.76]	0.163	0.06	2.13 [1.45, 3.11]	1.0×10^{-4}	0.06
3	rs7626254	1.27 [1.13, 1.42]	6.1×10^{-5}	0.22	1.27 [1.12, 1.45]	2.6×10^{-4}	0.41	1.30 [1.09, 1.53]	2.8×10^{-3}	0.18
6	rs2517600	0.87 [0.80, 0.95]	1.0×10^{-3}	0.44	0.82 [0.74, 0.90]	3.2×10^{-5}	0.41	0.98 [0.87, 1.10]	0.727	0.82
6	rs1419675	1.11 [1.03, 1.20]	7.0×10^{-3}	0.42	1.19 [1.09, 1.29]	9.0×10^{-5}	0.68	0.97 [0.87, 1.09]	0.660	0.41
6	rs2517592	0.88 [0.81, 0.96]	3.3×10^{-3}	0.71	0.83 [0.75, 0.91]	9.5×10^{-5}	0.44	1.00 [0.88, 1.12]	0.942	0.90
6	rs2523734	1.15 [1.04, 1.26]	4.8×10^{-3}	0.56	1.23 [1.11, 1.37]	8.4×10^{-5}	0.83	1.00 [0.87, 1.15]	0.999	0.43
6	rs1833080	1.13 [1.04, 1.22]	2.8×10^{-3}	0.27	1.19 [1.09, 1.30]	1.1×10^{-4}	0.48	1.02 [0.91, 1.15]	0.703	0.35
6	rs113533991	0.87 [0.81, 0.94]	2.0×10^{-4}	0.60	0.83 [0.77, 0.90]	1.0×10^{-5}	0.90	0.96 [0.86, 1.06]	0.401	0.43
6	rs9262492	1.15 [1.08, 1.24]	7.7×10^{-5}	0.19	1.20 [1.11, 1.30]	6.1×10^{-6}	0.08	1.07 [0.96, 1.19]	0.207	0.50
6	rs537160	0.85 [0.78, 0.92]	5.3×10^{-5}	0.19	0.83 [0.76, 0.91]	6.7×10^{-5}	0.20	0.89 [0.79, 1.00]	0.045	0.88
6	rs4151651	0.77 [0.64, 0.92]	3.5×10^{-3}	0.14	0.63 [0.51, 0.79]	4.8×10^{-5}	0.58	1.03 [0.82, 1.31]	0.778	0.49
6	rs9267798	0.77 [0.68, 0.88]	9.3×10^{-5}	0.03	0.72 [0.62, 0.84]	1.6×10^{-5}	0.11	0.91 [0.76, 1.09]	0.310	0.58
6	rs396960	1.14 [1.06, 1.23]	7.9×10^{-4}	0.54	1.20 [1.10, 1.31]	3.8×10^{-5}	0.48	1.05 [0.94, 1.18]	0.381	0.35
6	rs482759	1.16 [1.06, 1.27]	1.7×10^{-3}	0.48	1.24 [1.12, 1.37]	3.2×10^{-5}	0.59	1.02 [0.89, 1.17]	0.779	0.16
6	rs439303	1.16 [1.07, 1.26]	3.1×10^{-4}	0.05	1.20 [1.09, 1.31]	1.0×10^{-4}	0.07	1.11 [0.98, 1.25]	0.089	0.35
6	rs3817966	1.21 [1.12, 1.31]	7.5×10^{-7}	0.59	1.24 [1.14, 1.36]	6.3×10^{-7}	0.44	1.15 [1.03, 1.29]	0.014	0.76
7	rs117675241	1.57 [1.18, 2.08]	1.8×10^{-3}	0.37	1.40 [1.00, 1.96]	0.050	0.37	2.18 [1.50, 3.18]	4.6×10^{-5}	0.46
10	rs12778349	0.85 [0.79, 0.92]	1.1×10^{-4}	0.71	0.87 [0.80, 0.95]	2.4×10^{-3}	0.48	0.82 [0.73, 0.93]	1.4×10^{-3}	0.90
14	rs11625064	1.16 [1.08, 1.25]	8.4×10^{-5}	0.08	1.16 [1.07, 1.26]	5.1×10^{-4}	0.10	1.16 [1.04, 1.29]	9.2×10^{-3}	0.75
16	rs74608019	1.32 [1.04, 1.67]	0.020	0.31	1.17 [0.89, 1.53]	0.265	0.55	1.94 [1.40, 2.69]	6.6×10^{-5}	0.14

Table 5 (continued)

b) UC										
2	rs78767684	1.46 [1.05, 2.04]	2.6×10^{-2}	0.68	2.77 [1.77, 4.35]	8.7×10^{-6}	0.77	1.36 [0.91, 2.02]	0.131	0.53
3	rs67425923	1.21 [1.09, 1.36]	5.3×10^{-4}	0.83	1.38 [1.18, 1.62]	7.2×10^{-5}	0.65	1.15 [1.02, 1.30]	0.028	0.69
5	rs75772434	1.37 [1.13, 1.66]	1.1×10^{-3}	0.06	1.71 [1.31, 2.23]	8.7×10^{-5}	0.07	1.33 [1.07, 1.65]	9.0×10^{-3}	0.02
6	rs2429657	1.08 [0.99, 1.19]	0.079	0.29	1.35 [1.18, 1.55]	9.0×10^{-6}	0.32	0.98 [0.89, 1.09]	0.759	0.49
6	rs3129891	0.81 [0.73, 0.89]	1.1×10^{-5}	0.30	0.82 [0.71, 0.95]	9.7×10^{-3}	0.6	0.80 [0.72, 0.89]	7.1×10^{-5}	0.37
7	rs10275045	0.86 [0.79, 0.93]	1.0×10^{-4}	0.22	0.88 [0.78, 0.99]	0.037	0.77	0.84 [0.77, 0.92]	1.6×10^{-4}	0.25
7	rs4721190	0.85 [0.78, 0.92]	3.4×10^{-5}	0.05	0.89 [0.79, 1.00]	0.046	0.86	0.83 [0.76, 0.91]	3.6×10^{-5}	0.06
7	rs4380850	1.13 [1.04, 1.23]	3.8×10^{-3}	0.66	0.99 [0.87, 1.13]	0.861	0.45	1.21 [1.10, 1.33]	7.3×10^{-5}	0.79
8	rs56069917	1.26 [1.09, 1.45]	2.1×10^{-3}	0.46	1.06 [0.84, 1.33]	0.638	0.83	1.37 [1.17, 1.61]	9.4×10^{-5}	0.19
8	rs7831613	0.83 [0.76, 0.90]	8.1×10^{-6}	0.70	0.83 [0.73, 0.94]	4.1×10^{-3}	0.94	0.83 [0.75, 0.91]	6.6×10^{-5}	0.43
8	rs7016774	0.83 [0.77, 0.90]	1.2×10^{-5}	0.62	0.85 [0.75, 0.96]	9.1×10^{-3}	0.96	0.83 [0.75, 0.91]	4.9×10^{-5}	0.22
12	rs1447876	0.82 [0.75, 0.91]	6.3×10^{-5}	0.82	0.86 [0.74, 1.00]	0.049	0.1	0.82 [0.73, 0.91]	1.8×10^{-4}	0.67
12	rs7958802	1.35 [1.17, 1.55]	4.0×10^{-5}	0.72	1.34 [1.08, 1.67]	7.2×10^{-3}	0.82	1.36 [1.16, 1.59]	1.6×10^{-4}	0.76
12	rs6538534	0.85 [0.78, 0.92]	5.3×10^{-5}	0.27	0.83 [0.74, 0.94]	3.2×10^{-3}	0.95	0.85 [0.78, 0.93]	5.9×10^{-4}	0.27
12	rs117614539	1.21 [0.95, 1.54]	0.124	0.17	1.93 [1.40, 2.67]	6.3×10^{-5}	0.49	1.00 [0.76, 1.34]	0.975	0.22
15	rs8026358	1.18 [0.86, 1.64]	0.308	0.71	2.57 [1.65, 4.02]	3.1×10^{-5}	0.39	0.89 [0.60, 1.32]	0.566	1.00
15	rs10518987	1.64 [1.25, 2.14]	3.4×10^{-4}	0.7	2.17 [1.52, 3.11]	2.1×10^{-5}	0.92	1.45 [1.06, 1.96]	0.019	0.59
16	rs76391629	1.63 [1.31, 2.02]	1.0×10^{-5}	0.64	1.64 [1.19, 2.26]	2.3×10^{-3}	0.31	1.65 [1.30, 2.09]	3.8×10^{-5}	0.86
16	rs1895539	0.83 [0.76, 0.91]	5.0×10^{-5}	0.72	0.82 [0.72, 0.94]	5.3×10^{-3}	0.37	0.83 [0.76, 0.92]	4.0×10^{-4}	0.83

OR, p: Exposure odds ratio and p value from fixed-effects inverse-variance meta-analysis, based upon center-specific Wald tests. Before meta-analysis, p values were individually adjusted for possible population stratification, following a genomic control approach. p_h: heterogeneity p value from a Cochrane Q test

No notably different results were obtained in the sex- and age-adjusted analyses. This finding is explicable by the fact that the CO design is sensitive to interaction but not to effects. Since smoking is associated with both sex and age at the population level, we cannot determine the influence of these two covariates on an inferred gene-smoking interaction by including them as mandatory predictors in equation 1. Therefore, we consistently chose to report unadjusted results only (except for possible adjustments for population stratification). As of the time of our study, a total of 238 IBD-associated SNPs had been identified through GWAS [34]. To assess their possible overlap with gene-smoking interactions, we quantified the level of LD between the 42 SNPs identified in our gene-smoking interaction study with those 229 SNPs for which we had genotype data available. Some 13 interacting SNPs were found to be located within 1 Mb of a GWAS SNP (Table 6; 31% overlap). However, r^2 exceeded 0.15 for only one pair of SNPs, namely rs11625064 and rs194749 on chromosome 14. Our focused analyses of classical HLA-alleles revealed a gene-smoking interaction fulfilling the chosen significance criteria for four alleles for CD and for one allele for UC (Table 7).

Finally, 10 SNPs were identified as interacting differentially with smoking in CD and UC (Table 8), indicating that the same allele of one and the same SNPs may increase the risk of CD, but at the same time be protective against UC.

Table 6: Pair-wise linkage disequilibrium between disease-associated SNPs from a previous GWAS [Liu et al. 2015] and smoking-interacting SNPs

SNP Pair		Chromosome	Position (Mb)		r^2
SNP A	SNP B		SNP A	SNP B	
rs7608910	rs114129353	2	61.20	60.97	8.2×10^{-3}
rs12994997	rs78767684	2	234.17	234.47	0.011
rs3197999	rs67425923	3	49.72	50.57	8.2×10^{-3}
rs3197999	rs34166957	3	49.72	50.61	0.011
rs6863411	rs75772434	5	141.51	141.61	2.1×10^{-3}
rs1182188	rs10275045	7	2.87	1.92	6.9×10^{-4}
rs1182188	rs4721190	7	2.87	1.95	5.6×10^{-4}
rs653178	rs117614539	12	112.01	112.36	0.020
rs194749	rs11625064	14	69.27	69.23	0.158
rs17293632	rs8026358	15	67.44	67.44	3.2×10^{-3}
rs423674	rs76391629	16	11.37	11.27	6.7×10^{-3}
rs11641184	rs76391629	16	11.70	11.27	5.8×10^{-4}
rs1728785	rs74608019	16	68.59	68.82	4.8×10^{-3}

SNP A, SNP B: GWAS and interacting SNP, respectively. r^2: sample-size-weighted measure of pair-wise linkage disequilibrium

Table 7: HLA alleles involved in gene-smoking interaction

IBD type	HLA allele	never vs. ever			never vs. current			never vs. former		
		OR [95% CI]	p	p_h	OR [95% CI]	p	p_h	OR [95% CI]	p	p_h
CD	HLA_DQA1_02	1.24 [1.13, 1.36]	1.2×10^{-5}	0.49	1.29 [1.16, 1.43]	3.5×10^{-6}	0.67	1.16 [1.01, 1.33]	0.041	0.76
	HLA_DQA1_0201	1.24 [1.13, 1.36]	1.2×10^{-5}	0.49	1.29 [1.16, 1.43]	3.5×10^{-6}	0.67	1.16 [1.01, 1.33]	0.041	0.76
	HLA_DRB1_07	1.23 [1.12, 1.35]	2.3×10^{-5}	0.51	1.27 [1.15, 1.42]	7.7×10^{-6}	0.67	1.16 [1.00, 1.33]	0.043	0.74
	HLA_DRB1_0701	1.23 [1.12, 1.35]	2.3×10^{-5}	0.51	1.27 [1.15, 1.42]	7.6×10^{-6}	0.67	1.16 [1.00, 1.33]	0.043	0.74
UC	DRB3_9101	0.77 [0.68, 0.87]	2.2×10^{-5}	0.38	0.75 [0.62, 0.9]	2.6×10^{-3}	0.74	0.79 [0.69, 0.91]	7.4×10^{-4}	0.25

For details, see legend to Table 5

Table 8: Differential gene-smoking interaction in CD and UC for the three smoking status contrasts

Smoker contrast	Chromosome	SNP	Minor allele	CD			UC			p^*
				OR	p	p_h	OR	p	p_h	
never vs. current	2	rs17022433	A	1.14 [1.05, 1.23]	1.3×10^{-3}	0.12	0.82 [0.73, 0.93]	1.6×10^{-3}	0.57	1.1×10^{-5}
	6	rs915895	G	1.13 [1.04, 1.22]	3.9×10^{-3}	0.93	0.83 [0.73, 0.94]	3.5×10^{-3}	0.98	5.9×10^{-5}
	18	rs2919450	A	1.12 [1.03, 1.23]	9.1×10^{-3}	0.86	0.83 [0.72, 0.95]	8.2×10^{-3}	0.76	2.9×10^{-4}
	19	rs34561079	A	1.23 [1.09, 1.39]	1.1×10^{-3}	0.82	0.74 [0.60, 0.92]	6.4×10^{-3}	0.64	6.5×10^{-5}
never vs. former	8	rs4403369	G	0.79 [0.66, 0.93]	4.6×10^{-3}	0.82	1.19 [1.04, 1.36]	0.010	0.23	1.4×10^{-4}
	16	rs36094971	A	1.17 [1.04, 1.32]	8.7×10^{-3}	0.75	0.87 [0.79, 0.96]	6.8×10^{-3}	0.28	1.7×10^{-4}
never vs. ever	6	rs3117098	G	0.90 [0.83, 0.97]	5.5×10^{-3}	0.46	1.12 [1.04, 1.22]	4.0×10^{-3}	0.35	6.3×10^{-5}
	6	rs3127599	A	1.11 [1.03, 1.19]	9.0×10^{-3}	0.86	0.88 [0.80, 0.95]	2.3×10^{-3}	0.70	6.0×10^{-5}
	10	rs11596541	A	0.90 [0.84, 0.97]	4.7×10^{-3}	0.86	1.12 [1.04, 1.21]	3.6×10^{-3}	0.33	5.0×10^{-5}
	16	rs79748582	G	1.12 [1.03, 1.22]	6.0×10^{-3}	0.49	0.86 [0.78, 0.94]	1.0×10^{-3}	0.12	1.9×10^{-5}

OR, p: Exposure odds ratio and p value from a fixed-effects inverse-variance meta-analysis, based upon center-specific Wald tests; p_h: heterogeneity (across study centers) p value from a Cochrane Q test; p^*: p value from a heterogeneity test (Cochrane Q test) of OR differences in the CD and UC

Discussion

Genome-wide association studies (GWAS), by definition, follow an 'agnostic', hypothesis-free approach to identify genetic variants that play a role in a given complex disease. Consequently, GWAS have pointed towards various risk genes that had not been implicated in the respective disease aetiology before, and that would therefore never have been detected in candidate (single-) gene studies alone. Provided that the cohorts under study are appropriately characterized for environmental exposure(s) of interest as well, similar arguments should also apply to the search for gene-environment interactions. Since large cohorts have been collated for genetic studies of inflammatory bowel disease (IBD) in the past, we set out to use the existing data to analyse gene-smoking interactions in CD and UC at a genome-wide level.

In so doing, we adopted a powerful but rarely used case-only (CO) design. The CO approach relies upon two key assumptions, namely that (i) the disease is sufficiently rare in the general population and that (ii) G and E are uncorrelated in the general population. Case-only studies offer a number of methodological advantages compared to case-control analyses, including a higher per-sample power and a better exposure data quality [29-31]. Since the validity of results of CO analyses depends upon the G-E independence assumption, the latter has to be checked using available control data because many genetic variants are known to be associated with smoking behaviour in the general population. Indeed, we found 7 SNPs that violated the G-E independence assumption with at least one of the three smoking status contrasts used ('never vs. ever', 'never vs. current', and 'never vs. former'). The reasons for these G-E associations are however unclear because none of the respective SNPs coincides with a previously identified smoking association.

We adopted a recently proposed genomic control approach to allow for population stratification in individual centers. Since all λ values were small to moderate, with a maximum of 1.11 obtained in UK/Edinburgh, population stratification-adjusted ORs and p values were found to be similar to their unadjusted counterparts.

In our extensive meta-analyses, we identified 23 SNPs for CD and 19 SNPs for UC that interact with at least one of three smoking status contrasts ('never vs. ever', 'never vs. current', and 'never vs. former'). We observed that some 30% of the SNPs interacting with smoking lie in close vicinity (<1Mb) to SNPs identified as disease-associated in previous GWAS. However, owing to the lack of substantial LD between them, we may conclude that the respective association and interaction signals have different causes. In other words, even if they belong to the same functional unit, the disease risk of some of the genetic variants is modified by smoking whereas the effect of others is not. Along the same vein, in our specific analysis of the MHC region, only a subset of the HLA alleles recently shown to have a main effect [35] was found to interact with smoking as well.

We identified 10 SNPs that differentially interact with smoking in CD and UC. Since statistical interaction can be viewed from different angles, such difference may mean one of two things. Either the disease predisposing effect of a given genetic variant is rendered protective by the presence of an environmental stimulus (i.e. smoking), or the environmental effect is reversed in the presence of genetic variant. In the present case, this means that a certain genotype may render smoking a risk factor for

CD but protective against UC. The present study will add to an improved understanding of biological mechanisms underlying IBD which is required to develop new preventive strategies and to improve diagnostic and therapeutic measures [33]. A limitation of our study may be that the definition of 'smoking' may vary between centers and countries. Smoking status was assessed by country-specific questionnaires or interviews.

In summary, using a CO approach, we were able to identify 42 SNPs with strong evidence of an interaction with smoking in IBD. None of these polymorphisms had been reported as being involved in a gene-smoking interaction in IBD before. In addition, several nominally significant interactions were observed with other SNPs that failed to pass the significance criteria used for reporting in our study. The lack of overlap between the 42 strongly suggestive SNPs and known IBD-associated markers further highlights the potential of an agnostic, hypothesis free genome-wide search for gene-environment interactions. Given the hitherto observed effect sizes, however, such studies may even have to be larger than the present one, particularly when following a genome-wide approach. In any case, further functional and experimental studies are required to fully clarify the role of the identified gene-smoking interactions in IBD aetiology.

Acknowledgments

This study was carried out as part of the Research Training Group 'Genes, Environment and Inflammation', supported by the Deutsche Forschungsgemeinschaft (GRK 1743). The data was collected at the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC).

References

1. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. 2010;42(12):1118-25.

2. Franke A, Balschun T, Sina C, Ellinghaus D, Hasler R, Mayr G, et al. Genome-wide association study for ulcerative colitis identifies risk loci at 7q22 and 22q13 (IL17REL). Nat Genet. 2010;42(4):292-4.

3. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, Taylor KD, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet. 2011;43(3):246-52.

4. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491(7422):119-24. doi: 10.1038/nature11582.

5. Molodecky NA, Panaccione R, Ghosh S, Barkema HW, Kaplan GG. Challenges associated with identifying the environmental determinants of the inflammatory bowel diseases. Inflamm Bowel Dis. 2011;17(8):1792-9. Epub 2011/07/12. doi: 10.1002/ibd.21511. PubMed PMID: 21744435.

6. Mahid SS, Minor KS, Soto RE, Hornung CA, Galandiuk S. Smoking and inflammatory bowel disease: a meta-analysis. Mayo Clin Proc. 2006;81(11):1462-71.

7. Higuchi LM, Khalili H, Chan AT, Richter JM, Bousvaros A, Fuchs CS. A prospective study of cigarette smoking and the risk of inflammatory bowel disease in women. Am J Gastroenterol. 2012;107(9):1399-406.

8. Garcia Rodriguez LA, Gonzalez-Perez A, Johansson S, Wallander MA. Risk factors for inflammatory bowel disease in the general population. Aliment Pharmacol Ther. 2005;22(4):309-15. Epub 2005/08/16. doi: 10.1111/j.1365-2036.2005.02564.x. PubMed PMID: 3622939.

9. Birrenbach T, Bocker U. Inflammatory bowel disease and smoking: a review of epidemiology, pathophysiology, and therapeutic implications. Inflamm Bowel Dis. 2004;10(6):848-59.

10. Thomas GA, Rhodes J, Green JT, Richardson C. Role of smoking in inflammatory bowel disease: implications for therapy. Postgrad Med J. 2000;76(895):273-9.

11. Cosnes J. Tobacco and IBD: relevance in the understanding of disease mechanisms and clinical practice. Best Pract Res Clin Gastroenterol. 2004;18(3):481-96.

12. Parkes GC, Whelan K, Lindsay JO. Smoking in inflammatory bowel disease: Impact on disease course and insights into the aetiology of its effect. J Crohns Colitis. 2014; 8(8):717-25. doi: 10.1016/j.crohns.2014.02.002. PubMed PMID: 24636140.

13. Kikuchi H, Itoh J, Fukuda S. Chronic nicotine stimulation modulates the immune response of mucosal T cells to Th1-dominant pattern via nAChR by upregulation of Th1-specific transcriptional factor. Neurosci Lett. 2008;432(3):217-21.

14. Bergeron V, Grondin V, Rajca S, Maubert MA, Pigneur B, Thomas G, et al. Current smoking differentially affects blood mononuclear cells from patients with crohn's disease and ulcerative colitis: relevance to its adverse role in the disease. Inflamm Bowel Dis. 2012;18(6):1101-11.

15. Andersen V, Nimmo E, Krarup HB, Drummond H, Christensen J, Ho GT, et al. Cyclooxygenase-2 (COX-2) polymorphisms and risk of inflammatory bowel disease in a Scottish and Danish case-control study. Inflamm Bowel Dis. 2011;17(4):937-46.

16. Biedermann L, Zeitz J, Mwinyi J, Sutter-Minder E, Rehman A, Ott SJ, et al. Smoking cessation induces profound changes in the composition of the intestinal microbiota in humans. PLoS One. 2013;8(3):e59260.

17. Biedermann L, Brulisauer K, Zeitz J, Frei P, Scharl M, Vavricka SR, et al. Smoking cessation alters intestinal microbiota: insights from quantitative investigations on human fecal samples using FISH. Inflamm Bowel Dis. 2014;20(9):1496-501. Epub 2014/07/30. doi: 10.1097/mib.0000000000000129.

18. Benjamin JL, Hedin CR, Koutsoumpas A, Ng SC, McCarthy NE, Prescott NJ, et al. Smokers with active Crohn's disease have a clinically relevant dysbiosis of the gastrointestinal microbiota. Inflamm Bowel Dis. 2012;18(6):1092-100.

19. Klareskog L, Stolt P, Lundberg K, Kallberg H, Bengtsson C, Grunewald J, et al. A new model for an etiology of rheumatoid arthritis: smoking may trigger HLA-DR (shared epitope)-restricted immune reactions to autoantigens modified by citrullination. Arthritis Rheum. 2006;54(1):38-46.

20. Wegner N, Lundberg K, Kinloch A, Fisher B, Malmstrom V, Feldmann M, et al. Autoimmunity to specific citrullinated proteins gives the first clues to the etiology of rheumatoid arthritis. Immunol Rev. 2010;233(1):34-54.

21. Mahdi H, Fisher BA, Kallberg H, Plant D, Malmstrom V, Ronnelid J, et al. Specific interaction between genotype, smoking and autoimmunity to citrullinated alpha-enolase in the etiology of rheumatoid arthritis. Nat Genet. 2009;41(12):1319-24.

22. Wagner CA, Sokolove J, Lahey LJ, Bengtsson C, Saevarsdottir S, Alfredsson L, et al. Identification of anticitrullinated protein antibody reactivities in a subset of anti-CCP-negative rheumatoid arthritis: association with cigarette smoking and HLA-DRB1 'shared epitope' alleles. Ann Rheum Dis. 2015;74(3):579-86. Epub 2013/12/04. doi: 10.1136/annrheumdis-2013-203915. PubMed PMID: 24297382.

23. Andersen V, Vogel U. Systematic review: interactions between aspirin, and other nonsteroidal anti-inflammatory drugs, and polymorphisms in relation to colorectal cancer. Aliment Pharmacol Ther. 2014;40(2):147-59. doi: 10.1111/apt.12807.

24. Andersen V, Holst R, Vogel U. Systematic review: diet-gene interactions and the risk of colorectal cancer. Aliment Pharmacol Ther. 2013;37(4):383-91. doi: 10.1111/apt.12180.

25. Andersen V, Vogel U. Interactions between meat intake and genetic variation in relation to colorectal cancer. Genes Nutr. 2015;10(1):448. doi: 10.1007/s12263-014-0448-9. PubMed PMID: 25491747.

26. Helbig KL, Nothnagel M, Hampe J, Balschun T, Nikolaus S, Schreiber S, et al. A case-only study of gene-environment interaction between genetic susceptibility variants in NOD2 and cigarette smoking in Crohn's disease aetiology. BMC Med Genet. 2012;13:14. doi: 10.1186/1471-2350-13-14. PubMed PMID: 22416979.

27. Fowler EV, Doecke J, Simms LA, Zhao ZZ, Webb PM, Hayward NK, et al. ATG16L1 T300A shows strong associations with disease subgroups in a large Australian IBD population: further support for significant disease heterogeneity. Am J Gastroenterol. 2008;103(10):2519-26.

28. Doecke JD, Simms LA, Zhao ZZ, Roberts RL, Fowler EV, Croft A, et al. Smoking behaviour modifies IL23r-associated disease risk in patients with Crohn's disease. J Gastroenterol Hepatol. 2015;30(2):299-307. doi: 10.1111/jgh.12674.

29. 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(2):153-62. Epub 1994/01/30.

30. Gatto NM, Campbell UB, Rundle AG, Ahsan H. Further development of the case-only design for assessing gene-environment interaction: evaluation of and adjustment for bias. Int J Epidemiol. 2004;33(5):1014-24. doi: 10.1093/ije/dyh306. PubMed PMID: 15358745.

31. Yadav P, Freitag-Wolf S, Lieb W, Krawczak M. The role of linkage disequilibrium in case-only studies of gene-environment interactions. Hum Genet. 2015;134(1):89-96. doi: 10.1007/s00439-014-1497-2. PubMed PMID: 25304818.

32. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature. 2008;452(7187):638-42. doi: 10.1038/nature06846.

33. Andersen VC, Vogel U. [Inflammatory bowel diseases--can genetic studies assist in improving treatment?]. Ugeskr Laeger. 2009;171(47):3448-51.

34. Liu JZ, van Sommeren S, Huang H, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015;47(9):979-86. doi: 10.1038/ng.3359.

35. Goyette P, Boucher G, Mallon D, et al. High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. Nat Genet. 2015;47(2):172-9. doi: 10.1038/ng.3176.

36. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75.

37. Yadav P, Freitag-Wolf S, Lieb W, Dempfle A, Krawczak M. Allowing for population stratification in case-only studies of gene-environment interaction, using genomic control. Hum Genet. 2015;134(10): 1117-25. doi: 10.1007/s00439-015-1593-y.

38. Wang S, Chen W, Chen X, Hu F, Archer KJ, Liu HN, Sun S, Gao G. Double genomic control is not effective to correct for population stratification in meta-analysis for genome-wide association studies. Front Genet. 2012;3:300. doi: 10.3389/fgene.2012.00300.

Supplementary Tables

Table S1: Gene-smoking interactions identified using a 'never vs. current smoker' contrast

IBD type	Chromosome[*]	Top SNP	Minor allele	OR [95% CI]	p	p_h
CD	1	rs12022663	G	0.73 [0.63, 0.85]	4.9×10^{-5}	0.33
	6(24)	rs113533991	A	0.83 [0.77, 0.90]	1.0×10^{-5}	0.90
	6 (43)	rs3817966	G	1.24 [1.14, 1.36]	6.3×10^{-7}	0.44
	9	rs4979621	A	1.17 [1.08, 1.27]	7.7×10^{-5}	0.42
UC	2	rs78767684	C	2.77 [1.77, 4.35]	8.7×10^{-6}	0.77
	3	rs67425923	A	1.38 [1.18, 1.62]	7.2×10^{-5}	0.65
	5 (2)	rs75772434	T	1.71 [1.31, 2.23]	8.7×10^{-5}	0.07
	6 (10)	rs2429657	G	1.35 [1.18, 1.55]	9.0×10^{-6}	0.32
	12	rs117614539	A	1.93 [1.40, 2.67]	6.3×10^{-5}	0.49
	15	rs8026358	A	2.57 [1.65, 4.02]	3.1×10^{-5}	0.39
	15	rs10518987	A	2.17 [1.52, 3.11]	2.1×10^{-5}	0.92

*: Given in brackets is the number of SNPs located in a 1Mb region around the top SNP ('locus') that passed the applied significance criteria. OR, p: Exposure odds ratio and p value from a fixed-effects inverse-variance meta-analysis, based upon center-specific Wald tests; p_h: heterogeneity p value from a Cochrane Q test

Table S2: SNPs with a nominally significant gene-smoking interaction upon conditional analysis using a 'never vs. current smoker' contrast

IBD type	Chromosome	SNP	Top SNP	Minor allele	OR_c [95% CI]	p_c
CD	6	rs2517600	rs113533991	A	0.86 [0.78, 0.95]	3.1×10^{-3}
	6	rs1419675	rs113533991	C	1.12 [1.02, 1.23]	0.019
	6	rs2517592	rs113533991	A	0.89 [0.80, 0.99]	0.030
	6	rs2523734	rs113533991	C	1.17 [1.05, 1.30]	5.6×10^{-3}
	6	rs1833080	rs113533991	A	1.11 [1.00, 1.23]	0.041
	6	rs9262492	rs113533991	G	1.16 [1.07, 1.26]	3.9×10^{-4}
	6	rs4151651	rs3817966	A	0.67 [0.54, 0.84]	3.9×10^{-4}
	6	rs537160	rs3817966	A	0.88 [0.80, 0.96]	5.8×10^{-4}
	6	rs9267798	rs3817966	C	0.75 [0.64, 0.87]	1.7×10^{-4}
	6	rs396960	rs3817966	A	1.19 [1.10, 1.30]	5.4×10^{-5}
	6	rs482759	rs3817966	G	1.20 [1.08, 1.33]	6.1×10^{-4}
	6	rs439303	rs3817966	A	1.18 [1.08, 1.29]	3.2×10^{-4}
	6	rs3104389	rs3817966	A	1.13 [1.00, 1.27]	0.042
	6	rs3104407	rs3817966	G	0.89 [0.82, 0.97]	8.9×10^{-3}

OR_c, p_c: Exposure odds ratio and p value as obtained after conditioning upon the top SNP in the respective region (i.e. when including the top SNP as a mandatory predictor in each logistic regression analysis)

Table S3: SNPs that showed a significant gene-smoking interaction ('never vs current') in CO analyses but violated the G-E independence assumption in controls

Chromosome	SNP	Minor allele	OR [95% CI]	p
6	rs3104389	A	0.77 [0.62, 0.95]	0.014
6	rs3104407	G	1.20 [1.01, 1.42]	0.035
9	rs4979621	A	0.82 [0.69, 0.98]	0.029

For details, see legend to Table S1

Table S4: Gene-smoking interactions identified using a 'never vs. former smoker' contrast

IBD type	Chromosome*	Top SNP	Minor allele	OR [95% CI]	p	p_h
CD	2	rs114129353	A	2.04 [1.42, 2.93]	1.1×10^{-4}	0.38
	3	rs34166957	A	2.13 [1.45, 3.11]	1.0×10^{-4}	0.06
	4	rs11131613	G	1.25 [1.12, 1.39]	5.5×10^{-5}	0.16
	7	rs117675241	A	2.18 [1.50, 3.18]	4.6×10^{-5}	0.46
	16 (2)	rs74608019	A	1.94 [1.40, 2.69]	6.6×10^{-5}	0.14
UC	6	rs3129891	A	0.80 [0.72, 0.89]	7.1×10^{-5}	0.37
	7(6)	rs4721190	A	0.83 [0.76, 0.91]	3.6×10^{-5}	0.06
	7	rs4380850	G	1.21 [1.10, 1.33]	7.3×10^{-5}	0.79
	8(3)	rs56069917	A	1.37 [1.17, 1.61]	9.4×10^{-5}	0.19
	8(10)	rs7016774	T	0.83 [0.75, 0.91]	4.9×10^{-5}	0.22
	16	rs76391629	C	1.65 [1.30, 2.09]	3.8×10^{-5}	0.86

For details, see legend to Table S1

Table S5: SNPs that showed a significant gene-smoking interaction ('never vs former') in CO analyses but violated the G-E independence assumption in controls

Chromosome	SNP	Minor allele	OR [95% CI]	p
4	rs11131613	G	0.87 [0.75, 1.00]	0.055
6	rs3129891	A	0.85 [0.71, 1.00]	0.053

For details, see legend to Table S1

Table S6: Center-wise summary of λ values for the three smoking status contrasts

Study center	never vs. ever	never vs. current	never vs. current
CD			
US/Los Angeles	1.100	0.996	1.052
Italy/Florence	0.985	0.961	1.032
Belgium/Leuven	0.966	0.972	1.048
Germany/Kiel	0.947	0.974	0.938
UK/Newcastle	0.925	1.032	0.957
US/Pittsburgh	0.869	0.944	0.956
Australia/Brisbane	1.018	1.054	0.999
New Zealand/Christchurch	0.980	0.974	0.966
UK/Exeter	1.016	1.015	1.007
UK/London	0.908	1.006	1.014
UC			
US/Los Angeles	0.938	0.928	1.021
Italy/Florence	0.941	0.968	0.969
Germany/Kiel	0.887	0.943	0.916
UK/Exeter	0.923	0.962	0.889
UK/Newcastle	1.077	0.957	1.044
Belgium/Leuven	0.872	0.859	0.838
US/Pittsburgh	0.903	1.096	0.840
Australia/Brisbane	0.999	0.880	0.952
New Zealand/Christchurch	0.888	0.959	0.879
UK/Edinburgh	1.114	0.896	1.099

Chapter 3

Discussion

3.1 The Improved CO Approach

This work was largely motivated by our current enthusiasm for studying gene-environment interactions (G×E). One major challenge to G×E studies is the insufficient power of traditional epidemiological study designs including case-control and cohort studies. This thesis work dealt with a controversial albeit potentially efficient the case-only (CO) design and resolved two important issues to successfully employ this approach for genome-wide G×E studies. Publication (i) illustrated the role of linkage disequilibrium (LD) in CO studies of G×E. Publication (ii) was the first study to employ genomic control (GC) to correct for population stratification (PS) in the context of CO studies of G×E. Finally, Publication (iii) presented a real data application of the improved CO approach using an IBD data set. The key findings of this work have been already discussed in their respective publication; here they are revised together with some additional aspects.

3.1.1 LD in CO Studies of G×E

Publication (i) elucidated the impact of LD and physical distance on the power to detect G×E in CO studies of proxy SNPs. It was shown, with computer simulation and a real data example, that SNPs in LD with a truly interacting SNP can be used as proxies to indirectly detect the respective G×E signal. The power to detect G×E through proxies was found to be higher when the truly underlying interaction was strong. Although these findings seem to be expected in the light of previous observations from GWAS paradigm per se, it has to be noted that G×E studies follow a different strategy in principle. In GWAS, both causal and non-causal SNPs (i.e. proxies) in LD with one another are expected to exhibit a certain degree of disease association. This means that a causal SNP does not have to be tested itself in order to highlight its position in the genome. In fact, all GWAS published to date rely upon a manageable subset of markers that are assessed using genotyping arrays. Contrarily, in G×E studies, an additional non-genetic factor (i.e. environment) plays an important role, and it was by no means clear how the peculiarities of the genetic effect impact upon the power to detect the interaction. Simply stated, in G×E studies there is one link more to the chain that relates the proxy SNP to the epidemiological effect of interest - which is ‘interaction’, not disease association. Therefore, Publication (i) contributed significantly to our current understanding of the role of LD in G×E studies. It is clear now that SNPs in LD with a truly interacting SNP are capable of indirectly detecting the underlying G×E signal. In the future, genome-wide G×E studies will benefit from this revelation particularly in instances where the truly interacting variant is unknown or remained missing due to systematic and technical reasons.

3.1.2 PS in CO Studies of G×E

PS and LD are two closely related issues and the former can lead to the latter if allele frequencies at two or more loci differ among the underlying subpopulations. One example of this can be seen in Africa where the high levels of population substructure resulted in the observed divergent patterns of LD among African subpopulations [55]. In Publication (i), we learned that LD information is as fundamental to CO studies of G×E as to GWAS. Publication (ii) showed that assessment of PS is of interest to obtain reliable detection of G×E following the CO approach. Knowledge of PS also helps to quantify the differences in LD patterns between populations [56]. That is, if G×E found in one population is not replicated in another, it could be due to differences in the LD pattern between the two populations.

In Publication (ii), a qualitative assessment of the impact of PS in CO studies of G×E was performed and various means to correct for PS were explored. One important finding from this paper was that CO studies of G×E are not affected by PS as long as the frequency of only one factor (i.e. either G or E) differs between subpopulations. However, without further adjustment, the regular Wald test used in CO approach was found to be invalid under joint stratification by G and E factors (PS_{GE}), with seriously inflated type I error rates, as expected. This outcome is explicable by the fact that PS_{GE} creates an association at the population level between G and E and thus violates the key assumption of the CO design. It was shown that modified Wald test statistics based on the GC approach could be adopted to correct for PS in CO studies. However, the GC approach was found to be conservative particularly at high level of PS. Consequently, the modified Wald test statistics suffered from a loss of power to detect G×E under high level of PS. In addition, it was also shown that family-based approaches such as an extension of TDT [57] could be employed to correct for PS. However, in the context of G×E studies it would be difficult to obtain the relevant data from patients’ relatives, particularly for disease instances with long latent periods. In summary, Publication (ii) added significantly to our current understanding of the role of PS in CO studies of G×E and showed that the adaptation of the GC paradigm from gene-disease association studies can be extended to G×E to rectify this problem.

3.1.3 IBD Data Example

In Publication (iii), the improved CO approach was successfully applied to a real IBD data set to assess genome-wide gene-smoking interaction in BID. This study used 10 Immunochip data sets collated by the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC), comprising 12,776 cases of known smoking status. However, control samples with information on smoking history were available only from four of the participating centers (N=3730). Therefore, a CO study was preferred over a case-control study to make use of all the available case samples. Moreover, CO approach has been deemed superior to case-control studies by many scholars because of its greater per-sample power [36–38]. The CO analysis was performed center-wise to reduce the impact of PS. In addition, the GC-based approach as proposed in Publication (ii) was followed to allow for PS within each study center. The PS observed in this example was only small to moderate (with maximum λ=1.11), as expected. This is explicable by the fact that the PS can impact CO studies if the study population is stratified with respect to both G and E (i.e. if there is PS_{GE}) which was very less likely to occur within each center. However, if cases from different centers were pooled into one sample (i.e. mega-analysis), λ could be higher. This is because the smoking prevalence was found to be varying across the study centers and genetic differences across the centers are expected as well, thus creating a likely scenario for PS_{GE}.

Furthermore, the LD information was exploited in the study of IBD data set mainly in two ways. First, conditional analysis was done to verify whether a given region harbored a single or multiple independent gene-smoking interactions. SNPs with a nominally significant Wald test result ($p<0.05$) in the conditional analysis were identified as independent gene-smoking interaction partners. All logistic regression and meta-analyses were repeated with the SNP showing the most significant gene-smoking interaction (i.e. smallest p value) within a given 1 Mb region included as a mandatory predictor variable in the analysis of other SNPs in the same region. Although not exactly the same, the underlying principle of the conditional analysis is similar to that of LD. Second, to assess whether the genetic risks factors modified by smoking that were identified in this study overlap or coincide with previously reported IBD-associated variants [58], pair-wise LD was estimated in all available control samples. It was observed that some 30% of the SNPs interacting with smoking lie in close vicinity (<1Mb) to SNPs identified as disease-associated in previous GWAS. However, owing to the lack of substantial LD between them, it may conclude that the respective association and interaction signals have different causes. In other words, even if they belong to the same functional unit, the disease risk of some genetic variants is modified by smoking whereas the effect of others is not. The majority of these markers (16 SNPs) were located on chromosome 6, thus indicating a potential role of the HLA region in IBD. In fact, a focused analysis of classical HLA alleles revealed a suggestive gene-smoking interaction for four alleles in CD and for one allele in UC. Noteworthy, several interaction effect differences were observed between CD and UC thereby suggesting a differential role of tobacco smoking in the etiology of IBD. Since statistical interaction can be viewed from different angles, such difference may mean one of two things. Either the disease predisposing effect of a given genetic variant is rendered protective by the presence of an environmental stimulus (i.e. smoking), or the environmental effect is reversed in the presence of genetic variant. In the present case, this means that a certain genotype may render smoking a risk factor for CD but protective against UC.

3.2 Strengths and Limitations

The main analyses work in Publication (i) and Publication (ii) relied upon simulated data set. Often, it is of interest to evaluate the performance of a given methodology under certain scenarios. Simulations are clearly preferable to assess the effectiveness of statistical methods since the relevant parameters influencing the disease risk are known in advance. In the context of G×E studies, utility of simulated data set is of high importance since publicly available data sets comprising genome-wide data and reliable exposure information are rare - rendering the validation of novel G×E approaches difficult.

This work followed a convenient and inexpensive way to obtain valid and reliable answers to our research queries. In Publication (i), simulations were used to systematically vary the level of LD between an interacting SNP and a proxy SNP. In addition, real haplotypes from HapMap were taken to simulate the haplotypes with a realistic LD pattern. The outcomes obtained on simulated data sets were also validated on a real colorectal cancer data set. The consistent finding in both instances was that the level of LD present strongly affects the power of a proxy SNP to detect a given G×E. In Publication (ii), the entire work relied upon the simulated data sets. CO data sets were simulated under different PS scenarios in order to assess the performances of modified Wald test statistics. Simulations turned out to be very fruitful for this study because publicly available data sets comprising genome-wide data and reliable exposure information are rare, and for the few that exist, the degree of underlying PS is usually unknown. Moreover, trios (cases and their parent) data sets were also simulated to verify whether family-based methods such as an extension of TDT can be robust against PS in G×E studies. To perform such studies using real data is difficult in the first place because one needs to collect genetic and exposure data on cases plus the genetic data on their biological parents. Such data are unlikely to exist in practice, given that many of the diseases with environmental component are of late-onset. The use of real data could be illustrative, and therefore beneficial, in many ways. Publication (iii) used a real IBD data set to illustrate the genome-wide usability of the improved CO approach. Therefore, this work gained its strength from both the simulation and the application of real data set to support the simulation outcomes.

Although potentially useful, simulation frameworks usually capture only specific scenarios, but reality is often more complex. Typically, the performance of a given methodology can be influenced by many variables including sample size, the underlying model of inheritance, the allelic frequencies, the distributions of the environmental factors, and the relative strength of the different factors affecting the risk of disease. Often, it is difficult to cover all possible realistic scenarios within a single simulation study. Publication (ii), for instance, considered relatively simple PS scenarios involving two subpopulations but in practice underlying PS might involve more than two subpopulations or even admixtures in a given sample of cases. Furthermore, both Publication (i) and Publication (ii) considered the simplest (binary) type of the environmental exposure. However, many relevant environmental factors such as diet, physical activity, and air pollution parameters are multidimensional and difficult to measure. Moreover, environmental exposures can change over time and are not always measured at the relevant time period. Measurement at baseline or at interview may not reflect the relevant windows of exposure and will not reflect lifetime exposure. Simulating such complexities can be more challenging and computationally expensive. Therefore, the CO approach might be underpowered in practice due to these inevitable variations.

In a few simulation settings, relatively large interaction effects (i.e. G×E OR=7.39) were considered, but such effects are unlikely to exist in reality. Alternatively, one can chose smaller interaction effects, but this would require large sample sizes to be simulated in order to achieve similar qualitative assessment. The real data application was missing in the Publication (ii). The stratified analysis used to benchmark the GC-based statistics idealistically assumed that the true subpopulation affiliation was known. However, PS is usually hidden and true population affiliation unknown. The GC-based approach has the disadvantage of assuming that stratification creates uniform inflation across the genome, this might bias test statistics conservatively in some regions and freely in other regions. Beyond this, the CO approach is of great interest when objective is to assess the interaction, but the main effects cannot be measured with this approach.

3.3 Conclusion and Outlook

Selection of an optimal study design is crucial for conducting successful G×E studies like any other epidemiologic study. CO studies trade on the fact that cases are usually much easier to recruit and characterize in terms of their phenotype and environmental exposure than controls, particularly in retrospective studies. This work resolved two important issues in order to facilitate the usability of CO approach on genome-wide level. First, it is clear now that SNPs in LD with the truly interacting SNP are capable of indirectly detecting the underlying G×E signal (section 2.1). As in GWAS, genome-wide G×E studies can benefit from this outcome particularly in instances where the truly interacting variant is not genotyped. Second, it was shown that the genomic control (GC) paradigm from GWAS can be adopted to correct for PS in CO studies of G×E (section 2.2). Moreover, family-based methods such as extension of TDT can be employed to account for the PS. However, the feasibility of family-based approaches in the context of G×E studies is warranted as it would be difficult to obtain the relevant data from patients' biological relatives, particularly for disease instances with late-onset. The use of real data set throughout this work was illustrative. The work in Publication (i) and Publication (ii) provided valid answers to the questions raised in the introduction (section 1.4). Moreover, the work in Publication (iii) benefited from the outcomes of other two publications. For instance, the GC-based approach to correct for PS in CO studies as proposed in Publication (ii) was applied in the analysis of IBD data set in Publication (iii).

The improved CO approach will motivate the genome-wide detection of G×E. The G×E studies have several implications for future research. Most importantly, G×E studies are believed to shed light on what has been termed 'dark matter' of GWAS [22]. Furthermore, knowledge of G×E might improve our abilities to further develop personalized medicine based applications by targeting the individual's genetic and life-long exposure information. The detection of statistical interactions provides a good starting point for a more focused investigation of the joint involvement of the relevant factors, which can potentially be addressed and replicated in other types of experimental data [59–61]. For instance, in Publication (iii), using CO approach, 42 SNPs were identified to show suggestive interaction with smoking in IBD. None of these polymorphisms had been reported as being involved in a gene-smoking interaction in IBD before. Undoubtedly, further functional and experimental studies are required to fully clarify the role of identified gene-smoking interactions in IBD etiology.

There are a few issues that were not covered in this work and could be interesting to explore in the future. For instance, the effect of genotype imputation [62–64] on the validity and power of adjusted and unadjusted test statistics used in the CO studies can be explored in the future. Furthermore, the effect of joint genotype and exposure misclassification on the CO studies of G×E could be explored [65]. Moreover, comparison of the CO approach with other methods to detect G×E such as the machine learning based approaches [66] can be performed to verify whether different methods lead to similar outcomes or not.

The two issues addressed in this work have led to a better understanding of the CO design for performing genome-wide G×E studies. Possibly, some of the 'missing heritability' of common complex disease could be elucidated in the future employing this improved design. In summary, this work implies that future research of G×E may safely adopt a CO approach to exploit existing GWAS data sets. At the same time, this thesis work suggests that relevant data resources should aim at comprising large numbers of cases for whom genetic and environmental exposure data may be easier to obtain than for controls or patient relatives.

Bibliography

[1] Blot WJ, McLaughlin JK, Winn DM, et al. (1988). Smoking and drinking in relation to oral and pharyngeal cancer. Cancer Res 48:3282–7.

[2] Hart CL, Davey Smith G, Gruer L, Watt GC (2010). The combined effect of smoking tobacco and drinking alcohol on cause-specific mortality: a 30 year cohort study. BMC Public Health 10:789. doi: 10.1186/1471-2458-10-789

[3] Purdue MP, Hashibe M, Berthiller J, et al. (2009). Type of alcoholic beverage and risk of head and neck cancer—a pooled analysis within the INHANCE Consortium. Am J Epidemiol 169:132–42. doi: 10.1093/aje/kwn306

[4] García-Closas M, Malats N, Silverman D, et al. (2005). NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. Lancet (London, England) 366:649–59. doi: 10.1016/S0140-6736(05)67137-1

[5] Hwang SJ, Beaty TH, Panny SR, et al. (1995). Association study of transforming growth factor alpha (TGF alpha) TaqI polymorphism and oral clefts: indication of gene-environment interaction in a population-based sample of infants with birth defects. Am J Epidemiol 141:629–636.

[6] Siegert S, Hampe J, Schafmayer C, et al. (2013). Genome-wide investigation of gene-environment interactions in colorectal cancer. Hum Genet 132:219–31. doi: 10.1007/s00439-012-1239-2

[7] Molodecky NA, Kaplan GG (2010). Environmental risk factors for inflammatory bowel disease. Gastroenterol Hepatol (N Y) 6:339–46.

[8] Vineis P, Marinelli D, Autrup H, et al. (2001). Current smoking, occupation, N-acetyltransferase-2 and bladder cancer: a pooled analysis of genotype-based studies. Cancer Epidemiol Biomarkers Prev 10:1249–52.

[9] Hunter DJ (2005). Gene-environment interactions in human diseases. Nat Rev Genet 6:287–298. doi: 10.1038/nrg1578

[10] Genes and Environment Initiative (GEI). http://www.genome.gov/19518663#al-1. Accessed 6 Sep 2015

[11] German National Cohort (GNC) Consortium (2014). The German National Cohort: aims, study design and organization. Eur J Epidemiol 29:371–82. doi: 10.1007/s10654-014-9890-7

[12] Deutsche Forschungsgemeinschaft (DFG). http://www.dfg.de. Accessed 6 Sep 2015

[13] Lander ES, Schork NJ (1994). Genetic dissection of complex traits. Science 265:2037–48.

[14] Visscher PM, Brown MA, McCarthy MI, Yang J (2012). Five years of GWAS discovery. Am J Hum Genet 90:7–24. doi: 10.1016/j.ajhg.2011.11.029

[15] 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. doi: 10.1038/nature05911

[16] Risch N, Merikangas K (1996). The future of genetic studies of complex human diseases. Science 273:1516–7.

[17] Hindorff LA, Sethupathy P, Junkins HA, et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106:9362–7. doi: 10.1073/pnas.0903103106

[18] Chakravarti A (1999). Population genetics—making sense out of sequence. Nat Genet 21:56–60.

[19] Gray IC, Campbell DA, Spurr NK (2000). Single nucleotide polymorphisms as tools in human genetics. Hum Mol Genet 9:2403–2408. doi: 10.1093/hmg/9.16.2403

[20] Manolio TA, Brooks LD, Collins FS (2008). A HapMap harvest of insights into the genetics of common disease. J Clin Invest 118:1590–605. doi: 10.1172/JCI34772

[21] Klein RJ, Zeiss C, Chew EY, et al. (2005). Complement factor H polymorphism in age-related macular degeneration. Science 308:385–9. doi: 10.1126/science.1109557

[22] Manolio TA, Collins FS, Cox NJ, et al. (2009). Finding the missing heritability of complex diseases. Nature 461:747–53. doi: 10.1038/nature08494

[23] McCarthy MI, Abecasis GR, Cardon LR, et al. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356–369. doi: 10.1038/nrg2344

[24] Björkegren JL, Kovacic JC, Dudley JT, Schadt EE (2015). Genome-wide significant loci: how important are they? Systems genetics to understand heritability of coronary artery disease and other common complex disorders. J Am Coll Cardiol 65:830–45. doi: 10.1016/j.jacc.2014.12.033

[25] Eichler EE, Flint J, Gibson G, et al. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–50. doi: 10.1038/nrg2809

[26] Aschard H, Lutz S, Maus B, et al. (2012). Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. Hum Genet 131:1591–613. doi: 10.1007/s00439-012-1192-0

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

[28] Rothman KJ, Greenland S, Lash TL (2008). Modern Epidemiology, 3rd Edition. Lippincott Williams & Wilkins, Philadephia.

[29] Bhattacharjee S, Wang Z, Ciampa J, et al. (2010). Using principal components of genetic variation for robust and powerful detection of gene-gene interactions in case-control and case-only studies. Am J Hum Genet 86:331–42. doi: 10.1016/j.ajhg.2010.01.026

[30] Greenland S (2009). Interactions in epidemiology: relevance, identification, and estimation. Epidemiology 20:14–17. doi: 10.1097/EDE.0b013e318193e7b5

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

[32] Siemiatycki J, Thomas DC (1981). Biological models and statistical interactions: an example from multistage carcinogenesis. Int J Epidemiol 10:383–7.

[33] Clayton DG (2009). Prediction and interaction in complex disease genetics: experience in type 1 diabetes. PLoS Genet 5:e1000540. doi: 10.1371/journal.pgen.1000540

[34] Thomas D (2010). Gene-environment-wide association studies: emerging approaches. Nat Rev Genet 11:259–272. doi: 10.1038/nrg2764

[35] 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 13:153–162.

[36] Kraft P, Yen YC, Stram DO, et al. (2007). Exploiting gene-environment interaction to detect genetic associations. Hum Hered 63:111–119. doi: 10.1159/000099183

[37] Gauderman WJ (2002). Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol 155:478–484.

[38] Yang Q, Khoury MJ, Flanders WD (1997). Sample size requirements in case-only designs to detect gene-environment interaction. Am J Epidemiol 146:713–720.

[39] Dennis J, Hawken S, Krewski D, et al. (2011). Bias in the case-only design applied to studies of gene-environment and gene-gene interaction: A systematic review and meta-analysis. Int J Epidemiol 40:1329–1341. doi: 10.1093/ije/dyr088

[40] Chatterjee N, Kalaylioglu Z, Carroll RJ (2005). Exploiting gene-environment independence in family-based case-control studies: increased power for detecting associations, interactions and joint effects. Genet Epidemiol 28:138–56. doi: 10.1002/gepi.20049

[41] Liu X, Fallin MD, Kao WH (2004). Genetic dissection methods: designs used for tests of gene-environment interaction. Curr Opin Genet Dev 14:241–5. doi: 10.1016/j.gde.2004.04.011

[42] Le Marchand L, Wilkens LR (2008). Design considerations for genomic association studies: importance of gene-environment interactions. Cancer Epidemiol Biomarkers Prev 17:263–7. doi: 10.1158/1055-9965.EPI-07-0402

[43] Clayton D, McKeigue PM (2001). Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet (London, England) 358:1356–60. doi: 10.1016/S0140-6736(01)06418-2

[44] Goldstein AM, Falk RT, Korczak JF, Lubin JH (1997). Detecting gene-environment interactions using a case-control design. Genet Epidemiol 14:1085–1089. doi: 10.1002/(SICI)1098-2272(1997)14:6<1085::AID-GEPI87>3.0.CO;2-D

[45] Lewontin RC, Kojima K (1960). The evolutionary dynamics of complex polymorphisms. Evolution (N Y) 14:458–472. doi: 10.2307/2405995

[46] Slatkin M (2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477–85. doi: 10.1038/nrg2361

[47] Hästbacka J, de la Chapelle A, Kaitila I, et al. (1992). Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 2:204–11. doi: 10.1038/ng1192-204

[48] International HapMap Consortium (2005). A haplotype map of the human genome. Nature 437:1299–320. doi: 10.1038/nature04226

[49] Wang K (2009). Testing for genetic association in the presence of population stratification in genome-wide association studies. Genet Epidemiol 33:637–645. doi: 10.1002/gepi.20415

[50] Price AL, Patterson NJ, Plenge RM, et al. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909. doi: 10.1038/ng1847

[51] Pritchard JK, Rosenberg NA (1999). Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228. doi: 10.1086/302449

[52] Devlin B, Roeder K (1999). Genomic control for association studies. Biometrics 55:997–1004.

[53] Wang Y, Localio R, Rebbeck TR (2006). Evaluating bias due to population stratification in epidemiologic studies of gene-gene or gene-environment interactions. Cancer Epidemiol Biomarkers Prev 15:124–132. doi: 10.1158/1055-9965.EPI-05-0304

[54] Wang LY, Lee WC (2008). Population stratification bias in the case-only study for gene-environment interactions. Am J Epidemiol 168:197–201. doi: 10.1093/aje/kwn130

[55] Campbell MC, Tishkoff SA (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet 9:403–33. doi: 10.1146/annurev.genom.9.081307.164258

[56] Teo YY, Small KS, Fry AE, et al. (2009). Power consequences of linkage disequilibrium variation between populations. Genet Epidemiol 33:128–35. doi: 10.1002/gepi.20366

[57] Schaid DJ (1999). Case-parents design for gene-environment interaction. Genet Epidemiol 16:261–273.

[58] Liu JZ, van Sommeren S, Huang H, et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet 47:979–86. doi: 10.1038/ng.3359

[59] Xiong M, Feghali-Bostwick CA, Arnett FC, Zhou X (2005). A systems biology approach to genetic studies of complex diseases. FEBS Lett 579:5325–32. doi: 10.1016/j.febslet.2005.08.058

[60] Willis-Owen SA, Valdar W (2009). Deciphering gene-environment interactions through mouse models of allergic asthma. J Allergy Clin Immunol 123:14–23. doi: 10.1016/j.jaci.2008.09.016

[61] Knox SS (2010). From “omics” to complex disease: a systems biology approach to gene-environment interactions in cancer. Cancer Cell Int 10:11. doi: 10.1186/1475-2867-10-11

[62] Marchini J, Howie B, Myers S, et al. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–13. doi: 10.1038/ng2088

[63] Hao K, Chudin E, McElwee J, Schadt EE (2009). Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. BMC Genet 10:27. doi: 10.1186/1471-2156-10-27

[64] Anderson CA, Pettersson FH, Barrett JC, et al. (2008). Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. Am J Hum Genet 83:112–9. doi: 10.1016/j.ajhg.2008.06.008

[65] Cheng KF, Lin WJ (2009). The effects of misclassification in studies of gene-environment interactions. Hum Hered 67:77–87. doi: 10.1159/000179556

[66] Calle ML, Urrea V, Malats N, Van Steen K (2010). mbmdr: an R package for exploring gene-gene interactions associated with binary or quantitative traits. Bioinformatics 26:2198–9. doi: 10.1093/bioinformatics/btq352

Acknowledgements

Taking part in this project was a great experience for me. I had the chance to actively participate in an international and creative scientific research group. First and foremost, I would like to thank Prof. Dr. Michael Krawczak for giving me an opportunity to work on my thesis at the Institute of Medical Informatics and Statistics (IMIS). His friendly and attentive supervision have significantly contributed to steer me in the right direction throughout. The way from the initial idea to the final outcomes was very enlightening and introduced me a scientific way of thinking.

Without saying, my thanks go to my colleagues at the IMIS, who contributed immensely to my personal and professional time. I thank all the members for providing a friendly and supportive environment to carry out my research. In particular, I am grateful to Prof. Dr. Astrid Dempfle and Dr. Sandra Freitag-Wolf for their useful discussions. I am also thankful to Carolin Knecht (geb. Däumer) who helped with the translation of my thesis summary into German. Moreover, I am indebted to Prof. Dr. Wolfgang Lieb at the Institute of Epidemiology for his keen interest into my project.

I gratefully acknowledge the funding source that made my thesis work possible. This work was supported by Research Training Group (RTG) 1743 "*Genes, Environment and Inflammation*" as a part of *Deutsche Forschungsgemeinschaft* (DFG). I thank RTG for providing the overall excellent working conditions, especially for financial support, enabling self-improvement by participation at some soft-skill courses and allowing me to visit several national and international conferences.

Lastly, I would like to thank all my family members and friends for their love and for providing moral support and encouragement throughout my life.

Declaration

I hereby declare that this thesis is the outcome of my own work and effort. Apart from the advice and guidance of my supervisors, all third party help and any text or contents from other sources have been cited appropriately. I affirm in lieu of oath that up to this date I have not failed any dissertation procedures and that this thesis has not been previously submitted elsewhere. This work has been carried out in strict accordance with the rules of good scientific practice of the *Deutsche Forschungsgemeinschaft*.

Kiel, ____________ ____________________________

Pankaj Yadav

Curriculum Vitae

Personal Details

Name:	Pankaj Yadav
Date of birth:	Dec 05,1987
Place of birth:	Dulana, India
Nationality:	Indian
Marital status:	Married

Academic Training

12/2012 – present	PhD student at Institute of Medical Informatics and Statistics, Kiel
10/2010 – 11/2012	Master of Science in Life Science Informatics, University of Bonn, Germany
05/2005 – 04/2009	Bachelor of Engineering in Biotechnology, Panjab University, India
03/2003 – 02/2005	Senior secondary school (main subjects: physics, chemistry, mathematics), Chandigarh, India

Work Experience

12/2012 – present	PhD research student at Institute of Medical Informatics and Statistics, Kiel DFG project title: *Genome-wide investigation of gene-environment interactions using a case-only design for inflammatory bowel disease*
04/2012 – 11/2012	Master thesis at Helmholtz Zentrum München, Munich title: *A compound centric modeling approach for QSAR prediction*
01/2010 – 03/2010	Research Trainee at IBI Biosolutions Private Limited, India topic: PERL programming and its application in bioinformatics and database development using MySQL
07/2008 – 05/2009	Bachelor thesis at Panjab University, India title: *Anti-tumor activity of Tinospora cordifolia with focus on deep aspects of cell and molecular biology*
06/2008 – 07/2008	Research Trainee at Ranbaxy (R & D), India topic: Basic aspects for gene cloning and protein purification methods
06/2007 – 07/2007	Research Trainee at IBI Biosolutions Private Limited, India topic: *In silico* modeling and drug docking analysis of enzyme Lipoate Biosynthate of *Mycobacterium tuberculosis*