Gene-Environment Interaction and Extension to Empirical Hierarchical Bayes Models in Genome-Wide Association Studies

---

Dissertation
zur Erlangung des humanwissenschaftlichen Doktorgrades
in der Medizin
der Georg-August-Universität Göttingen

vorgelegt von
Elena Viktorova
aus Ufa

Göttingen, 2014

Members of the Thesis Committee

Supervisor:	Prof. Dr. Heike Bickeböller Institut für Genetische Epidemiologie Universitätsmedizin Georg-August-Universität Göttingen
Second Thesis Committee Member:	Prof. Dr. Tim Friede Abteilung Medizinische Statistik Universitätsmedizin Georg-August-Universität Göttingen
Third Thesis Committee Member:	Prof. Dr. Dieter Kube Abteilung Hämatologie & Onkologie Universitätsmedizin Georg-August-Universität Göttingen
Day of Disputation:	17th June, 2014

Affidavit

Here I declare that my doctoral thesis entitled "Gene-Environment Interaction and Extension to Empirical Hierarchical Bayes Models in Genome-Wide Association Studies" has been written independently with no other sources and aids than quoted.

Elena Viktorova

Göttingen, April 2014

elena.viktorova@med.uni-goettingen.de
Georg-August-University Göttingen
Enjoy reading!

Acknowledgement

The writing of this doctoral thesis has been one of the biggest academic challenges I have ever faced in my life. Clearly, I would not have succeeded without the great help, support, and guidance of the following people.

First of all, I would like to express my deepest gratitude to my supervisor Professor Dr. Heike Bickeböller, who has accompanied me throughout this challenging journey. Her guidance, experience, and knowledge were a great aid during my research work as well as while writing this thesis. She acted as my mentor for the last few years despite her many other academic and professional commitments. Professor Bickeböller always offered me her support in all scientific dealings and my desire to participate at a number of conferences, workshops, or additional scientific training courses. She was always willing to share her experience, knowledge and wise with me in many other aspects of life. The advices and help that she gifted to me in the last three years are priceless. At the end of my studies, she spent countless hours correcting and significantly improving this dissertation. I remain truly indebted to her for this collaboration.

My thanks go to the TRICL consortium (NIH U19CA148127) for providing the lung cancer GWAS data. I would like to mention my appreciation of the research training group "Cancer Pharmacogenomics" (GRK 1034) and in particular the speaker Professor Dr. med. Jürgen Brockmöller for supporting my work. I am particularly grateful to GRK 1034 for giving me the opportunity to complete a short research stay at the University of Southern California (USC) by covering my travel expenses. I also enjoyed our annual scientific retreat days, which allowed me the opportunity to interact with other PhD students and professors of the research training group.

I would specifically like to express my thanks to Professor Duncan Thomas, with whom I spent two months at USC. During my time in Professor Thomas' research team, I received a lot of advice, guidance, and help from Professor Thomas himself and Dr. Juan Pablo Lewinger. I am very grateful to both of them.

I would like to thank many professors at the University of Göttingen, but in particular Professor Dr. Dieter Kube for agreeing to sit on my PhD thesis committee and always finding the time to participate in my presentations, as well as for his kind comments and questions. I extend my thanks to Professor Tim Beißbarth for reading my thesis. I say thank you to Professor Dr. med.

Ralf Dressel for the opportunity to work as statistical consultant and perform analyses for the projects of his team. I appreciate the time and agreement of Professor Tim Friede to talk with me on the topics of my dissertation and for his participation in my progress report talks. I am grateful to Professor Martin Schlather, who originally suggested me to apply for this doctorate candidate position, as speaker for the Center of Statistics and who invited me to move to Germany to complete my PhD degree.

I would also like to recognize the assistance towards completion of this thesis by one person, a colleague and friend of mine Andrew Entwistle, who kindly agreed to proofread my dissertation and who improved the language of the work a lot. He also assisted with both of my publications, each time carefully reading and correcting them. We experienced a lot of funny moments and conversations. This all made the working atmosphere enjoyable.

I would like to say a few words of thanks to the University of Göttingen and in particular to the Center of Statistics and the University Medical Center for providing me with an excellent working environment and facilities.

I feel happy to have worked with so many nice people during my time here. I appreciate my colleagues for the wonderful time that we had together, in particular with Dr. Dörthe Malzahn while working on statistical consulting projects. This work was particularly productive with Stanislav Syekirin on the statistical R package.

I feel blessed to have met a number of friends I truly love from my deepest heart here in Göttingen and with whom I did a lot together. They always supported, cared, and had a free ear for me. Svetlana Miller and Thuy Ha are among them. Special words of appreciation go to Dr. Mehran Rafigh who was always there for me, every time I needed help, who encouraged me to learn German and invested a lot of effort in my professional skills.

Last but not least, I would like to express the appreciation from all of my heart to my loving family, my father Professor Vitaliy Viktorov, my mother Professor Tatiana Viktorova and my brother Dr. Sergey Viktorov. Even though so far away, they still managed to deliver incredible support, love, and help to every step I took on my way. I always knew they pray for me, they love me, and will do everything they can to help me succeed. I am speechless when I think about everything my family has done to make all this happen.

Finally, I would like to say thank you to all those I am unable to name who participated in this success and without whom I would have been unable to complete this work.

Quotes

"*What we observe is not nature itself, but nature exposed to our method of questioning"*

Werner Heisenberg (1901-1976)

"*Some men have constitutions that are like wooded mountains running with springs, others like those with poor soil and little water, still others like land rich in pastures and marshes, and yet others like the bare, dry earth of the plain*."

Hippocrates (5th century)

Contents

Abstract

There are over 100,000 human diseases of which only around 10,000 are known to be monogenic, resulting from modification in a single gene. Many multifactorial diseases, such as cancer and lung cancer in particular, are outcomes of the interplay between genetic and environmental factors. It is well known that smoking is the major environmental risk factor in lung cancer.

In recent years, great progress in genotyping technology and cost control has enabled researchers to perform large-scale association studies, involving thousands of individuals genotyped on millions of markers. To date, genome-wide association studies (GWAS) have identified hundreds of genetic risk factors in complex diseases. However, the detected variants explain only a small part of the total heritability. Unexplained phenotypic variance may be partly attributed to undetected gene-environment (G×E) interactions. Therefore, there has been a rapid evolution in the development of statistical tools to discover biologically credible G×E interactions in a genome-wide context.

The analysis of G×E interactions remains one of the greatest challenges in the post-genome-wide-association-studies era. Uncovered population stratification in large association and interaction studies may lead to false positive results or masks true signals via under (over)-estimation of the true effects. In this dissertation, we began by evaluating the robustness or the magnitude of the bias due to population stratification in case-control studies of G×E interaction. A simple equation was derived to measure the population stratification bias of the interaction effect for the case-control estimator of G×E interaction.

Another great challenge to G×E interaction research remains the ability to maintain adequate power, while accounting for gene-environment (G-E) correlation in the source population. G-E correlation occurs when exposure to the environmental condition depends on the individual's genotype or vice versa, irrespective of the disease status of that individual. The empirical hierarchical Bayes approach to G×E interaction (EHB-GE_{CHI}) benefits from greater power than the classical case-control test, while accounting for population based G-E correlation. We developed extensions of EHB-GE_{CHI} with respect to covariate adjustment, general exposure and genotype and to performance under an additive mode of inheritance.

In this dissertation, we finally introduce an alternative to EHB-GE_{CHI} which is computationally more efficient, using a more stable model to obtain the posterior estimates of G-E correlation

in controls. Incorporating a parametric Bayes inference framework, with a normal distribution in a hierarchical model, we developed an approach that corrects for G-E correlations, gathering information across all markers simultaneously (as does EHB-GE_{CHI}). We named it the empirical hierarchical Bayes approach for G×E interaction EHB-GE_{NN}. Our simulation study demonstrates that EHB-GE_{NN} controls type I error better than EHB-GE_{CHI} while remaining powerful.

The last objective of this thesis is to consider the joint tests for genetic marginal and G×E interaction effects. Previous studies suggest that G×E interactions might help to detect genetic variants missed by a test for association with main effects. Specifically, some SNPs may have a moderate genetic and a G×E interaction effect and thus joint tests for marginal association and G×E interaction were developed to gain additional power over tests of main effects. Here we present how EHB-GE_{NN} can be adopted for joint testing, resulting in the EHB-GE_{NN}^{J} test.

The application of EHB-GENN and joint tests on four lung cancer GWASs from the ILCCO/TRICL consortia is presented and the results are discussed. We detected known markers for lung cancer, e.g. rs1051730 in CHRNA3, rs8034191 in AGPHD1 and suggestive signals, e.g. rs7982922 in ENOX1, rs2736100 in TERT, applying joint tests, using either case-control, case-only, MUK-EB or EHB-GE_{NN} for the G×E interaction component.

List of Tables

List of Figures

Chapter 1

1. Introduction

Charles Darwin in his opus "*On the Origin of Species"* stated that there are two factors responsible for biological variation-"the nature of the organism and the nature of the conditions" (Darwin 1869). Darwin represents the idea of genes and environment as being two forces acting synergistically to design our individual characteristics. Nowadays, it is well known that most of the multifactorial human traits and diseases, such as asthma, diabetes, cardiovascular diseases, depression, rheumatoid arthritis, and cancer, result from a complex interplay of the individual genetic and various environmental factors.

Cancer is the leading cause of worldwide mortality. All cancer forms together were responsible for 8.2 million deaths and 14.1 million new cancer cases around the globe in 2012 (WHO) (http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx). Even though the cancer mortality rate has continued to decline within the last two decades, the prognosis remains poor (Jemal, Simard et al. 2013). In the European Union, the predicted number of cancer deaths for 2013 is 1.3 million (Malvezzi, Bertuccio et al. 2013)

Lung cancer is the most lethal malignant disease, having caused 1.37 million deaths worldwide annually according to figures from 2008 (WHO) (World Health Organization Report on the Global Tobacco Epidemic, 2008). Lung cancer alone is responsible for more cancer-related deaths than breast, prostate, and colon cancers together (Jemal, Siegel et al. 2008). In Germany, it is the third most common cancer type after prostate, colon, and breast cancers. According to the population-based cancer registries in Germany, 35,040 men and 17,030 women newly developed lung cancer in 2010 (Krebsregister and (GEKID) 2013). Lung cancer is a complex disease of the uncontrolled cancer cell growth in tissues of the lung. Lung cancer is classified in two main types: small cell (SCLC) and the more common non-small cell (NSCLC) lung

cancer. These two types differ in their growth rates and are treated differently. The most abundant of the three histological forms of NSCLC is adenocarcinoma, which is also the most common type of lung cancer in lifelong non-smokers, so-called "never smokers" (Subramanian and Govindan 2007).

Various environmental factors may affect the risk of lung cancer development, such as exposure to tobacco smoke, radon, asbestos, arsenic, diesel exhaust, silica, and chromium.

Lung cancer in non-smokers may occur due to a combination of genetic factors (Gorlova, Weng et al. 2007) with radon (Catelinois, Rogel et al. 2006), asbestos (O'Reilly, McLaughlin et al. 2007) and air pollution (Chiu, Cheng et al. 2006, Coyle, Minahjuddin et al. 2006, Kabir, Bennett et al. 2007), including second-hand smoke (WHO , Smoking and and Health 2006).

In USA, *the major environmental risk factor* for lung cancer is exposure to tobacco smoke. Smoking contributes to 80% and 90% of lung cancer deaths in women and men, respectively (US Department of Health and Human Services, 2004). In Germany, 90% of lung cancer cases in men and 60% in women are attributed to active smoking (Robert Koch-Institut und die Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. 2012).

However, not only smoking alone increases the risk of developing lung cancer. Nowadays, it is well recognized that genetic factors also play a role in lung cancer development. Single nucleotide polymorphisms (SNPs) in a number of genome regions have been reported as being associated with lung cancer. Mutations in the genes *CHRNA5*, *CHRNA5*, *CHRNB4* located on chromosome *15q25.1* (Amos, Wu et al. 2008, Hung, Christiani et al. 2008, Thorgeirsson, Geller et al. 2008), SNPs on chromosome *5p13.3* (McKay, Hung et al. 2008, Wang, Broderick et al. 2008, Landi, Chatterjee et al. 2009), mutations in *BAT3* on chromosome *6p21.33* (Wang, Broderick et al. 2008), *RAD52* on *12p13.3* (Shi, Chatterjee et al. 2012), and in the *CDKN2A/CDKN2B* genes on chromosome *9p21.3* (Timofeeva, Hung et al. 2012) were discovered to affect the risk of lung cancer in genome-wide association studies (GWAS). Even

though active smoking is the lead environmental risk factor in lung cancer, only 10% of heavy smokers are known to develop the disease (Sauter, Rosenberger et al. 2008). This together with the discovery of genetic causes of the disease suggests that the inter-individual genetic variability affects the metabolism of tobacco-smoke carcinogens and leads to risk modification for some groups (Matakidou, Eisen et al. 2005, Amos 2007, Sun, Schiller et al. 2007). Therefore, studies of G×E interactions play an important role in public health, especially in the context of cancer research. They may further help to understand the nature of many complex diseases (Thomas 2010a, Thomas 2010b) and more specifically of the above lung cancer, for which smoking has such a considerable impact.

The initial completion and ongoing development of the International HapMap Project (International HapMap Consortium 2003, International HapMap Consortium 2005) and Human Genome Project (International Human Genome Sequencing Consortium 2004) lists human genetic variation at millions of polymorphic locations in several human populations, supporting more powerful association study designs. Recent advances in genotyping technologies, together with a significant reduction in the associated costs, has enabled researchers to genotype millions of common and rare single nucleotide polymorphisms (SNPs) both rapidly and accurately (International HapMap Consortium 2005, Frazer, Murray et al. 2009, Spencer, Su et al. 2009). A direct consequence is the opportunity to perform genome-wide association studies (GWAS), investigating the role of individual genetic variability in the etiology of complex diseases such as cancer. The genome-wide association study was originally designed to investigate DNA variations associated with common diseases (Hardy and Singleton 2009, Manolio, Collins et al. 2009). Nowadays, a new generation of GeneChips (Affymetrix) and BeadChips (Illumina) not only target common and rare SNPs but also known copy number variations (CNV), based on the maps available for the human genome (Redon, Ishikawa et al. 2006, McCarroll 2008, Itsara, Cooper et al. 2009). Recently, a lot of effort was

undertaken in developing methods for low-cost whole-genome next generation sequencing (NGS) (Mardis 2008, Schuster 2008, von Bubnoff 2008), which will capture even more rare variants, previously missing.

In addition to the technological advances in the field, genetic association studies and studies of gene-environment interactions can benefit from improvements in study design and the development of novel statistical approaches. In the following, I list statistical methods commonly used in G×E interaction studies for a case-control design. Consider a case-control study with a total of *N* individuals. Let *G* denote a genotype, *E* denote the exposure variable, and *D* the disease outcome variable. Many of the existing association tests including interaction tests are based on logistic regression models such as

$$logit\,(P(D = 1|\,G)) = \alpha_0 + \beta_G G + \beta_Z Z^t \qquad \textbf{(1.1)}$$

$$logit\,(P(D = 1|\,G, E)) = \alpha_{0CC} + \beta_{G\_CC} G + \beta_{E\_CC} E + \beta_{CC} G \times E + \beta_{ZCC} Z^t \qquad \textbf{(1.2)}$$

Equation (1.1) models the association between *D* and *G*, therefore $\beta_G = 0$ tests for the presence of a genetic main effect, while equation (1.2) includes genetic, environmental, and G×E interaction effects. Both (1.1) and (1.2) are adjusted for the covariate Z.

The classic case-control (CC) method for G×E interactions estimates the corresponding coefficient β_{CC} per SNP, which is equal to the natural logarithm of the odds ratio for G×E interaction from a logistic regression model (1.2).

The case-control test analyzes G×E interaction as departure from the multiplicative odds ratio model. It is often underpowered to detect G×E interactions, especially in situations in which genetic and environmental factors are rare and the interaction effect is weak (Mukherjee, Ahn et al. 2012).

Some researchers have addressed the lower power issue of the case-control test for G×E interactions by developing statistical tools designed to increase power to detect such associations besides marginal genetic effects. One powerful proposal is the case-only design, in which tests for G×E interaction are performed without considering controls (Piegorsch, Weinberg et al. 1994, Khoury and Flanders 1996). Under the assumption of the absence of population-based gene-environment correlation (G-E), the case-only (CO) test provides a valid procedure to test for G×E interaction, characterized by the more precise estimate of G×E interaction and therefore more powerful alternative to the CC test. As proposed by the case-only method, under G-E independence the odds ratio of G×E interaction can be estimated using information only from cases (Piegorsch, Weinberg et al. 1994). However, when the assumption of G-E independence is violated, the CO test produces a large number of false positive results; in other words the CO test has a highly inflated type I error rate.

Generally, on genome-wide level one would expect to see only a small number of genes and therefore a moderate number of SNPs with true detectable G-E correlation. However, this may be different for diseases such as lung cancer with a strong behavioral component. It is also well known that population stratification leads to such spurious dependence between genotype and environment in the general population (Thomas 2010a). Therefore, in the presence of population stratification thousands of markers may induce population-based G-E correlation. These correlations result from the difference in the genetic origin of individuals, i.e. differences in minor allele frequencies across the subgroups and cultural differences leading to the specific behavior and favor of the specific exposures, resulting in differences in the environment distribution. Since confounding owing to the population stratification leads to biased G×E effect estimates, it is important to control for the ancestry covariates in the analyses (Bhattacharjee, Wang et al. 2010).

In recent years, further methods to test for G×E interaction have been proposed, aiming to increase the power while keeping type I error at the nominal 5% level, mainly exploiting the assumption of G-E independence or trying to account for G-E correlation.

The two-step approach to scan for G×E interactions was developed by Murcray in 2009 (Murcray, Lewinger et al. 2009), which we will refer to as Murcray's two-step test (MUR). During the first step, the approach screens for G-E correlation in the combined sample of cases and controls. Then only a subset of SNPs that exceed a given significance threshold in step one is selected and tested for G×E interaction in step two. This test combines power protection from bias of the case-control estimator in a two-step procedure with the test statistics being independent from each other. A disadvantage of MUR is that the power of the first step depends on the case-control ratio. A large number of controls compared to cases leads to a decrease in power in step one and hence a loss of power for the overall procedure (Murcray, Lewinger et al. 2011).

The empirical Bayes type shrinkage estimator (MUK-EB) proposed by Mukherjee and Chatterjee (Mukherjee, Ahn et al. 2008, Mukherjee and Chatterjee 2008) combines the robust case-control estimator with the efficient case-only estimator in a single Bayes type shrinkage estimator. This estimator is approximately robust to the presence of G-E correlation in the source population and performs comparably to the case-control estimator under large departures from independence. This method does not strictly adhere to nominal type I error rate level under violation of the G-E independence assumption and moderate sample size. However, it does maintain a smaller mean squared error (MSE) compared to the other estimators listed above irrespective of the true state of the G-E correlation.

Recently Sohns and colleagues developed the empirical hierarchical Bayes approach to G×E interaction (EHB-GE_{CHI}) (Sohns 2012, Sohns, Viktorova et al. 2013). EHB-GE_{CHI} is based on a two-level hierarchical model with a parametric distribution assigned to the parameters during

both stages; the *chi distribution* and a mixture distribution with the point mass at zero. EHB-GE_{CHI} does not require the assumption of G-E independence. In fact, the approach estimates the G-E correlation effect by borrowing information across all SNPs.

The EHB-GE_{CHI} test has inflated type I error in the presence of a large number of G-E correlations. The approach is therefore not recommended for significance testing. EHB-GE_{CHI} is however proposed as a powerful ranking method to identify biologically plausible signals worth further detailed investigation (Sohns, Viktorova et al. 2013). "Rank power" is defined as the percentage of simulated replicates in which the true interacting SNP is within the top ranking positions, according to the absolute value of the corresponding test statistics, for example top 25 (Sohns, Viktorova et al. 2013). EHB-GE_{CHI} was shown to be the most powerful procedure in most of the cases in terms of rank power (Kuo and Zaykin 2011) when compared to the other G×E methods listed above.

All the approaches discussed above were designed specifically to study G×E interactions and therefore do not involve the estimation of genetic marginal or joint effects. Nevertheless, it is also interesting to know if G×E interaction may help to uncover additional genetic variants associated with disease, markers with moderate G×E interaction and main effects. This idea is based on the belief that even though a disease locus only modifies the disease risk in presence of the environment, the locus may still have a detectable marginal effect on the disease (Clayton and McKeigue 2001). Joint tests were recently investigated by some research groups in terms of achieved power and type I error (Chatterjee, Kalaylioglu et al. 2006, Kraft, Yen et al. 2007, Dai, Logsdon et al. 2012). Joint tests are performed to address simultaneous testing for the presence of a genetic main effect combined with a test for G×E interaction (Vanderweele, Ko et al. 2013).

Genome-wide studies of G×E interactions are challenging, since there are many pitfalls that can arise. We attempted to address some of these pitfalls. As a rule of thumb in case-control samples, the detection of an interaction requires a sample size at least four times larger than that required for the detection of a main effect of comparable size (Smith and Day 1984). Therefore, non-homogeneity of the study sample usually arises as an issue. The presence of population stratification in the study sample is the first problem in studies of G×E interactions, as it leads to a loss of power to identify true signals, spurious association signals, and can mask true associations. Principal component analysis (PCA) is currently the most powerful procedure to correct for population stratification in genetic main effects case-control GWASs (Price, Patterson et al. 2006). PCA was also shown to be an attractive approach to correct for the bias in studies of gene-gene (G-G) interactions (Bhattacharjee, Wang et al. 2010). In case-control studies of G×E interactions, we investigated the bias due to population stratification, deriving an analytical measure of the population stratification bias for case-control studies of G×E interactions. PCA was performed to correct for population stratification. We proposed PCA as a useful tool to correct for population stratification bias in GWAS of G×E interactions.

Another prominent problem in the study of G×E interactions is the occurrence of population-based G-E correlation for as many as thousands of markers. In a genome-wide context, the assumption of G-E independence cannot be surely stated and therefore statistical tools need to be able to relax this constraint. In this dissertation, we generalized the originally proposed EHB-GE_{CHI} method in three important ways: with respect to covariate adjustment; performance under the additive risk model assumption; and regarding applications with multilevel or continuous exposure, or genotype variables. However, some limitations remain, such as, for example, the complexity of the EHB-GE_{CHI} method and its relatively poor performance in the GWAS context and last but not least the inappropriateness of the approach to significance testing. Therefore, we proposed an alternative empirical hierarchical Bayes

approach for G×E interactions, naming it EHB-GE_{NN}. All three extensions mentioned above are valid in our modified EHB-GE_{NN}. Just as its predecessor EHB-GE_{CHI}, this novel approach does not require any assumption of independence between genotype and environment in the general population. It is characterized by a smaller number of hyperparameters requiring estimation on the dataset and by the ability to derive an exact equation for the posterior variance of the statistics. The asymptotic distribution of EHB-GE_{NN} test statistics is available as well. We propose EHB-GE_{NN} as a powerful tool keeping type I error rate at an approximately nominal level in contrast to EHB-GE_{CHI} and MUK-EB in samples in which a large number of G-E correlation signals with moderate to large effect size are suspected to occur. Moreover, to address the joint testing issue, we constructed a joint test EHB-GE_{NN}^{J} similar to that proposed by Dai and colleagues (Dai, Logsdon et al. 2012).

This thesis is motivated by lung cancer GWAS data from the International Lung and Cancer Consortium (ILCCO) and the Transdisciplinary Research in Cancer of the Lung (TRICL) consortium and is illustrated on four GWASs (Holle, Happich et al. 2005, Wichmann, Gieger et al. 2005, Amos, Wu et al. 2008, Hung, Christiani et al. 2008, Hung, McKay et al. 2008, Sauter, Rosenberger et al. 2008, Wang, Broderick et al. 2008) with smoking as the exposure factor. On analysis, we searched for G×E interactions applying the EHB-GE_{NN} approach. Findings following the application of competing methods on the same data including EHB-GE_{CHI} can be found in (Sohns 2012, Sohns, Viktorova et al. 2013). The discovery and understanding of G×E interactions clearly is a key to the future of personalized medicine. Novel findings in this area of research will very likely prove to be a direct benefit to public health, as they have the potential to lead to the future development of individualized treatments.

This dissertation is structured as follows: Chapter 2 includes a review of the literature and presents the necessary definitions and methods. Chapter 3 discusses issues concerning bias resulting from population stratification in studies of G×E interaction. A simple equation is

presented to evaluate the degree of population stratification bias in case-control studies of G×E interaction. A description and the results of the calculation as well as a simulation study are presented. The advantage of applying PCA to correct for population stratification in G×E interaction studies is discussed. Chapter 4 introduces the EHB-GE_{CHI} approach (Sohns 2012, Sohns, Viktorova et al. 2013) and describes limitations of the originally proposed method. Newly developed generalizations of the EHB-GE_{CHI} method are also presented in this chapter. Chapter 5 introduces our alternative approach "Empirical hierarchical Bayes approach for G×E (EHB-GE_{NN})" to studies of G×E interaction in the presence of many population-based G-E correlation signals with moderate to strong effect size. A description and the results of a simulation study comparing EHB-GE_{NN} versus other G×E interaction methods are presented. The same chapter describes the joint tests for genetic main and G×E interaction effects. Joint tests as proposed in (Dai, Logsdon et al. 2012) are described. Similarly, a joint EHB-GE_{NN}^{J} test was built. In Chapter 6, we present the lung cancer analyses and results. We applied EHB-GE_{NN} and four joint tests on four GWASs from the ILCCO/TRICL consortia. The data are described and the methods and results of these genome-wide studies are discussed. The thesis is concluded by Chapter 7 with a discussion and suggestions of future research questions in this field.

Chapter 2

2. Fundamentals of Human Genetics and Association Studies

This chapter reviews basic concepts of population genetics as well as case-control genetic association and gene-environment interaction (G×E) studies. This chapter also presents the necessary definitions and principles to understand the main challenges in the area of case-control genome-wide G×E interaction studies and our approach to addressing some of them. The statistical methods described in this chapter are standard methods to analyze G×E interactions in genome-wide association studies (GWAS) for a case-control design. Later in this thesis, these methods are employed in a comparative performance evaluation of our novel EHB-GE_{NN} approach to studies of G×E interaction, and are applied to analyze lung cancer GWAS data.

2.1. Population Genetics

2.1.1. Hardy-Weinberg Equilibrium

A keystone of population genetics is outlined in the Hardy-Weinberg law, a principle independently formulated by G.H. Hardy and W. Weinberg in 1908. The Hardy-Weinberg law relies on the assumption of random mating in a population. A random mating represents the situation, in which a mating occurs between individuals at random and implies absence of selection. The Hardy-Weinberg law describes the mathematical relationship between frequencies of alleles and frequencies of genotypes in a population at a locus (l). To illustrate the law, assume that q_A and q_a are the corresponding frequencies of alleles A and a at a biallelic locus l, so that $q_A+q_a=1$. The Hardy-Weinberg law postulates that in a random mating population the allele and genotype frequencies are in stable equilibrium, which is called *Hardy-*

Weinberg Equilibrium (HWE). The frequencies of the corresponding genotypes *AA*, *Aa* and *aa* are q_A^2, $2q_Aq_a$, and q_a^2, respectively. It indicates that the frequencies remain stable from generation to generation. On the other hand, allele frequencies can be derived from genotype frequencies under HWE by allele counting.

To check if population allele and genotype frequencies satisfy HWE, a χ^2 -test can be performed, which compares expected genotype frequencies derived from allele frequencies with those observed. Deviation from HWE may suggest e.g. the presence of selection or admixture of different populations. All markers, including single nucleotide polymorphisms (SNPs), are often tested for HWE during the quality control (QC) steps to avoid possible genotyping errors. Only control samples are used when testing for deviations from HWE. The threshold for declaring SNPs to be outside HWE varies significantly among studies; *p-values* between 0.001 and 5×10^{-8} (Zeggini and Morris 2010) are common depending on the number of SNPs under consideration.

2.1.2. Minor Allele Frequency

The minor allele frequency (MAF) refers to the frequency at which the least common allele occurs in a population or in the sample at hand. The frequency of alleles in the population can be estimated from their frequencies in a reference population, such as HapMap samples (International HapMap Consortium, Frazer et al. 2007). More often, MAF is estimated on the data on hand, and thus is only representative of cases or of controls. One of the alleles appears less frequently than the other and therefore is called minor allele. For a locus that is in Hardy-Weinberg Equilibrium in a diploid population, an allele that is at a frequency of 0.3 will be present in 51% of the population $[1 - (1 - 0.3)^2]$ and absent in 49% of the population $[(1 - 0.3)^2]$. Low MAF leads to poor performance of the genotype-calling algorithms (Weale 2010).

Therefore, during quality control of the data, it is reasonable to exclude markers with a MAF $\leq 5\%$ from further consideration depending on the sample size (Ziegler, König et al. 2008).

2.1.3. Linkage Disequilibrium

Genetic linkage represents violation of Mendel's Second Law, the law of independent assortment of genes, and is reflected in segregation of alleles at loci located close to each other on the same chromosome. Under independence, the frequency of *haplotypes*, for close loci defined as pairs of alleles at different loci on the same gamete, is the product of their respective allele frequencies. Therefore, when an excess or deficiency of some haplotypes exist, the loci are said to be in *linkage disequilibrium* (LD) (Khoury, Beaty et al. 1993). In other words, LD may be defined as an existing correlation between alleles located at nearby loci, owing to the possible joint inheritance (Ardlie, Kruglyak et al. 2002). For simplicity, assume that we have only two loci l_1 and l_2 with corresponding alleles *A/a* and *C/c* and allele frequencies q_A, q_a, q_C, q_c. Four haplotypes can be present for these two loci: *AC, Ac, aC,* and *ac*, with corresponding frequencies q_{AC}, q_{Ac}, q_{aC}, and q_{ac}. Hence, l_1 and l_2 are in equilibrium if

$$q_{AC} = q_A q_C,\ q_{Ac} = q_A q_c,\ q_{aC} = q_a q_C,\ q_{ac} = q_a q_c.$$

LD can be measured by the disequilibrium coefficient $D_{AC} = q_{AC} - q_A q_C$, which deviates from 0 in the presence of LD. Another measure of LD, which does not depend on the allele frequency is the squared correlation coefficient, r^2 (Ardlie, Kruglyak et al. 2002). It is defined as

$$r^2 = D^2 / (q_A q_a q_C q_c)$$

and ranges from 0 to 1. The HapMap database (http://www.hapmap.org) provides LD information across the whole human genome including the position of recombination hotspots (Zeggini and Morris 2010).

2.1.4. Population Stratification

A confounder is a variable that is not itself the object of a study, but is associated with the phenotype and at the same time with the variable under consideration. For example, a person's ethnicity can be a confounder associated with the marker allele under investigation. If the confounder is the ethnic affiliation of the individual, this is termed confounding by ethnicity or *population stratification* (PS) (Ziegler and König 2006). PS in case-control studies can occur when cases and controls are sampled from different populations in different proportions and the allele frequencies of genetic markers, often SNPs, are distributed unequally in these populations (Ziegler and König 2006).

Population stratification can act as a confounder when the genetic effect is assumed to be uniform across admixed subpopulations. On the other hand, PS can act as an *effect modifier* when the existing genetic effect is different in the subpopulations. In other words, the homogeneity of genetic effects in all subpopulations is assumed for a confounder, whereas for an effect modifier, heterogeneity across subpopulations is present. In addition to producing false-positives, population stratification might also mask a true association, thus reducing the power to detect a genetic effect (Ziegler and König 2006).

To test for the presence of population stratification in the study sample, Pritchard and Rosenberg (Pritchard and Rosenberg 1999) proposed to select randomly a set *M* of neutral markers in linkage equilibrium and construct χ^2-statistics for each marker, testing for association between the phenotype and the marker. Then, the sum of all statistics (χ_l^2, *l=1..M*) is formed $\chi_{PS}^2 = \sum_{l=1}^{M} \chi_l^2$ and it is asymptotically distributed as χ^2 with *M* degrees of freedom (df) under the null hypothesis. Failure to reject the null hypothesis by this test means that the sample is assumed to be homogeneous.

There are three well-known approaches in the literature to test for association in case-control studies while adjusting for unobserved population stratification. The first approach is the method of *Genomic Control* (GC), proposed by Devlin and Roeder (Devlin and Roeder 1999). The idea of GC is to use additionally genotyped marker loci ("null loci") to estimate empirically the variance inflation under the null hypothesis of no association. For this, an inflation factor λ is estimated as

$$\hat{\lambda} = \frac{median(\chi_1^2, \chi_2^2, \ldots \chi_M^2)}{0.4549},$$

in which 0.4549 is the median of the χ_{1df}^2 distribution. It is assumed that this remains constant across the genome. Then, the test statistic for any locus *l* is corrected to $\chi_l^2/\hat{\lambda}$.

The second approach proposed is the structured association (SA), (Pritchard, Stephens et al. 2000, Pritchard, Stephens et al. 2000). SA is a two-step procedure. The first step involves inferring details of the population structure from the sample using unlinked loci. The number of subpopulations and allele frequencies in each of them, as well as the mixed genetic ancestry of each individual are estimated employing a Monte-Carlo method at this step. In the second step, the information obtained is used to test for association within subpopulations (Ziegler and König 2006). Several different SA methods were subsequently proposed, for example as in (Köhler and Bickeböller 2006).

The third approach to correct for population stratification, which is applicable to genome-wide association case-control studies, is based on principal component analysis (PCA) and was proposed by Price and colleagues (Price, Patterson et al. 2006). To perform a PCA, more than 10,000 SNPs are necessary for the principal components estimation. The analysis is therefore only applicable in the GWAS context. The advantage of PCA over GC or SA is that the ancestry adjustment is performed per SNP. This allows us to correct for both false positive and false

negative associations (Weale 2010). Nowadays, PCA is the most commonly used and the most appropriate method to correct for PS in genetic association studies. We implemented PCA to account for population stratification in our study and as such, it is explained in more detail below.

2.1.5. Principal Component Analysis

Principal component analysis is a method of data dimensionality reduction. It is a roadmap of how to transform a large set of related variables into a new smaller set of independent variables to reveal hidden substructure in the original data. The main idea of PCA is that most of the variance in the original dependent variables, in the GWAS context genotypes, can be explained by a significantly smaller number of independent variables, termed principal components. Principal components are ordered according to the amount of the variance in the full set of original variance that they explain.

PCA can be performed on case-control data and can be summarized in the following steps. Let a GWAS dataset be coded in the form of a large $n \times m$ matrix with one row $i=1,...,n$ for an individual and one column $j=1,...,m$ for every SNP. Each cell ij of the original data matrix is the genotype of individual i at a particular SNP j, coded as (0,1,2) according to the minor allele count (g_{ij}).

Step 1 Normalize the original $n \times m$ matrix by subtracting column means and dividing by standard deviation.

Step 2 Calculate the covariance matrix for the normalized data variables. Assume Σ to be the $m \times m$ covariance matrix of $M=(\boldsymbol{m}_1 \ldots \boldsymbol{m}_m)$, so that $\Sigma_{jj'}=cov(\boldsymbol{m}_j,\boldsymbol{m}_{j'})$, where $\boldsymbol{m}_j=(g_{1j} \ldots g_{nj})^t$ is a jth SNP column-vector, $j=1,...,m$, $j'=1,...,m$, g_{ij} is each cell entry in $n \times m$ matrix.

Step 3 Calculate the eigenvectors and eigenvalues of Σ. To do so, let $a_1 \in \mathbb{R}^M$ be the first eigenvector and λ to be an eigenvalue, then we search for the vector maximizing

$$var(a_1^t M) = a_1^t \Sigma a_1$$

with $a_1^t a_1 = 1$. This defines an optimization problem with one constraint and can be solved using the method of Lagrange multipliers. Consider the function

$$a_1^t \Sigma a_1 = \lambda(a_1^t a_1 - 1)$$

where λ is a constant termed the Lagrange multiplier. Differentiating the equation above with respect to a_1 leads to $\Sigma a_1 - \lambda a_1 = (\Sigma - \lambda I_M) a_1 = 0$, with I_M being the $M \times M$ identity matrix. From this it follows that λ is an eigenvalue of Σ and a_1 is corresponding eigenvector. From the above and the fact that $\lambda \in \mathbb{R}$, the equation below follows

$$a_1^t \Sigma a_1 = a_1^t \lambda a_1 = \lambda a_1^t a_1 = \lambda.$$

Therefore, λ is the largest eigenvalue of Σ and a_1 is the first eigenvector, explaining the largest proportion of variance. Once a_1 is derived, the transformation $a_1^t M$ yields the first principal component. To obtain the second, third and finally *mth* principal components, we proceed in the same manner, choosing vector $a_2 \in \mathbb{R}^M$ maximizing the variance, such that $a_2^t M$ and $a_1^t M$ are uncorrelated, i.e. orthogonal. Then, $a_2^t M \ldots a_M^t M$ are *m* principal components. Mathematically speaking, this process is equivalent to a singular value decomposition of the original data matrix. In 2006, Price and colleagues demonstrated in application on case-control genetic data that the inclusion of the set of significant principal components as covariates into the analysis corrects for population stratification in genome-wide association studies, of the genetic main effect (Price, Patterson et al. 2006). PCA for GWAS data is integrated in the EIGENSOFT software package (Patterson, Price et al. 2006).

2.2. Case-Control Association Studies

2.2.1. Genome-Wide Association Studies

In a case-control design, the aim of a GWAS is to compare genetic variants in cases to those in controls and answer the question as to whether there is any association of these variants with the outcome status (cases/controls) (Witte 2010). Even though there is an increasing tendency to apply GWA methodologies to population-based cohorts, most published GWASs employ the case-control design (McCarthy, Abecasis et al. 2008). Genetic variation in such studies is often measured using single nucleotide polymorphisms (SNPs). GWASs are possible nowadays because millions of SNPs in the human genome have been identified.

2.2.2. Measures of Association

Consider the following data representation in an epidemiological study. Let *G=(0, 1, 2)* represent the minor allele count for an individual genotype. Let *D* denote the disease status with 1 for cases and 0 for controls. Let n_{ij} denote the number of subjects with *D=i*, *G=j* and *N* is the total number of individuals. Replacing any subscript with a dot (.) denotes summation over the subscript. We can summarize our data for each SNP in **Table 2.1**.

Table 2.1 Data representation in a case-control study with a SNP

	G=0	***G=1***	***G=2***	
D=1	n_1	n_{11}	n_{12}	$\mathbf{n_{1.}}$
D=0	n_0	n_{01}	n_{02}	$\mathbf{n_{0.}}$
	$\mathbf{n_{.0}}$	$\mathbf{n_{.1}}$	$\mathbf{n_{.2}}$	**N**

The most common measure of association between a categorical characteristic and a disease the "*relative risk"* (RR) of a member with the characteristic developing the disease compared to a member without this characteristic. For example, genetic association represents association

between a specific genotype and the disease and can be measured by the relative risk of a person with such a genotype developing the disease compared to a person with the reference genotype. To identify risk factors for disease development, the risks of contracting or developing the disease among people exposed to potential risk factors, such as genotype or environment, and those of an unexposed individual, such as wild-type genotype or absence of environment, are related to each other. The corresponding measure of risk is the relative risk.

The relative risk is the probability that a member of an exposed group will develop a disease (D=1) relative to the probability that a member of an unexposed group will develop that same disease.

$$RR = \frac{P(disease|exposed)}{P(disease|non-exposed)}$$

There are three different penetrances $r_g = P(D=1|G=g)$, $g=0,1,2$, that can be estimated by

$$\hat{r}_0 = \frac{n_{10}}{n_{.0}},\ \hat{r}_1 = \frac{n_{11}}{n_{.1}},\ \hat{r}_2 = \frac{n_{12}}{n_{.2}},$$

from the data presented in **Table 2.1**, where a penetrance is the disease risk given a specific genotype. Therefore the genotype relative risks compared to the G=0 reference genotype are defined as:

$$RR_{G=1} = \frac{P(D=1|G=1)}{P(D=1|G=0)} \text{ and } RR_{G=2} = \frac{P(D=1|G=2)}{P(D=1|G=0)}$$

can be estimated by

$$\widehat{RR}_{G=1} = \frac{\hat{r}_1}{\hat{r}_0} = \frac{n_{11}}{n_{.1}} / \frac{n_{10}}{n_{.0}} \text{ and } \widehat{RR}_{G=2} = \frac{\hat{r}_2}{\hat{r}_0} = \frac{n_{12}}{n_{.2}} / \frac{n_{10}}{n_{.0}}.$$

The genetic risk of a specific mode of inheritance is defined according to the relationship between the corresponding genotype RRs.

The *dominant* mode of inheritance satisfies $RR_{G=2}=RR_{G=1}>1$.

The *co-dominant* mode of inheritance satisfies $RR_{G=2}\neq RR_{G=1}>1$.

The *recessive* mode of inheritance satisfies $RR_{G=2}>1$, $RR_{G=1}=1$.

The *additive* mode of inheritance satisfies $RR_{G=2}=(2RR_{G=1}-1)>1$ (additive scale).

The *multiplicative* mode of inheritance satisfies $RR_{G=2}=(RR_{G=1})^2>1$.

In a prospective cohort study, the numbers of individuals in the exposed and non-exposed groups are representative of the whole population. This is not the case in retrospective case-control studies, since the number of individuals in each group is decided upon by the investigator and can therefore differ from the population case-control ratio. It is therefore impossible to estimate risks and thus relative risks from case-control data directly. However, association can then be measured by the so-called *odds ratio* (OR).

If an event takes place with probability *P*, the odds in favor of that event are *P to (1-P)*. The odds ratio relates two odds to each other. In our example, OR is the odds of exposed individuals among cases divided by the odds of exposed individuals among controls:

$$OR = \frac{P(exposed|disease)/(1-P(exposed|disease))}{P(exposed|non-disease)/(1-P(exposed|non-disease))}$$

For the data in **Table 2.1**, $OR_{G=1}$ and $OR_{G=2}$ can be estimated by

$$\widehat{OR}_{G=1} = \frac{n_{11}}{n_{01}} / \frac{n_{10}}{n_{00}} \text{ and } \widehat{OR}_{G=2} = \frac{n_{12}}{n_{02}} / \frac{n_{10}}{n_{00}}.$$

Generally, OR overestimates RR when RR>1 and underestimates it when RR<1. The two converge with decreasing disease prevalence. Under the assumption of a rare disease in the population, OR is a good approximation of RR and is therefore often implemented in case-

control studies. In practice, even with a disease prevalence of 10%, scientists estimate $\widehat{OR}$ from the data collected by a case-control study and use it to approximate RR.

2.2.3. Case-Control and Case-Only Studies

In a case-control study, sampling data are collected retrospectively and conditional on the disease status of the individuals. The common practice in association studies is to analyze such data ignoring the fact of the retrospective nature of the sampling. In 1956, Cornfield demonstrated that prospective and retrospective odds ratios are equivalent. Therefore, odds ratios estimation based on the case-control data is valid as according to (Cornfield 1956).

The efficiency of the approach was established in two other research papers by Andersen (Andersen 1970) and Prentice and Pyke (Prentice and Pyke 1979). They demonstrated that classic prospective analysis of the case-control data yields the correct maximum-likelihood estimates of the odds ratio parameter under the retrospective sampling design when the distribution of the underlying covariates is nonparametric.

Later, in 1994, Piegorsch and colleagues proposed the case-only approach to estimate the G×E interaction effect (Piegorsch, Weinberg et al. 1994). Under the population-based G-E independence assumption, it was shown that efficient estimates of G×E interaction for the categorical exposure and binary genotype variables can be derived through logistic regression in a case-only approach (Piegorsch, Weinberg et al. 1994, Umbach and Weinberg 1997). The CO approach was later extended to continuous environment and categorical genotype variables employing logistic, ordinal, and multinomial regression techniques (Albert, Ratnasinghe et al. 2001, Armstrong 2003, Cheng 2006).

2.2.4. Single Nucleotide Polymorphism

Molecular markers revealing polymorphisms at the deoxyribonucleic acid **(**DNA) level are essential in human genetic studies. Over the last ten years, the revolution in biological science, advanced genotyping and sequencing technologies, together with a substantial reduction in their cost, have enabled the research community grow significantly in terms of knowledge regarding genetic and genomic variation, as more and more genomes have been sequenced. One of the essential steps towards greater knowledge was the completion of the Human Genome Project in 2003 (Collins, Green et al. 2003). As a consequence of this, great progress in the discovery of genes influencing the risks of contracting and/or developing monogenic and complex human diseases has been made (Johnson 2009). The post-genome era is beginning to unravel the function of the human genome and explain how the circa 21,000 human genes interact with each other and the environmental conditions. Comparison of genomic DNA sequences in a variety of people reveals many positions at which two or sometimes more different nucleotide bases can be observed (Syvanen 2001). Such variation at a single position of a DNA sequence is called a single nucleotide polymorphism, or simply SNP. SNPs are very abundant in the human genome and are estimated to appear approximately once within every thousand bases (Sachidanandam, Weissman et al. 2001, Syvanen 2001, Venter, Adams et al. 2001). The effect of a SNP on a phenotype depends on the genome position at which the SNP occurs, be it a non-coding region or the coding region of a gene or its regulatory region. Multifactorial human diseases do not follow a simple Mendelian mode of inheritance, but are the result of the complex interplay between a number of genetic and environmental factors (Buselmaier and Tariverdian 1999, Thomas and Kejariwal 2004). There is increasing evidence that many complex diseases demonstrate association with various SNPs and a number of environmental factors. Identifying the molecular causes of multifactorial diseases has become the focus of many researchers. Association studies are rapidly gaining ground for human traits,

with the human Haplotype Map Project (International HapMap Consortium 2003) being funded to support these findings (Thomas and Kejariwal 2004).

2.2.5. Gene-Environment Interaction and Gene-Environment Correlation

The vast majority of common diseases occur as a result of the complex interplay between genetic and environmental factors. In genetic studies, gene-environment interaction (G×E) is present when genetic and environmental factors interact to cause a disease. In other words, the effect of the genotype and particular environment together on the disease risk differs from the separate effects of these factors (Ober and Vercelli 2011). For example, in cancer biology the susceptibility to particular external toxic elements depends on the efficiency of the DNA repair process, which can be different among the people with a different genetic signature. Another example is individual response to drug therapy or nutrition. Genetics may affect the response to a particular medication via drug metabolism and can also lead to medication or therapy intolerance (Hunter 2005). So far, numerous gene-environment associations with various complex diseases have been discovered through candidate gene or genome-wide association studies. For example, the *GST* superfamily polymorphisms have been demonstrated to be associated with an elevated risk of smoking-related lung cancers (Haugen, Ryberg et al. 2000, Raimondi, Paracchini et al. 2006). It was also demonstrated that female smokers develop a substantially higher expression level of *CYP1A1* in the lung when compared to males (Haugen, Ryberg et al. 2000). The variant alleles of the *NAT2* gene increase the risk of colorectal cancer only in combination with red meat consumption (Chen, Stampfer et al. 1998). Furthermore, variants of the *MC1R* gene, responsible for skin color, combined with UV radiation result in an increased skin cancer risk (Rees 2004), while on their own the genetic and environmental factors have no effect on the disease risk.

To understand the scope of this dissertation, it is of at most importance to distinguish between G×E interaction and gene-environment (G-E) correlation in the source population. In this dissertation, G×E interaction will always refer to gene-environment interaction and G-E correlation to gene-environment correlation. Population-based G-E correlation occurs when exposure to the environmental condition depends on an individual's genotype or vice versa, irrespective of the disease status of the individual. This can be either causal or spurious. An example of a causal G-E correlation would be smoking addiction genes, which favor smoking, such as *GPR51* and *CYPR51* (Caporaso, Gu et al. 2009), or the genes *GABRA2* and *ADH1C* correlated with alcohol addiction (Online Mendelian Inheritance in Man 2012). Generally, one would expect only a small number of genes to have a true causal G-E correlation, detectable on a genome-wide level. However, this can be different for some diseases such as lung cancer for which many SNPs may correlate with nicotine addiction. It is also well known that population stratification leads to a spurious dependence between genotype and environment in a general population, owing to non-causal mechanisms (Thomas, Lewinger et al. 2012), and may lead to a large number of G-E correlations. Understanding the difference between G×E interaction and G-E correlation is crucial to this dissertation. It is therefore important to separate these two terms. However, G×E interaction and G-E correlation are not mutually exclusive in reality and can occur simultaneously.

To introduce the approach to measure G×E interaction and G-E correlation in case-control studies, we restrict to the binary disease (D), a binary exposure (E) and the three level genotype (G) variables. As previously introduced, let $G=(0, 1, 2)$ be an individual genotype. Let E denote an exposure variable with 1 for exposed subjects and 0 otherwise. Let D denote the disease status with 1 for cases and 0 for controls. Let n_{ijk} denote the number of subjects with $D=i$, $G=j$ and $E=k$ and N is the total number of individuals. Then, data for each SNP may be presented in a 2×6 contingency table (**Table 2.2**).

Table 2.2 Data representation in a case-control study with a SNP and a single environment as factor

	E=1			***E=0***			
	G=0	***G=1***	***G=2***	***G=0***	***G=1***	***G=2***	
D=1	n_{101}	n_{111}	n_{121}	n_{100}	n_{110}	n_{120}	$\mathbf{n_{1..}}$
D=0	n_{001}	n_{011}	n_{021}	n_{000}	n_{010}	n_{020}	$\mathbf{n_{0..}}$
	$\mathbf{n_{.01}}$	$\mathbf{n_{.11}}$	$\mathbf{n_{.21}}$	$\mathbf{n_{.00}}$	$\mathbf{n_{.10}}$	$\mathbf{n_{.20}}$	**N**

The observed vector of cell counts for cases in the sub table $n_1=(n_{121}, n_{111}, n_{101}, n_{120}, n_{110}, n_{100})$ and respectively for controls $n_0=(n_{021}, n_{011}, n_{001}, n_{020}, n_{010}, n_{000})$ can be seen as realizations from two independent multinomial distributions $n_1 \sim MN(n_1, p_1)$ and $n_0 \sim MN(n_0, p_0)$, where $p_1=(p_{121}, p_{111}, p_{101}, p_{120}, p_{110}, p_{100})$ and $p_0=(p_{021}, p_{011}, p_{001}, p_{020}, p_{010}, p_{000})$ are the cell probabilities of the underlying case-control population. Then the following ORs per SNP may be calculated:

$OR_{G=1} = \frac{p_{110}}{p_{010}} / \frac{p_{100}}{p_{000}}$ and $OR_{G=2} = \frac{p_{120}}{p_{020}} / \frac{p_{100}}{p_{000}}$ for the genetic main effect, when *E=0*

$OR_E = \frac{p_{101}}{p_{001}} / \frac{p_{100}}{p_{000}}$ environmental main effect, at the reference genotype level *G=0*

$OR_{G=1E} = \frac{p_{111}}{p_{011}} / \frac{p_{100}}{p_{000}}$ and $OR_{G=2E} = \frac{p_{121}}{p_{021}} / \frac{p_{100}}{p_{000}}$ joint effect of genotype and environment.

Assuming a multiplicative risk model, G×E can be measured as follows

$$OR_{G\times E} = \frac{OR_{GE}}{OR_G OR_E}$$

Where OR_G is for G=1 or 2, likewise for OR_{GE}.

$$OR_{G\times E} \in (-\infty,\infty) \text{ and } OR_{G\times E} \begin{cases} > 1 & \text{positive G} \times \text{E, increasing disease risk} \\ = 1 & \text{no G} \times \text{E} \\ < 1 & \text{negative G} \times \text{E, decreasing disease risk} \end{cases}$$

Gene-environment correlation separately within cases or controls, respectively, can also be measured employing ORs, which we denote OR_{cases} and $OR_{controls}$ from now on

$$OR_{cases|G=1} = \frac{p_{111}}{p_{110}}\frac{p_{100}}{p_{101}} \text{ and } OR_{cases|G=2} = \frac{p_{121}}{p_{120}}\frac{p_{100}}{p_{101}}$$

$$OR_{controls|G=1} = \frac{p_{011}}{p_{010}}\frac{p_{000}}{p_{001}} \text{ and } OR_{controls|G=2} = \frac{p_{021}}{p_{020}}\frac{p_{000}}{p_{001}}$$

If G-E correlation is absent for a SNP in cases or in controls, then $OR_{controls}=1$ or $OR_{cases}=1$ for that SNP. As before, departure from 1 indicates the presence of G×E interaction.

It is very important for this thesis that G×E can be expressed by the ORs measuring G-E correlation within cases and within controls as

$$OR_{G\times E} = \frac{OR_{GE}}{OR_G OR_E} = \frac{\frac{p_{111}p_{000}}{p_{100}p_{011}}}{\frac{p_{110}p_{000}}{p_{100}p_{010}}\frac{p_{101}p_{000}}{p_{100}p_{001}}} = \frac{\frac{p_{111}p_{100}}{p_{110}p_{101}}}{\frac{p_{011}p_{000}}{p_{001}p_{010}}} = \frac{OR_{cases}}{OR_{controls}}.$$

Therefore, if $OR_{G\times E} = 1$, G×E is absent if $OR_{cases} = OR_{controls} = 1$ or if $OR_{cases} = OR_{controls} \neq 1$. And G×E is present if $OR_{cases} \neq 1$ and $OR_{controls} = 1$ or if $OR_{cases} \neq OR_{controls}$ and $OR_{cases} \neq 1$. We say that G-E correlation is present in a source population if $OR_{controls} \neq 1$ and this correlation is independent from the respective disease status of the individual. If the prevalence of the disease is small, i.e. the disease is rare in the population, $OR_{controls}$ in the presence of G×E and absence of population G-E converges to 1 (Schmidt and Schaid 1999).

Generally, ORs of genetic main effect, environmental main effect, and G×E interaction can be estimated via logistic regression models. Assume we want to model the probability P(D=1|G,E) for a SNP and a single environment (data as in **Table 2.2**).

$$logit\big(\mathrm{P}(\mathrm{D}=1|\mathrm{G},\mathrm{E})\big) = \log\left(\frac{\mathrm{P}(\mathrm{D}=1|\mathrm{G},\mathrm{E})}{\mathrm{P}(\mathrm{D}=0|\mathrm{G},\mathrm{E})}\right) = \alpha + \beta_E + \beta_G + \beta_{G\times E}GE, \quad \textbf{(2.1)}$$

where $\beta_E = log(OR_E)$, $\beta_G = log(OR_G)$, and $\beta_{G\times E} = log(OR_{G\times E})$.

The OR of the G-E correlation in cases and controls can also be modeled via logistic regression.

$$logit\big(\mathrm{P}(\mathrm{E}=1|\mathrm{G},\mathrm{D}=1)\big)=\log\left(\frac{\mathrm{P}(\mathrm{E}=1|\mathrm{G},\mathrm{D}=1)}{\mathrm{P}(\mathrm{E}=0|\mathrm{G},\mathrm{D}=1)}\right)=\alpha_{cases}+\beta_{cases}G \quad \text{and}$$

(2.2)

$$logit\big(\mathrm{P}(\mathrm{E}=1|\mathrm{G},\mathrm{D}=0)\big)=\log\left(\frac{\mathrm{P}(\mathrm{E}=1|\mathrm{G},\mathrm{D}=0)}{\mathrm{P}(\mathrm{E}=0|\mathrm{G},\mathrm{D}=0)}\right)=\alpha_{controls}+\beta_{controls}G,$$

(2.3)

where $\beta_{cases}=log(OR_{cases})$ and $\beta_{controls}=log(OR_{controls})$.

It is easy to see from the previous page that G×E interaction can be measured

$$\beta_{G\times E}=\log(\Psi)=log\left(\frac{OR_{cases}}{OR_{controls}}\right)=\beta_{cases}-\beta_{controls}. \qquad \textbf{(2.4)}$$

Equation (2.4) is crucial to this dissertation.

The *βs* can be estimated from the data by the maximum likelihood estimates (MLE) $\hat{\beta}$, which would then approximately follow a normal distribution, by

$$\hat{\beta}_{G=(1,2)}=log\left(\frac{n_{1G0}n_{000}}{n_{0G0}n_{100}}\right)\sim N(\beta_G,\sigma_G^2), \qquad \text{with } \hat{\sigma}_G^2=\sum_D\sum_G\frac{1}{n_{DG0}}$$

$$\hat{\beta}_E=log\left(\frac{n_{101}n_{000}}{n_{100}n_{001}}\right)\sim N(\beta_E,\sigma_E^2), \qquad \text{with } \hat{\sigma}_E^2=\sum_D\sum_E\frac{1}{n_{D0E}}$$

$$\hat{\beta}_{cases}=log\left(\frac{n_{1G1}n_{100}}{n_{1G0}n_{101}}\right)\sim N(\beta_{cases},\sigma_{cases}^2), \qquad \text{with } \hat{\sigma}_{cases}^2=\sum_G\sum_E\frac{1}{n_{1GE}}$$

$$\hat{\beta}_{controls}=log\left(\frac{n_{0G1}n_{000}}{n_{0G0}n_{001}}\right)\sim N(\beta_{controls},\sigma_{controls}^2), \text{ with } \hat{\sigma}_{controls}^2=\sum_G\sum_E\frac{1}{n_{0GE}}$$

And finally,

$$\hat{\beta}_{G\times E}=log\left(\frac{n_{1G1}n_{100}}{n_{1G0}n_{101}}\Big/\frac{n_{0G1}n_{000}}{n_{0G0}n_{001}}\right)=\hat{\beta}_{cases}-\hat{\beta}_{controls}\sim N(\beta_{G\times E},\sigma_{G\times E}^2),$$

with $\hat{\sigma}_{G\times E}^2=\sum_D\sum_G\sum_E\frac{1}{n_{DGE}}=\hat{\sigma}_{cases}^2+\hat{\sigma}_{controls}^2$.

Logistic regression is a very flexible approach in association studies and is therefore is widely used in genetic main effect, environmental main effect, and G×E and gene-gene interaction (G×G) studies. It allows for adjusted analysis by simple inclusion of additional covariables.

For a binary disease outcome such as case-control status, most existing association tests, including interaction tests, are based on logistic regression models. To test for the presence of G×E interaction for a SNP, one needs to construct a test statistic testing whether the null hypothesis (H_0) is followed for each SNP.

H_0: $\beta_{G\times E}$= 0, no G×E interaction at the SNP

The corresponding $\hat{\beta}_{G\times E}$ can be estimated from the data.

2.2.6. Statistical Tests for G×E Interaction in Case-Control Genome-Wide Association Studies

Case-Control Test

The classic case-control test (CC) for G×E interaction tests H_0 using the standard Wald-type test statistics, constructed for each SNP. This test statistic, T_{CC}, is distributed in an approximately standard normal fashion.

$$T_{CC} = \frac{\hat{\beta}_{G\times E}}{\hat{\sigma}_{G\times E}} = \frac{\hat{\beta}_{cases} - \hat{\beta}_{controls}}{\sqrt{\hat{\sigma}^2_{cases} + \hat{\sigma}^2_{controls}}} \sim N\ (\beta_{G\times E}, 1)$$

Case-Only Test

Piegorsch and colleagues proposed the case-only test (CO) for gene-environment interaction, seeking to achieve greater power than the case-control test (Piegorsch, Weinberg et al. 1994). They used equation (2.4) as a basis for their estimator of G×E interaction and additionally introduced two critical constraints to construct a valid test. They assume that the disease of interest is rare in the population and that G-E correlation is absent, i.e. genotypes and environment are independent and thus $OR_{controls}=1 \Rightarrow \beta_{controls}=0$. These assumptions allow the construction of a test statistic, which is distributed as $N(0,1)$ under H_0, and is characterized by a reduced variance and is therefore more powerful than the case-control test.

$$T_{CO} = \frac{\hat{\beta}_{cases}}{\hat{\sigma}_{cases}} = \frac{\hat{\beta}_{cases}}{\sqrt{\hat{\sigma}^2_{cases}}} \sim N\,(\beta_{G\times E}\,,1)$$

However, when the assumptions are violated, the case-only method leads to biased estimates and T_{CO} has highly inflated type I error rate. Thus, testing for significance is no longer trustworthy.

Mukherjee's Shrinkage Estimator

Mukherjee and Chatterjee proposed another method to test for G×E interaction, relying on empirical Bayes models (please refer to empirical Bayes in the subsequent sections). They named the G×E interaction estimator based on their approach an empirical Bayes type shrinkage estimator for G×E and introduced the corresponding test statistic (MUK-EB), (Mukherjee and Chatterjee 2008). The MUK-EB estimator combines the robust case-control and powerful case-only estimators into a single estimator as

$$\hat{\beta}_{MUK-EB} = (1-B)\hat{\beta}_{cases} + B\hat{\beta}_{G\times E}.$$

The weight *B* is chosen according to the evidence in the data on the G-E correlation. If G-E is present in the controls then B→1 and $\hat{\beta}_{MUK-EB}$ converges to $\hat{\beta}_{G\times E}$. When no evidence of G-E is present, then B→0 and $\hat{\beta}_{MUK-EB}$ converges to $\hat{\beta}_{cases}$.

To derive the shrinkage factor *B*, Mukherjee and Chatterjee demonstrated that the G-E correlation for each SNP can be modeled by the use of $\hat{\beta}_{controls}$. They used a variance parameter τ^2 representing the degree of uncertainty with respect to G-E correlation per SNP.

$$\hat{\beta}_{controls} \mid \beta_{controls} \sim N(\beta_{controls}, (\sigma_{controls})^2)$$

$$\beta_{controls} \mid \tau^2 \sim N(0, \tau^2)$$

and estimate the parameter τ^2, by $\hat{\tau}^2 = \widehat{\beta}^2_{controls}$.

Therefore

$$\hat{\beta}_{MUK-EB} = \frac{(\widehat{\sigma_{G\times E}})^2}{\hat{\tau}^2 + (\widehat{\sigma}_{G\times E})^2}\hat{\beta}_{cases} + \frac{\hat{\tau}^2}{\hat{\tau}^2 + (\widehat{\sigma_{G\times E}})^2}\hat{\beta}_{G\times E}.$$

This estimator, even though derived from the Bayesian perspective, is neither Bayes nor empirical Bayes, but a pure function of the observed data.

The variance is estimated by the following term:

$$\hat{\sigma}_{MUK-EB} = \hat{\sigma}^2_{cases} + \left(\frac{\hat{\beta}^2_{controls}\left(\hat{\beta}^2_{controls} + 3\hat{\sigma}^2_{G\times E}\right)}{\left(\hat{\beta}^2_{controls} + \hat{\sigma}^2_{G\times E}\right)^2}\right)^2 \hat{\sigma}^2_{G\times E}$$

The Wald-type test statistic can be constructed for MUK-EB as follows

$$T_{MUK-EB} = \frac{\hat{\beta}_{MUK-EB}}{\hat{\sigma}_{MUK-EB}} \sim N\,(\beta_{G\times E}\,, 1)$$

The MUK-EB test was shown to be more powerful than the case-only test. At the same time the type I error for MUK-EB is substantially less inflated as for the case-only test, which makes the test applicable to significance testing in the presence of G-E correlation (Mukherjee and

Chatterjee 2008). However, the type I error of MUK-EB can still be inflated in the presence of a large number of G-E correlations in the source population (Mukherjee, Ahn et al. 2008).

Murcray's Two Step Approach

Murcray and colleagues introduced a two-step procedure to test for G×E interaction (MUR). At the first step, they proposed to screen for the correlation between genotype and environment irrespective of the disease status by

$$logit\big(\mathrm{P}(\mathrm{E}=1|\mathrm{G})\big)=\alpha_{all}+\beta_{all}G$$

Therefore, $OR_{all}=\frac{(p_{100}+p_{000})(p_{1G1}+p_{0G1})}{(p_{101}+p_{001})(p_{1G0}+p_{0G0})}$, for G=1 or 2 and data from **Table 2.2**.

The maximum likelihood estimate (MLE) of β_{all},

$$\hat{\beta}_{all}=log\left(\frac{(n_{100}+p_{000})(n_{1G1}+n_{0G1})}{(n_{101}+n_{001})(n_{1G0}+n_{0G0})}\right) \sim N(\beta_{all},\sigma^2_{all}),\ \hat{\sigma}^2_{all}=\sum_G\sum_E\frac{1}{(n_{0GE}+n_{1GE})}.$$

The first step test statistic is

$$T_{MUR}=\frac{\hat{\beta}_{all}}{\hat{\sigma}_{all}}\sim N(\beta_{all},1)$$

Only SNPs passing the first step proceed to the second step, meaning only those SNPs with $\hat{\beta}_{all}$ significantly different from zero. At the second step of the MUR procedure, SNPs passed on from step one are tested for G×E interaction using the classic case-control test. Since steps one and two are independent, the overall procedure provides a valid test for interaction (Murcray, Lewinger et al. 2009). Furthermore, given that a substantially reduced number of SNPs is passed to the second level compared to the CC or CO test, the common Bonferroni multiple testing adjustments are performed based on that number of second level SNPs, leading to the power gain over the CC test. However, the power of the test depends on the case-control

ratio. An excess number of controls compared to cases leads to an overall loss in power for the MUR method. Nevertheless, Murcray's two step method keeps type I error at the nominal level.

Chapter 3

3. Population Stratification in Studies of G×E Interaction

A large number of naturally occurring populations are heterogeneous and stratified, meaning that a population is composed of discrete homogeneous subpopulations or continuous admixture is present. These subpopulations have possibly different ethnic backgrounds and therefore different genetic makeup as well as environmental exposures. For such admixed populations with uncovered substructure, the assumption of G-E independence is often violated as a result of confounding or effect modification. However, within a specific substratum, the assumption of independence may still hold.

Quite a few statistical methods have been proposed to estimate G×E interaction in large-scale case-control studies, including those described in Chapter 2. However, not all of the proposed methods are robust to the presence of hidden substructure in the study sample, such as population stratification. As a consequence, their performance often leads to biased effect estimates. Unfortunately, population stratification (PS) is not easily identifiable and is hard to control for using classic approaches such as matching or stratified analysis. The extent of population stratification bias depends on certain characteristics of the study sample, specifically on the number of admixed ethnicities, differences in genotype and exposure frequencies, and differences in disease prevalence across the strata.

In the following, we derive an equation to measure the theoretical population stratification bias of G×E interaction in a case-control design. We investigated the magnitude of the bias due to population stratification for G×E interaction in case-control studies and compared estimates of G×E interaction to the genetic main effect estimates and to the case-only estimates of interaction in terms of robustness to the presence of PS. An analytical study of various realistic situations was performed to measure population stratification bias using our derived equation.

We named this measure of bias the confounding interaction ratio for case-control estimator of G×E interaction (CIR_{CC}). We used equations derived earlier to evaluate the bias of the genetic main effect by the so-called confounding rate ratio (CRR) (Lee and Wang 2008) and of G×E interaction in a case-only design by the confounding interaction ratio for the case-only estimator of G×E interaction (CIR_{CO}) (Wang and Lee 2008).

Furthermore, we compared four common methods for G×E interaction (Chapter 2) in terms of their robustness to the presence of population stratification in the study sample. We performed a simulation study for a set of different scenarios admixing similar or more divergent sub-populations. The EHB-GE_{CHI} approach was compared to CC, CO MUK-EB (see Chapter 1).

We also evaluated the ability of principal component analysis (PCA) (Price, Patterson et al. 2006) as integrated in the EIGENSOFT statistical package to correct for population stratification bias in studies of G×E interaction.

3.1. Measures of Population Stratification Bias

3.1.1. Notation

Assume, that a study population consists of $j = 1 \ldots J$ discrete subpopulations. Let $E(\overline{E})$ and $G(\overline{G})$ denote the presence (absence) of the exposure and of the susceptibility genotype for a person. We define p_j to be the prevalence of the environment *E*, q_j to be the frequency of the susceptible genotype *G*, and b_j to be the background disease risk (a risk for non-carriers of the risk allele, unexposed to the environment in subpopulation *j*). Then, $e_j = \frac{p_j}{1-p_j}$ denote the exposure prevalence odds and $g_j = \frac{q_j}{1-q_j}$ denotes the genotype frequency odds. Let n_j denote the total number of individuals in the *jth* subgroup. Note that in Section 3.1 we consider the

situation of collecting all individuals (cases and controls) from the entire population. This gives us the opportunity to estimate risk in the case-control or the case-only study settings instead of operating only with odds ratios. Please note that certain notation is redefined newly for each section of this dissertation and is valid only for that particular section.

3.1.2. Confounding Rate Ratio for Case-Control Design and Confounding Interaction Ratio for the Case-Only Design

Lee and Wang in (Lee and Wang 2008) derived an equation to quantify population stratification bias for genetic main effect estimation in a case-control study. They termed the newly introduced measure of population stratification bias the confounding rate ratio (CRR). Here we outline their derivation. Let RR_G denote the relative risk of disease for individuals carrying the susceptibility genotype as compared to those who do not. Assume RR_G is constant across the strata, meaning that in this case population stratification is a confounder only and is not an effect modifier. In the total population, the disease rate for a person carrying the susceptible genotype is

$$DR_G = \frac{\sum_{j=1}^{J} n_j q_j b_j RR_G}{\sum_{j=1}^{J} n_j q_j},$$

and for those who do not

$$DR_{\bar{G}} = \frac{\sum_{j=1}^{J} n_j (1-q_j) b_j}{\sum_{j=1}^{J} n_j (1-q_j)}.$$

The confounded relative risk RR_G^C is defined as the overall risk in the admixed population

$$RR_G^C = \frac{DR_G}{DR_{\bar{G}}}.$$

Define weights $w_j = \frac{n_j(1-q_j)}{\sum_k n_k(1-q_k)}$, and finally define the confounded rate ratio CRR as follows:

$$\mathrm{CRR} = \frac{\mathrm{RR}_{\mathrm{G}}^{\mathrm{c}}}{\mathrm{RR}_{\mathrm{G}}} = \frac{\sum_{\mathrm{j}=1}^{\mathrm{J}} \mathrm{w_j g_j b_j}}{\sum_{\mathrm{j}=1}^{\mathrm{J}} \mathrm{w_j b_j} \sum_{\mathrm{j}=1}^{\mathrm{J}} \mathrm{w_j g_j}} \quad \textbf{(3.1)}$$

To introduce a measure of population stratification bias in case-only studies of G×E interaction, we present once again the work of Wang and Lee described in (Wang and Lee 2008). Let $\mathrm{RR_{GE}}$ denote the relative risk of disease for those subjects with (G,E) compared to $(\overline{\mathrm{G}}, \overline{\mathrm{E}})$ individuals. Similarly $\mathrm{RR_G}$ denotes the relative risk of disease for individuals with $(\mathrm{G}, \overline{\mathrm{E}})$ compared to $(\overline{\mathrm{G}}, \overline{\mathrm{E}})$ and $\mathrm{RR_E}$ denotes the relative risk of disease for individuals with $(\overline{\mathrm{G}}, \mathrm{E})$ compared to $(\overline{\mathrm{G}}, \overline{\mathrm{E}})$. Let $\mathrm{RR_{GE}}$, $\mathrm{RR_G}$, and $\mathrm{RR_E}$ be constant across the strata of admixed population. Once again, PS is acting as confounder here and not like an effect modifier (Chapter 1). Assume that genotype and environment are independent within each stratum, for validity of the case-only G×E estimate. The G×E interaction effect on the multiplicative scale can be measured by $\mathrm{RR_{G\times E}} = \frac{\mathrm{RR_{GE}}}{\mathrm{RR_G RR_E}}$ (Chapter 2) and can be estimated by the case-only approach. If a study collects each and every case in the whole population, then the number of disease carriers would be $\mathrm{n_{GE}} = \sum_{\mathrm{j}=1}^{\mathrm{J}} \mathrm{n_j q_j p_j b_j\, RR_{GE}}$, $\mathrm{n_{\overline{G}E}} = \sum_{\mathrm{j}=1}^{\mathrm{J}} \mathrm{n_j}(1-\mathrm{q_j})\mathrm{p_j b_j\, RR_E}$, $\mathrm{n_{G\overline{E}}} = \sum_{\mathrm{j}=1}^{\mathrm{J}} \mathrm{n_j q_j}(1-\mathrm{p_j})\mathrm{b_j\, RR_G}$, and $\mathrm{n_{\overline{G}\overline{E}}} = \sum_{\mathrm{j}=1}^{\mathrm{J}} \mathrm{n_j}(1-\mathrm{q_j})(1-\mathrm{p_j})\mathrm{b_j}$

for $(\mathrm{G}, \mathrm{E}), (\overline{\mathrm{G}}, \mathrm{E}), (\mathrm{G}, \overline{\mathrm{E}}), (\overline{\mathrm{G}}, \overline{\mathrm{E}})$ subjects, respectively.

Thus, the confounded G×E interaction effect can be estimated by the case-only approach as

$$\mathrm{RR}_{\mathrm{G\times E}}^{\mathrm{c}} = \frac{\mathrm{n_{GE}} \mathrm{n_{\overline{G}\overline{E}}}}{\mathrm{n_{\overline{G}E}} \mathrm{n_{G\overline{E}}}} = \frac{\sum_{\mathrm{j}} \mathrm{n_j q_j p_j b_j RR_{GE}}}{\sum_{\mathrm{j}} \mathrm{n_j q_j}(1-\mathrm{p_j})\mathrm{b_j RR_G}} \frac{\sum_{\mathrm{j}} \mathrm{n_j}(1-\mathrm{q_j})(1-\mathrm{p_j})\mathrm{b_j}}{\sum_{\mathrm{j}} \mathrm{n_j}(1-\mathrm{q_j})\mathrm{p_j b_j RR_E}}$$

Therefore, the confounding interaction ratio for the case-only estimator of G×E interaction ($\mathrm{CIR_{CO}}$) is defined as

$$\mathrm{CIR_{CO}} = \frac{\mathrm{RR}_{\mathrm{INT}}^{\mathrm{c}}}{\mathrm{RR_{INT}}} = \frac{\sum_{\mathrm{j}} \mathrm{w_j}(\mathrm{e_j}-\overline{\varphi}_{\mathrm{E}})(\mathrm{g_j}-\overline{\varphi}_{\mathrm{G}})}{\mathrm{SD}(\varphi_{\mathrm{E}})\times\mathrm{SD}(\varphi_{\mathrm{G}})} \frac{\mathrm{SD}(\varphi_{\mathrm{E}})}{\overline{\varphi}_{\mathrm{E}}} \frac{\mathrm{SD}(\varphi_{\mathrm{G}})}{\overline{\varphi}_{\mathrm{G}}} + 1 = \mathrm{r_{GE}} \mathrm{CV_G} \mathrm{CV}_E + 1 \quad \textbf{(3.2)}$$

where $w_j = \frac{n_j(1-p_j)(1-q_j)b_j}{\sum_{k=1}^{J} n_k(1-p_k)(1-q_k)b_k}$ are weights, $\overline{\varphi}_E = \sum_{j=1}^{J} w_j e_j$, $\overline{\varphi}_G = \sum_{j=1}^{J} w_j g_j$, $\overline{\varphi}_E, \overline{\varphi}_G$ denote the means, $SD(\varphi_E) = \sqrt{\sum_{j=1}^{J} w_j (e_j - \overline{\varphi}_E)^2}$ and $SD(\varphi_G) = \sqrt{\sum_{j=1}^{J} w_j (g_j - \overline{\varphi}_G)^2}$ the standard deviations, CV_E and CV_G denote the coefficients of variation of the exposure prevalence odds and the genotype frequency odds, respectively, and r_{GE} denotes the correlation coefficient between the exposure prevalence odds and genotype frequency odds.

3.1.3. Derivation of Confounding Interaction Ratio for the Case-Control Design

To derive an equation for the confounded interaction ratio for a case-control study, CIR_{CC}, we followed the method and used the notation as described above (Lee and Wang 2008, Wang and Lee 2008). In a case-control study, the G×E interaction effect on the multiplicative scale can be measured by $RR_{G\times E} = \frac{RR_{GE}}{RR_G RR_E}$. Note that in this section we newly redefine DR, RR, CV, SD and all notations from the previous section accordingly for the case-control design. Assume that the study was able to collect all the affected subjects, from here on cases, and controls from the entire population. Thus, we are still deriving risks and not odds ratios. Let DR_G denote disease rate for carriers of a susceptible genotype given absence of any environmental factor in the whole population and $DR_{\overline{G}}$ for non-carriers. Then $RR_G^C = \frac{DR_G}{DR_{\overline{G}}}$ is the confounded relative risk for carriers of the genotype in the absence of environmental exposure compared to the non-carriers,

where $DR_G = \frac{\sum_{j=1}^{J} n_j q_j (1-p_j) b_j RR_G}{\sum_{j=1}^{J} n_j q_j (1-p_j)}$ and $DR_{\overline{G}} = \frac{\sum_{j=1}^{J} n_j (1-q_j)(1-p_j) b_j}{\sum_{j=1}^{J} n_j (1-q_j)(1-p_j)}$.

In the same manner, let DR_E denote the disease rate for an individual exposed to the environment at the reference level of the genotype and $DR_{\overline{E}}$ for those unexposed, therefore

$$RR_E^C = \frac{DR_E}{DR_{\overline{E}}},$$

where $DR_E = \frac{\sum_{j=1}^{J} n_j(1-q_j)p_j b_j \, RR_E}{\sum_{j=1}^{J} n_j(1-q_j)p_j}$ and $DR_{\overline{E}} = \frac{\sum_{j=1}^{J} n_j(1-q_j)(1-p_j)b_j}{\sum_{j=1}^{J} n_j(1-q_j)(1-p_j)}$.

The confounded relative risk RR_{GE}^C for exposed carriers of the susceptible genotype is now

$$RR_{GE}^C = \frac{DR_{GE}}{DR_{\overline{GE}}}, \text{ where } DR_{GE} = \frac{\sum_{j=1}^{J} n_j q_j p_j b_j RR_{GE}}{\sum_{j=1}^{J} n_j q_j p_j} \text{ and } DR_{\overline{GE}} = \frac{\sum_{j=1}^{J} n_j(1-q_j)(1-p_j)b_j}{\sum_{j=1}^{J} n_j(1-q_j)(1-p_j)}.$$

Finally, we define the confounded interaction effect as follows

$$RR_{G\times E}^C = \frac{RR_{GE}^c}{RR_G^c RR_E^c} = \frac{DR_{GE}/DR_{\overline{GE}}}{DR_G/DR_{\overline{G}} \; DR_E/DR_{\overline{E}}}.$$

Thus, the ratio of the confounded effect to the true effect of G×E interaction CIR_{CC} in a case-control G×E interaction study is defined as

$$CIRcc = \frac{RR_{G\times E}^c}{RR_{G\times E}} =$$

$$= \left(\frac{\sum_j n_j q_j p_j b_j}{\sum_j n_j q_j p_j} \frac{\sum_j n_j(1-q_j)(1-p_j)}{\sum_j n_j(1-q_j)(1-p_j)b_j}\right) \Big/ \left(\frac{\sum_j n_j q_j (1-p_j) b_j}{\sum_j n_j q_j (1-p_j)} \cdot \frac{\sum_j n_j(1-q_j)(1-p_j)}{\sum_j n_j(1-q_j)(1-p_j)b_j} \frac{\sum_j n_j(1-q_j)p_j b_j}{\sum_j n_j(1-q_j)p_j} \frac{\sum_j n_j(1-q_j)(1-p_j)}{\sum_j n_j(1-q_j)(1-p_j)b_j}\right)$$

Define weights w_j by

$$w_j = \frac{n_j(1-p_j)(1-q_j)}{\sum_k n_k(1-p_k)(1-q_k)},$$

then

$$DR_G/DR_{\overline{G}} = \frac{\sum_j w_j g_j b_j}{\sum_j w_j g_j \sum_j w_j b_j},$$

$$DR_E/DR_{\overline{E}} = \frac{\sum_j w_j e_j b_j}{\sum_j w_j e_j \sum_j w_j b_j},$$

$$DR_{GE}\Big/DR_{\overline{GE}} = \frac{\sum_j w_j e_j g_j b_j}{\sum_j w_j e_j g_j \sum_j w_j b_j}.$$

Therefore,

$$\frac{R_{GE}\Big/R_{\overline{GE}}}{R_G\Big/R_{\overline{G}}\;\;R_E\Big/R_{\overline{E}}} = \frac{\sum_j w_j e_j g_j b_j}{\sum_j w_j e_j g_j \sum_j w_j b_j} \Big/ \left(\frac{\sum_j w_j g_j b_j}{\sum_j w_j g_j \sum_j w_j b_j}\frac{\sum_j w_j e_j b_j}{\sum_j w_j e_j \sum_j w_j b_j}\right)$$

and

$$\mathrm{CIR_{CC}} = \frac{\sum_{j=1}^{J} w_j e_j g_j b_j}{\sum_{j=1}^{J} w_j e_j b_j \sum_{j=1}^{J} w_j g_j b_j}\frac{\sum_{j=1}^{J} w_j g_j \sum_{j=1}^{J} w_j e_j}{\sum_{j=1}^{J} w_j g_j b_j}\sum_{j=1}^{J} w_j b_j$$

Define separate weights for cases w_{j1} and for controls w_{j0}.

$$w_{j0} = \frac{n_j(1-p_j)(1-q_j)}{\sum_{k=1}^{J} n_k\,(1-p_k)(1-q_k)}$$

$$w_{j1} = \frac{n_j(1-p_j)(1-q_j)b_j}{\sum_{k=1}^{J} n_k\,(1-p_k)(1-q_k)b_k}$$

Let $\overline{\varphi}_E^0 = \sum_{j=1}^{J} w_{j0} e_j$, $\overline{\varphi}_E^1 = \sum_{j=1}^{J} w_{j1} e_j\, b_j$, $\overline{\varphi}_G^0 = \sum_{j=1}^{J} w_{j0} g_j$, $\overline{\varphi}_G^1 = \sum_{j=1}^{J} w_{j1} g_j b_j$, $\overline{\varphi's}$ denote the means of genotype and exposure frequencies, subscript 1 refers to cases and 0 to controls.

SDs are standard deviations defined as

$$\mathrm{SD}(\varphi_E) = \sqrt{\sum_{j=1}^{J} w_j (e_j - \overline{\varphi}_E)^2} \text{ and } \mathrm{SD}(\varphi_G) = \sqrt{\sum_{j=1}^{J} w_j (g_j - \overline{\varphi}_G)^2},$$

then

$$\mathrm{CIRcc} = \frac{\frac{\sum_j w_{j1}\left(e_j - \overline{\varphi}_E^1\right)\left(g_j - \overline{\varphi}_G^1\right)}{\mathrm{SD}(\varphi_E^1)\mathrm{SD}(\varphi_G^1)}\frac{\mathrm{SD}(\varphi_E^1)}{\overline{\varphi}_E^1}\frac{\mathrm{SD}(\varphi_G^1)}{\overline{\varphi}_G^1} + 1}{\frac{\sum_j w_{j0}\left(e_j - \overline{\varphi}_E^0\right)\left(g_j - \overline{\varphi}_G^0\right)}{\mathrm{SD}(\varphi_E^0)\mathrm{SD}(\varphi_G^0)}\frac{\mathrm{SD}(\varphi_E^0)}{\overline{\varphi}_E^0}\frac{\mathrm{SD}(\varphi_G^0)}{\overline{\varphi}_G^0} + 1}$$

Finally, following some simplification this leads to

$$\mathrm{CIRcc} = \frac{r_{GE1} CV_{G1} CV_{E1} + 1}{r_{GE0} CV_{G0} CV_{E0} + 1} \qquad \textbf{(3.3)}$$

where CV_E and CV_G are the coefficients of variation of the exposure prevalence odds and the genotype frequency odds; and r_{GE} is the correlation coefficient between the exposure prevalence odds and genotype frequency odds occurrence. Note that the mathematical form of CIR_{CC} derived here is similar to the CIR_{CO} measures of population stratification bias for the case-only design (Wang and Lee 2008), see previous section. It can be seen from the equation that there would be no population stratification bias when the exposure prevalence odds and the genotype frequency odds are uncorrelated in cases and controls, when there is no variation in the exposure prevalence odds, or when there is no variation in the genotype frequency odds across subpopulations. For CIR_{CC}, overestimation of RR ($CIR_{CC} > 1$) occurs when genotype and exposure are negatively correlated. Underestimation ($CIR_{CC} < 1$) occurs when exposure and genotype have positive correlation and the range of the background disease risks is considerably smaller than the range of both genotype and exposure frequencies. For CIR_{CO}, overestimation ($CIR_{CO} > 1$) of the parameter occurs when genotype and exposure are positively correlated, while underestimation ($CIR_{CO} < 1$) occurs when exposure and genotype are negatively correlated (Wang and Lee 2008).

3.1.4. Calculation Settings

To investigate the potential size of the confounding interaction ratio in a case-control study, CIR_{CC}, we calculated this measure over a range of realistic scenarios. Generally, we followed the procedures described by Wacholder et al. (Wacholder, Rothman et al. 2000), and Wang and Lee (Amos, Wu et al. 2008). Additionally, we investigated the bias for samples including *j = 2, 3, 5* or *8* subpopulations. For each scenario we assumed that there are *j* = 2, 3, 5 or *8*

strata each of equal size. We allowed for different genotype and exposure frequencies and different baseline disease risks across the strata. Genotype q_j and exposure frequencies p_j were set to one of the three intervals *0.01–0.3*, *0.1-0.4* and *0.3–0.6*. These intervals reflect the range of frequencies of different alleles for a large number of genes in European populations reported in (Cavalli-Sforza, Menozzi et al. 1994). The background risk of the disease b_j was chosen from intervals *1.0–1.5* or *1.0–3.0,* representing a realistic range of cancer rates among Europeans (Wacholder, Rothman et al. 2000). In each of the corresponding intervals, values of the genotype and exposure frequencies were set to be equally distant on the *logit* scale and values of the baseline disease risk were set to be equally distant on the *logarithmic* scale. Therefore, we obtained unique values of three parameters for each stratum. Such a choice of the parameters is unique for each interval.

Next, we calculated the bias due to population stratification employing the following approach: We set eight values for the exposure and genotype frequencies as well as for baseline disease risk as described above from the corresponding intervals. Then for *j=2* subpopulations we considered all possible combinations of the corresponding pairs of genotype and exposure frequencies and disease risks out of eight possible values for each parameter (all possible combinations of two values in each interval out of eight). For *j=3* subpopulations; we considered all possible combinations of triples from eight values of genotype, exposure frequencies, and background disease risks. To investigate both possible situations when genotype and exposure are positively and negatively correlated, we fixed disease risks and randomly permuted values for genotype and exposure frequencies in triples. For *j=5* subpopulations, we repeated the procedure as described for *j=3*, but for combinations of five values of parameters. For *j=8* subpopulations, we fixed the background disease risk and randomly permuted eight values for the genotype frequencies and eight values for the exposure

prevalence. Finally, we found the distribution of CIR_{CC}, CIR_{CO}, and CRR for each of the 18 scenarios and obtained the minimum, maximum, and the quartiles of its distribution.

3.1.5. Results

Table 3.2 summarizes the results for CIR_{CC} calculated for the admixture of two and eight subpopulations. The bias due to population stratification on average does not reach alarming values for the G×E interaction term in a case-control design, meaning it is always below 10%. However it can stretch up to 50% in the situations in which the ranges of genotype frequency, exposure prevalence and background disease risks are wide, such as for example in scenarios 10 to 14. To evaluate the degree of population stratification bias in case-only studies, we calculated CIR_{CO} for the same 18 scenarios. The results are summarized **in Table 3.3**. It is clear that the case-control estimator of G×E interaction is more robust to the presence of population stratification compared to the case-only estimator for all considered scenarios. On average, the degree of population stratification bias in a case-control study is tolerable. However, it can reach 50% or higher for the case-only estimator. Comparison of **Table 3.2** and **Table 3.3** demonstrates that the bias due to population stratification of the case-control estimator depends on the range of the background disease risks across the strata. In contrast, this statement is false for the case-only estimator. Calculations of CIR_{CC} for the admixture of 3 and 5 strata are presented in **Table 3.4**.

Wacholder (Wacholder, Rothman et al. 2000) mentioned that the bias of the interaction term is generally bigger than the bias in genetic main effects. We investigated the situations in which the population stratification bias of G×E interaction effect estimates were greater, smaller, or comparable to genetic main effects. We calculated the bias in main effects for the same set of 18 scenarios as in **Table 3.2** using CRR as a measure of the population stratification bias. The

results for scenarios 1 to 18 are represented graphically in **Figure 3.1** to **Figure 3.5**. This reveals that population stratification bias decreases for each scenario and for all three measures of bias (CRR, CIR_{CC}, CIR_{CO}) when the number of admixed subpopulations increases in the study sample from 2 to 8. The largest bias appears for the admixture of 2 subgroups. We can see that CIR_{CC} is greater than CRR in scenarios 1 and 10, is smaller in scenarios 4, 7, 8, 13, 14, 15, 16, 17, 18 and is comparable in scenarios 2, 3, 5, 6, 9, 12. Therefore, bias as measured by CIR_{CC} is greater than that for CRR when the exposure prevalence range in terms of variation of the odds ratios of the largest and the smallest values are extremely disparate. CIR_{CC} is generally smaller than CRR when genotype frequency range in terms of the variation of the odds ratios is considerably wider than the exposure prevalence odds ratios range. Finally, CIR_{CC} is comparable in size to the CRR when the genotype frequency odds ratios range is similar to the exposure frequency odds ratios range across the strata.

The grey-shaded areas in **Figure 3.1** to **Figure 3.5** represent theoretical bounds for CIR_{CC}, CIR_{CO}, and CRR, derived in (Amos, Wu et al. 2008, Lee and Wang 2008). We calculated theoretical bounds for CIR_{CC} in the same way. **Table 3.1** presents equations to calculate the corresponding lower (L) and upper (U) theoretical bound. We do not provide details on the boundary derivations, because they were derived in the same way as already published. In contrast to the bias in the case-only design (CIR_{CO}), the magnitude of variation in background disease risk affects the degree of the population stratification bias for both CIR_{CC} and the CRR. The bias is larger for a larger variation in the disease prevalence (scenarios 10 to 18). It is clear from the figures that the case-control design is significantly more robust to population stratification than the case-only design.

The bias of G×E interaction effect due to population stratification is usually small. However, it can still reach extreme values in realistic situations even for the robust case-control design, for example, when two divergent subpopulations are admixed.

Table 3.1 Theoretical bounds for CRR, CIR_{CC} and CIR_{CO}

bounds of CRR	$U = \frac{\sqrt{Q \times B} \times (\sqrt{Q \times B} + 1)^2}{(\sqrt{Q \times B} + Q) \times (\sqrt{Q \times B} + B)}$	$L = \frac{1}{U}$
bounds of CIR_{CC}	$U = B^2$	$L = \frac{1}{U}$
bounds of CIR_{CO}	$U = \frac{\sqrt{Q \times P} \times (\sqrt{Q \times P} + 1)^2}{(\sqrt{Q \times P} + Q) \times (\sqrt{Q \times P} + P)}$	$L = \frac{1}{U}$

U, theoretical upper bound; *L*, theoretical lower bound; $Q=max(g_j)/min(g_j)$; $P=max(e_j)/min(e_j)$; $B=max(b_j)/min(b_j)$; *j*, subgroup indicator; g_j, genotype frequency odds; e_j, exposure frequency odds; b_j, background disease risk;

Table 3.2 Confounding interaction ratio for case-control CIR_{CC}, evaluated for 18 scenarios admixture of 2 and 8 subpopulations

Scenario	Parameter intervals			CIR_{CC} 2 from 8 **		CIR_{CC} for 100 000 simulations of random permutation of 8 values***				
	b_j	p_j	q_j	min	max	min	25th	50th	75th	max
1	1.0-1.5	0.01-0.3	0.01-0.3	0.80	1.28	0.86	0.97	1.00	1.03	1.16
2	1.0-1.5	0.01-0.3	0.10-0.4	0.86	1.16	0.92	0.98	1.00	1.01	1.08
3	1.0-1.5	0.01-0.3	0.30-0.6	0.90	1.11	0.94	0.99	1.00	1.01	1.06
4	1.0-1.5	0.1-0.4	0.01-0.3	0.86	1.16	0.92	0.98	1.00	1.01	1.08
5	1.0-1.5	0.1-0.4	0.1-0.4	0.91	1.07	0.96	0.99	1.00	1.01	1.04
6	1.0-1.5	0.1-0.4	0.3-0.6	0.95	1.05	0.97	0.99	1.00	1.01	1.03
7	1.0-1.5	0.3-0.6	0.01-0.3	0.90	1.11	0.94	0.99	1.00	1.01	1.06
8	1.0-1.5	0.3-0.6	0.1-0.4	0.95	1.05	0.97	0.99	1.00	1.01	1.03
9	1.0-1.5	0.3-0.6	0.3-0.6	0.98	1.01	0.98	1.00	1.00	1.00	1.02
10	1.0-3.0	0.01-0.3	0.01-0.3	0.59	1.97	0.61	0.93	1.00	1.10	1.47
11	1.0-3.0	0.01-0.3	0.1-0.4	0.68	1.46	0.79	0.95	1.00	1.05	1.24
12	1.0-3.0	0.01-0.3	0.3-0.6	0.74	1.34	0.85	0.96	1.00	1.04	1.17
13	1.0-3.0	0.1-0.4	0.01-0.3	0.68	1.46	0.78	0.96	1.00	1.05	1.23
14	1.0-3.0	0.1-0.4	0.1-0.4	0.77	1.15	0.89	0.98	1.00	1.03	1.12
15	1.0-3.0	0.1-0.4	0.3-0.6	0.84	1.17	0.92	0.98	1.00	1.02	1.09
16	1.0-3.0	0.3-0.6	0.01-0.3	0.74	1.34	0.85	0.97	1.00	1.04	1.17
17	1.0-3.0	0.3-0.6	0.1-0.4	0.84	1.17	0.92	0.98	1.00	1.02	1.08
18	1.0-3.0	0.3-0.6	0.3-0.6	0.91	1.08	0.94	0.99	1.00	1.01	1.06

b_j, disease risk ratio, q_j, genotype frequency; p_j, exposure frequency; both G and E ranges are spaced to be equidistant on the logarithmic scale; * study cohort consists of 2 discrete, admixed populations; ** study cohort consists of 8 discrete, admixed populations; min, minimum of CIR_{CC}; max, maximum of CIR_{CC}; 25^{th}, 50^{th}, 75^{th}, percentile of the CIR_{CC};

Table 3.3 Confounding interaction ratio for case-only CIR_{CO}, evaluated for 18 scenarios admixture of 2 and 8 subpopulations

Scenario	Parameters			CIR_{CO} 2 from 8 *		CIR_{CO} for 100 000 simulations of random permutation of 8 values**				
	b_j	p_j	q_j	min	max	min	25th	50th	75th	max
1	1.1-1.5	0.01-0.3	0.01-0.3	0.09	3.5	0.29	.69	0.95	1.31	2.59
2	1.1-1.5	0.01-0.3	0.1-0.4	0.3	2.57	0.55	0.83	0.98	1.18	1.77
3	1.1-1.5	0.01-0.3	0.3-0.6	0.45	2.16	0.67	0.89	0.99	1.13	1.53
4	1.1-1.5	0.1-0.4	0.01-0.3	0.3	2.57	0.53	0.83	0.98	1.18	1.76
5	1.1-1.5	0.1-0.4	0.1-0.4	0.5	1.96	0.73	0.92	1.00	1.09	1.39
6	1.1-1.5	0.1-0.4	0.3-0.6	0.6	1.67	0.80	0.94	1.00	1.07	1.28
7	1.1-1.5	0.3-0.6	0.01-0.3	0.45	2.16	0.66	0.89	0.99	1.13	1.54
8	1.1-1.5	0.3-0.6	0.1-0.4	0.60	1.67	0.79	0.94	1.00	1.07	1.27
9	1.1-1.5	0.3-0.6	0.3-0.6	0.70	1.44	0.85	0.96	1.00	1.04	1.19
10	1.1-3.0	0.01-0.3	0.01-0.3	0.09	3.49	0.27	0.69	0.96	1.34	2.93
11	1.1-3.0	0.01-0.3	0.1-0.4	0.28	3.19	0.51	0.83	0.99	1.18	1.90
12	1.1-3.0	0.01-0.3	0.3-0.6	0.41	2.41	0.63	0.88	0.99	1.14	1.63
13	1.1-3.0	0.1-0.4	0.01-0.3	0.28	3.19	0.50	0.84	0.98	1.19	1.91
14	1.1-3.0	0.1-0.4	0.1-0.4	0.52	2.04	0.71	0.91	0.99	1.09	1.47
15	1.1-3.0	0.1-0.4	0.3-0.6	0.62	1.68	0.77	0.94	1.00	1.07	1.33
16	1.1-3.0	0.3-0.6	0.01-0.3	0.41	2.20	0.62	0.88	0.99	1.13	1.62
17	1.1-3.0	0.3-0.6	0.1-0.4	0.62	1.62	0.78	0.94	1.00	1.07	1.32
18	1.1-3.0	0.3-0.6	0.3-0.6	0.70	1.36	0.83	0.96	1.00	1.05	1.22

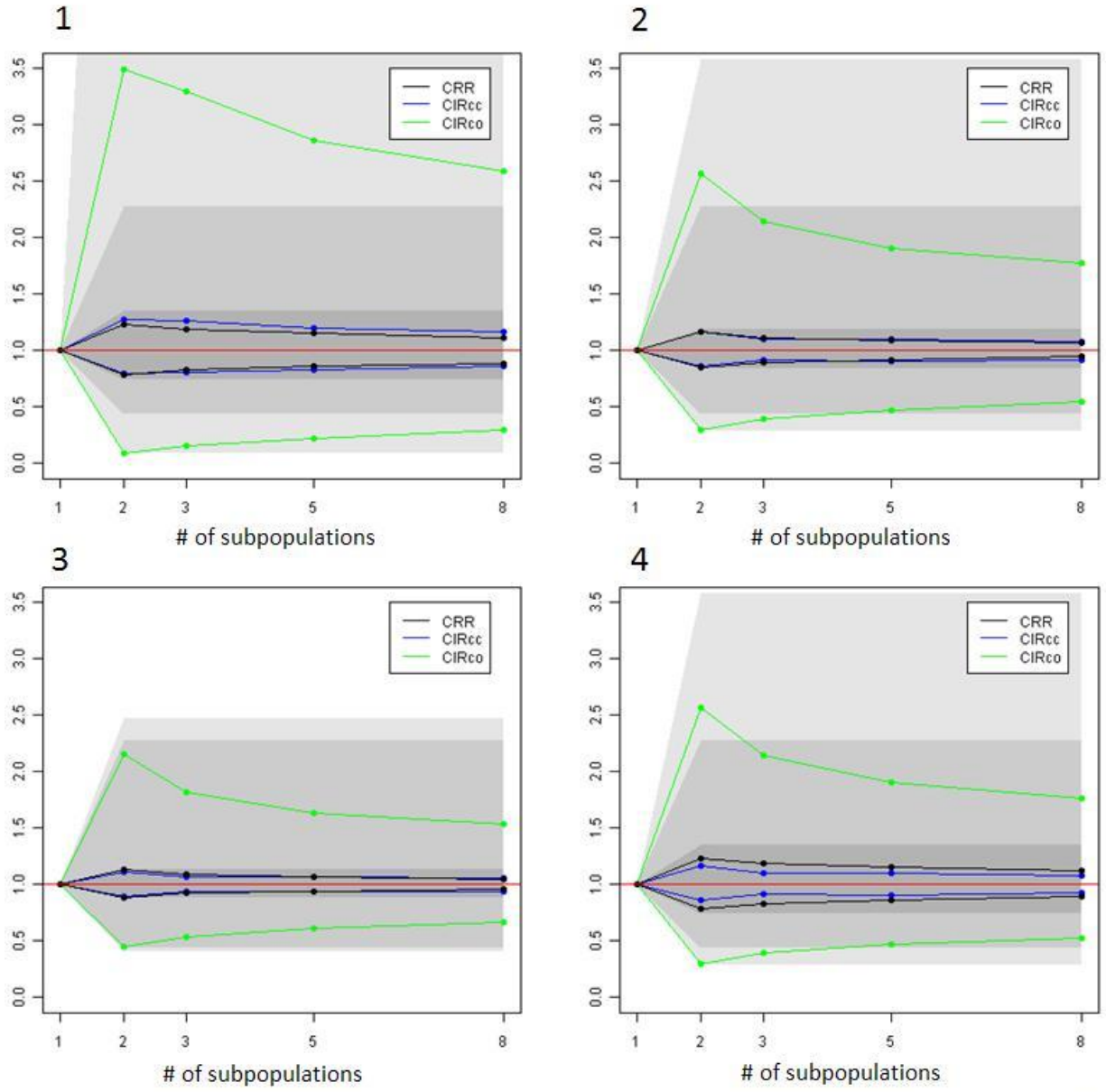
b_j, disease risk ratio, q_j, genotype frequency; p_j, exposure frequency; both G and E ranges are spaced to be equidistant on the logarithmic scale; * study cohort consists of 2 discrete, admixed populations; ** study cohort consists of 8 discrete, admixed populations; min, minimum of CIR_{CO}; max, maximum of CIR_{CO}; 25^{th}, 50^{th}, 75^{th}, percentile of the CIR_{CO};

Table 3.4 Confounding interaction ratio for case-control CIR_{CC}, evaluated for 18 scenarios, admixture of 3 and 5 subpopulations

Scenario	Parameters			CIR_{CC} for 100 000 simulations of random permutation for all possible combinations of 3 values out of 8*					CIR_{CC} for 100 000 simulations of random permutation for all possible combinations of 5 values out of 8**				
	b_j	p_j	q_j	min	25th	50th	75th	max	min	25th	50th	75th	max
1	1.0-1.5	0.01-0.3	0.01-0.3	0.8	0.98	1.00	1.03	1.17	0.85	0.97	0.99	1.02	1.17
2	1.0-1.5	0.01-0.3	0.1-0.4	0.89	0.99	1.00	1.02	1.11	0.91	0.98	0.99	1.01	1.08
3	1.0-1.5	0.01-0.3	0.3-0.6	0.91	0.99	1.00	1.01	1.08	0.94	0.99	1.00	1.01	1.05
4	1.0-1.5	0.1-0.4	0.01-0.3	0.89	0.99	1.00	1.02	1.11	0.93	0.99	1.00	1.01	1.09
5	1.0-1.5	0.1-0.4	0.1-0.4	0.92	0.99	1.00	1.01	1.07	0.95	0.99	1.00	1.01	1.05
6	1.0-1.5	0.1-0.4	0.3-0.6	0.94	0.99	1.00	1.01	1.05	0.97	0.99	1.00	1.00	1.04
7	1.0-1.5	0.3-0.6	0.01-0.3	0.91	0.99	1.00	1.01	1.08	0.94	0.99	1.00	1.01	1.06
8	1.0-1.5	0.3-0.6	0.1-0.4	0.94	0.99	1.00	1.01	1.05	0.96	0.99	1.00	1.01	1.04
9	1.0-1.5	0.3-0.6	0.3-0.6	0.96	1.00	1.00	1.00	1.04	0.97	1.00	1.00	1.00	1.03
10	1.0-3.0	0.01-0.3	0.01-0.3	0.57	0.95	1.01	1.09	1.59	0.63	0.91	0.97	1.06	1.68
11	1.0-3.0	0.01-0.3	0.1-0.4	0.72	0.97	1.00	1.04	1.35	0.75	0.94	0.98	1.03	1.28
12	1.0-3.0	0.01-0.3	0.3-0.6	0.79	0.98	1.00	1.03	1.22	0.81	0.96	0.98	1.02	1.23
13	1.0-3.0	0.1-0.4	0.01-0.3	0.72	0.97	1.00	1.04	1.35	0.77	0.96	1.00	1.04	1.30
14	1.0-3.0	0.1-0.4	0.1-0.4	0.82	0.98	1.00	1.02	1.2	0.85	0.97	1.00	1.02	1.17
15	1.0-3.0	0.1-0.4	0.3-0.6	0.86	0.99	1.00	1.01	1.14	0.89	0.98	1.00	1.02	1.13
16	1.0-3.0	0.3-0.6	0.01-0.3	0.79	0.98	1.00	1.03	1.22	0.83	0.97	1.00	1.03	1.24
17	1.0-3.0	0.3-0.6	0.1-0.4	0.86	0.99	1.00	1.01	1.14	0.90	0.98	1.00	1.02	1.10
18	1.0-3.0	0.3-0.6	0.3-0.6	0.89	0.99	1.00	1.01	1.10	0.91	0.99	1.00	1.01	1.08

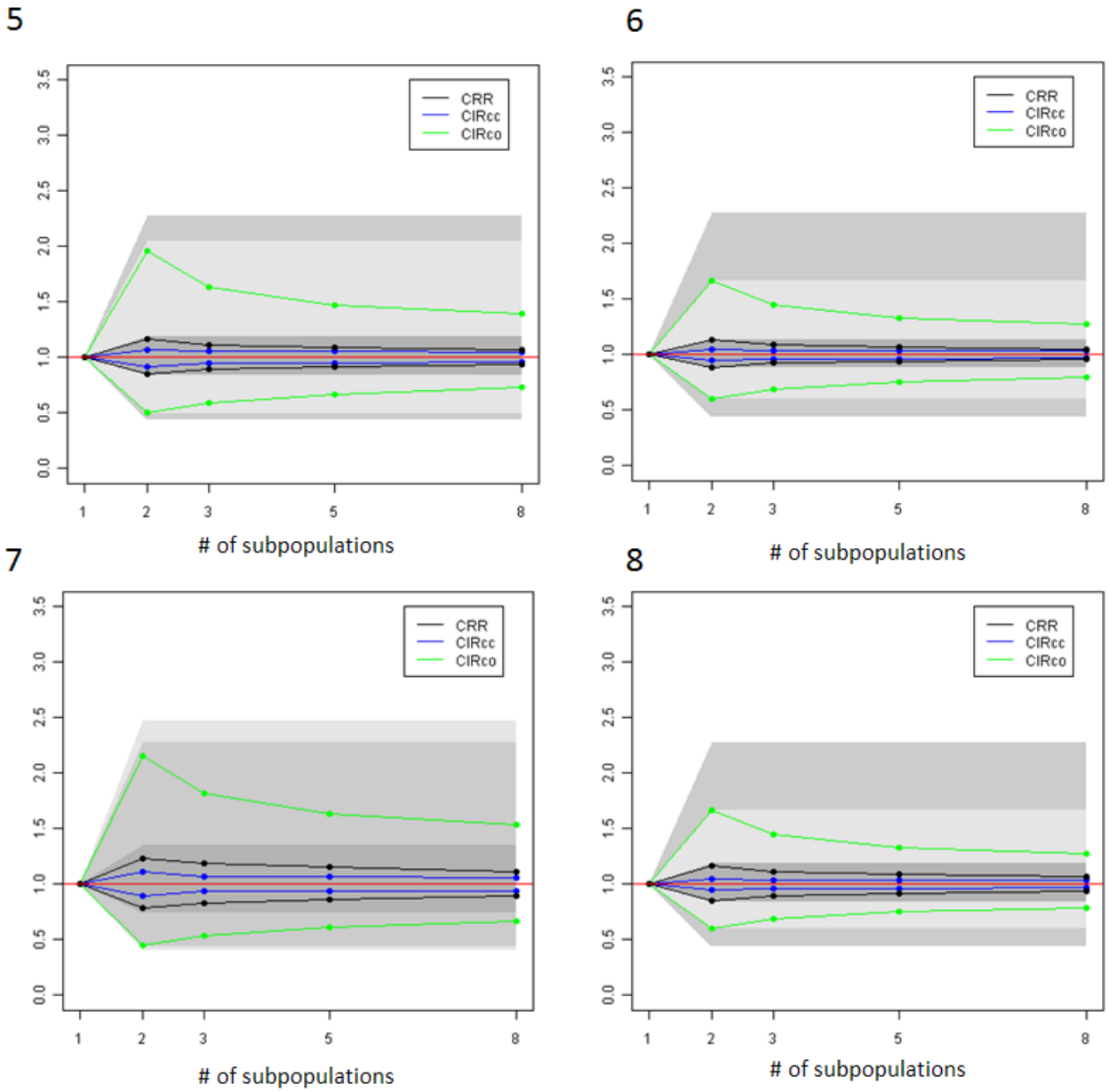
b_j, risk ratio, p_j, genotype frequency; q_j, exposure frequency; both G and E ranges are spaced to be equidistant on the logarithmic scale; * study cohort consists of 3 discrete, admixed populations; ** study cohort consists of 5 discrete, admixed populations; min, minimum CIR_{CC}; max, maximum CIR_{CC}; 25th, 50th, 75th, percentile of the CIR_{CC}

Figure 3.1 Scenarios 1-4, degree of population stratification for G×E interaction and genetic main effects



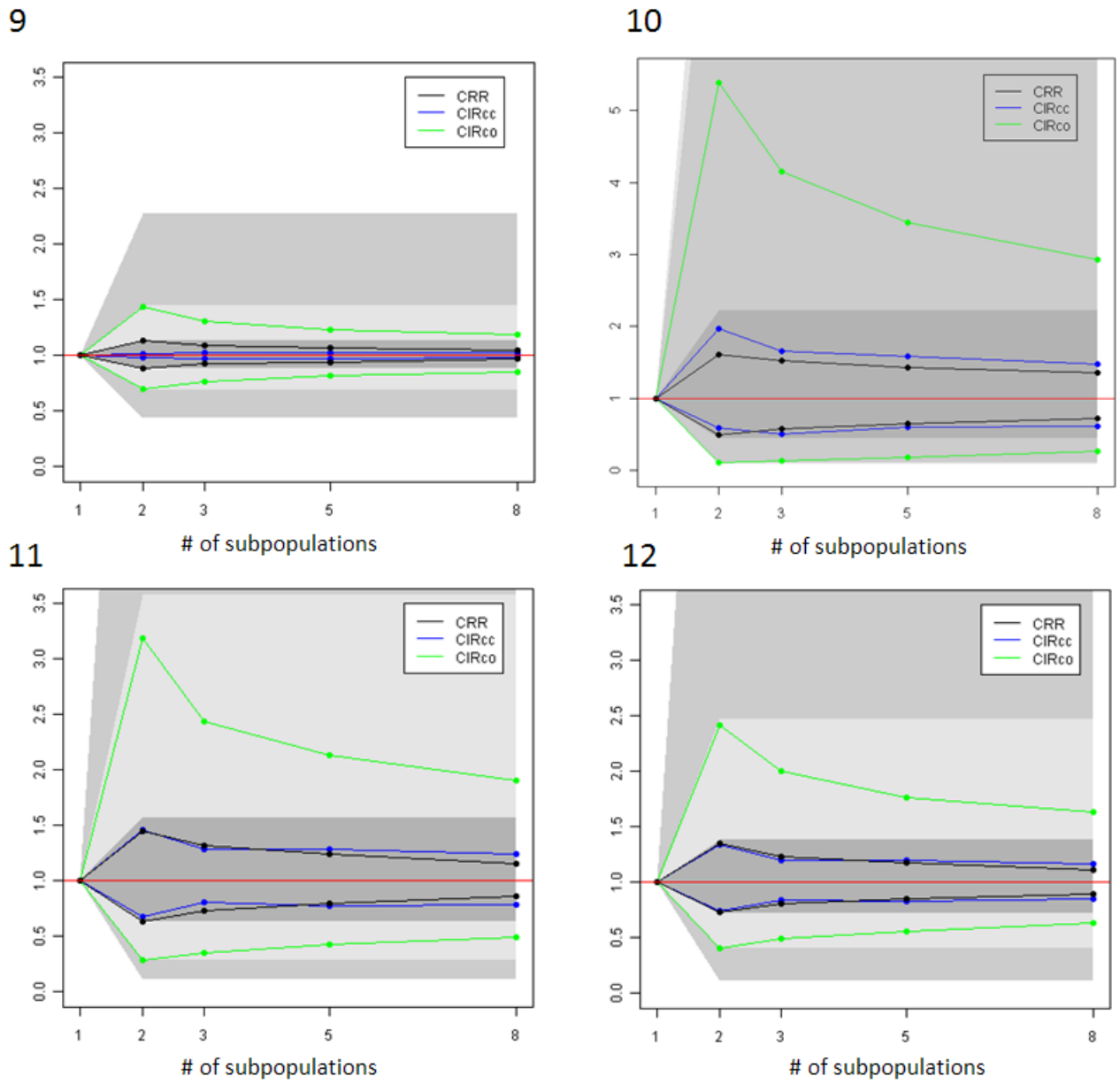
On the *x-axis* is the number of admixed subpopulations and on the y-axis are the minimum and maximum values of CRR, CIR_{CC}, and CIR_{CO} over 1000 replicates. In shades of grey are theoretical bounds of CRR, CIR_{CC}, and CIR_{CO} are depicted. Light grey corresponds to CIR_{CO} theoretical bounds for the scenario, medium grey CIR_{CC}, and dark grey CRR. The number at the corner denote the scenario, the order is the same as in Tables 3.2-3.4

Figure 3.2 Scenarios 5-8, degree of population stratification for G×E interaction and genetic main effects



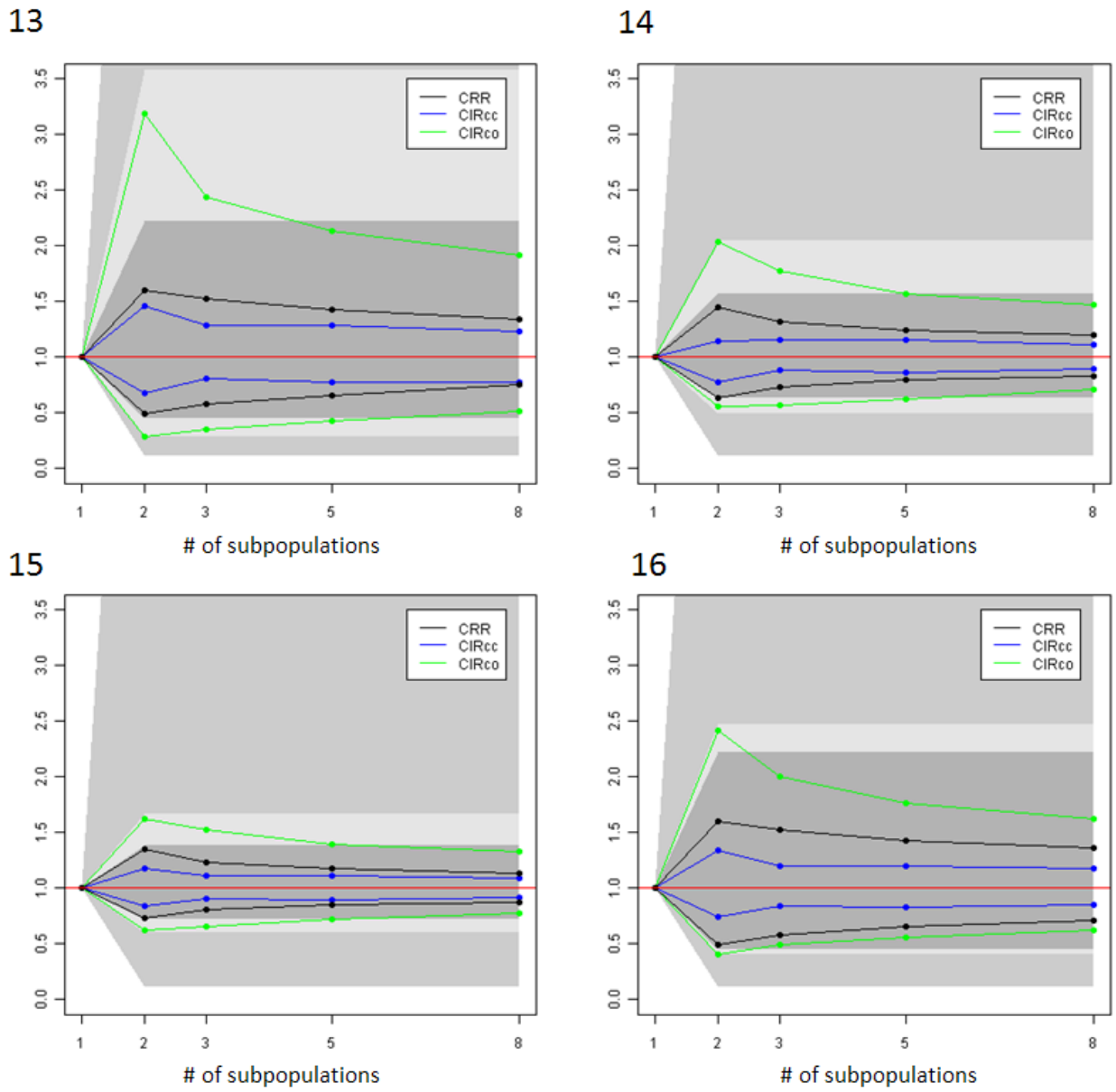
On the *x-axis* is the number of admixed subpopulations and on the y-axis are the minimum and maximum values of CRR, CIR_{CC}, and CIR_{CO} over 1000 replicates. In shades of grey are theoretical bounds of CRR, CIR_{CC}, and CIR_{CO} are depicted. Light grey corresponds to CIR_{CO} theoretical bounds for the scenario, medium grey CIR_{CC}, and dark grey CRR. The number at the corner denote the scenario, the order is the same as in Tables 3.2-3.4

Figure 3.3 Scenarios 9-12, degree of population stratification for G×E interaction and genetic main effects



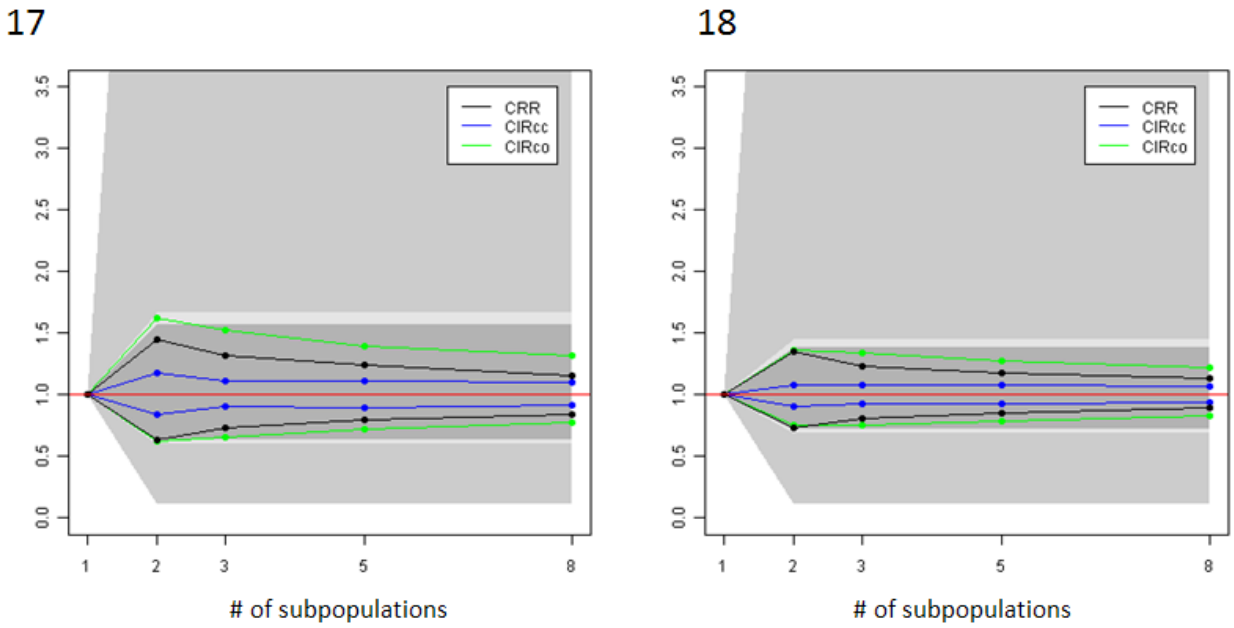
On the *x-axis* is the number of admixed subpopulations and on the y-axis are the minimum and maximum values of CRR, CIR_{CC}, and CIR_{CO} over 1000 replicates. In shades of grey are theoretical bounds of CRR, CIR_{CC}, and CIR_{CO} are depicted. Light grey corresponds to CIR_{CO} theoretical bounds for the scenario, medium grey CIR_{CC}, and dark grey CRR. The number at the corner denote the scenario, the order is the same as in Tables 3.2-3.4

Figure 3.4 Scenarios 13-16, degree of population stratification for G×E interaction and genetic main effects



On the *x-axis* is the number of admixed subpopulations and on the y-axis are the minimum and maximum values of CRR, CIR_{CC}, and CIR_{CO} over 1000 replicates. In shades of grey are theoretical bounds of CRR, CIR_{CC}, and CIR_{CO} are depicted. Light grey corresponds to CIR_{CO} theoretical bounds for the scenario, medium grey CIR_{CC}, and dark grey CRR. The number at the corner denote the scenario, the order is the same as in. Tables 3.2-3.4

Figure 3.5 Scenarios 17-18, degree of population stratification for G×E interaction and genetic main effects



On the *x-axis* is the number of admixed subpopulations and on the y-axis are the minimum and maximum values of CRR, CIR_{CC}, and CIR_{CO} over 1000 replicates. In shades of grey are theoretical bounds of CRR, CIR_{CC}, and CIR_{CO} are depicted. Light grey corresponds to CIR_{CO} theoretical bounds for the scenario, medium grey CIR_{CC}, and dark grey CRR. The number at the corner denote the scenario, the order is the same as in Tables 3.2-3.4

3.2. Degree of the Population Stratification Bias for G×E Interaction Methods

In the previous section we have seen that the biggest bias occurs when there are only two subgroups admixed. Therefore we decided to investigate the magnitude of the bias only for j = 2 strata in the population comparing four common methods for G×E interaction in GWAS. We included into our study CC, CO, MUK-EB and EHB-GE_{CHI} (see Chapters 1 and 2).

Since an analytical solution to evaluate the degree of population stratification bias for other G×E interaction methods than case-control or case-only design is not readily available we undertook the following approach. For each method estimates were obtained by fitting the "adjusted" logistic regression model, which accounts for population stratification in the sample

and an "unadjusted" model, which omits information about individual's membership in two subgroups. These two models are outlined in following Section 3.2.1.

Furthermore, PCA, integrated in the EIGENSOFT software (Patterson, Price et al. 2006), was performed on each simulated data set to estimate the principal components. After simulation of an admixed sample, we included the first two principal components in the logistic regression models as covariates. As explained in Chapter 2 the first two principal components explain the most variation in the sample. Since we have only two admixed subpopulation it is enough to use only first two principal components. The G×E interaction effect estimates were recalculated for each method after principal components adjustment and the bias due to population stratification was thus re-evaluated.

3.2.1. Methods

Let G=1 (G=0) denote carriers (non-carriers) of the susceptibility genotype and E=1 (E=0) for exposed (non-exposed) individuals. We let D be a binary phenotype, such that D=1, cases and D=0, controls. We assumed that the study sample consists of two (j=1, 2) strata represented by S_1 and S_2, and S_j is an indicator variable, such that $S_j=1$ if the individual is in the j^{th} subgroup and zero otherwise. We did not consider any issues concerning variance or precision of estimates, assuming that for large samples $E(\widehat{\beta}_l) = \beta_l$, l = CC, CO, MUK-EB, EHB-GE_{CHI}.

For the case-control study, an association between G×E interaction and the outcome D can be modeled in the following form using logistic regression

"Adjusted" model for *CC*: $logit(P(D= 1| G, E)) = \alpha_{1\_CC}+\alpha_{2\_CC}S_2+\beta_G G+\beta_E E+\beta_{CC} G\times E,$

where regression coefficient β_G is a measure of genetic main effect, β_E is a measure of the environmental main effect, β_{CC} is an estimated $G\times E$ interaction effect. Without loss of

generality let α_{1_CC} specify the log odds of the disease (i.e. *logit* function of the baseline disease risk) in the "low"-risk ethnicity S_1 and $0 < \alpha_2$, where α_2 specifies the log odds ratio of the disease risk comparing ethnicity S_2 versus S_1. Therefore to evaluate the observed bias we can omit the term from the model that is responsible to reflect the ethnic status of the individual and define "unadjusted" model for the CC study as follows

"Unadjusted" model for *CC*: $logit(P(D = 1| G, E)) = \alpha_{CC}^{*}+\beta^{*}_{G}G+\beta^{*}_{E}E+\beta^{*}_{CC}G\times E,$

where * denotes the regression in the unadjusted model.

Observed population stratification bias of the parameter estimate is defined as the difference between the corresponding parameters for the "Unadjusted" model and the "Adjusted" model. Therefore population stratification bias of the G×E interaction for the case-control design is equal to

$$bias_{CC} = \beta^{*}_{CC} - \beta_{CC}.$$

In a similar manner for case-only study, the "Adjusted" and "Unadjusted" models are given by

"Adjusted" model for *CO*: $logit(P(E = 1| G, D = 1)) = \alpha_{1\_cases}+\alpha_{2\_cases}S_2+\beta_{cases}G,$

"Unadjusted" model for *CO*: $logit(P(E = 1| G, D = 1)) = \alpha_{cases}^{*}+\beta^{*}_{cases}G.$

Therefore, bias for the case-only design is equal to

$$bias_{CO}=\beta^{*}_{cases} -\beta_{cases}.$$

Models for the subgroup of controls are only needed for other G×E methods.

The "Adjusted" and "Unadjusted" models for control are given by

"Adjusted" model for controls: $logit(P(E=1|G,D=0))=\alpha_{1\_controls}+\alpha_{2\_controls}S_2+\beta_{controls}G,$

"Unadjusted" model for controls: $logit(P(E = 1| G, D = 0)) = \alpha_{controls}^{*}+\beta^{*}_{controls}G.$

Mukherjee's Empirical Bayes type shrinkage estimator of the G×E interaction is given by

"Adjusted" and "Unadjusted" G×E parameter estimates of Mukherjee's Empirical Bayes approach are given by

$$\hat{\beta}_{MUK-EB} = \frac{\widehat{\beta}^2_{controls}}{\widehat{\beta}^2_{controls}+\widehat{\sigma}_{CC}}\hat{\beta}_{cases} + \frac{\widehat{\beta}^2_{controls}}{\widehat{\beta}^2_{controls}+\widehat{\sigma}_{CC}}\hat{\beta}_{CC}$$

$$\hat{\beta}^*_{MUK-EB} = \frac{\widehat{\beta}^{2*}_{controls}}{\widehat{\beta}^{2*}_{controls}+\widehat{\sigma}^*_{CC}}\hat{\beta}^*_{cases} + \frac{\widehat{\beta}^{2*}_{controls}}{\widehat{\beta}^{2*}_{controls}+\widehat{\sigma}^*_{CC}}\hat{\beta}^*_{CC},$$

where $\hat{\sigma}_{CC,}$ $\hat{\sigma}^*_{CC}$ are variance estimates from "Adjusted" and "Unadjusted" case-control model respectively. For MUK-EB the bias is defined as

$$bias_{MUK\text{-}EB} = \beta^*_{MUK\text{-}EB} - \beta_{MUK\text{-}EB}.$$

The EHB-GE_{CHI} G×E effect estimates are calculated using the following equation

$$\beta_{EHB\text{-}GECHI} = \beta_{cases} - posterior(\beta_{controls}),$$

therefore bias is given by

$$bias_{EHB\text{-}GE_{CHI}} = \beta^*_{EHB\text{-}GECHI} - \beta_{EHB\text{-}GECHI}.$$

Theoretical background for EHB-GE_{CHI} approach is given (Sohns 2012, Sohns, Viktorova et al. 2013) and summarized in Chapter 4 of this dissertation.

3.2.2. Simulation Study Set-up

We simulated case-control samples consisting of 1000 cases and 1000 controls sampled from an admixed population with different proportions of each of two subpopulations. Three different admixture sampling ratios were implemented 0.2, 0.4 and 0.5 reflecting an extreme (0.2-0.8), moderate (0.4-0.6) or balanced (0.5-0.5) level of admixture as follows. Admixture

sampling ratio 0.2, for example, means that 20% cases and 80% controls were sampled from "low"-risk ethnicity and 80% cases, 20% controls from "high"-risk subgroup. The usage of the two others ratios 0.4 or 0.5 is the same. We replicated each scenario 1000 times.

In each simulated sample we created binary genotype data at 5000 independent, random SNPs. We simulated three different types of SNPs. A single SNP with G×E interaction effect, and 1000 SNPs, with substantially varying frequency across two admixed subpopulations, denoted as differentiated SNPs. We called the rest of the SNPs dummy SNPs as they had no effect on the disease. A single binary environmental factor *E* was simulated, having no main effect on the disease. In each stratum susceptibility genotype frequencies at a marker locus for dummy SNPs were generated independently from a *beta* distribution following the Balding-Nichols model as in (Devlin and Roeder 1999). In this Balding-Nichols model two parameters are employed, $p(1-F_{st})/F_{st}$ and $(1-p)(1-F_{st})/F_{st}$, where F_{st} =0.01 and *p* is the ancestry population allele frequency from the *uniform* [0.1, 0.9] distribution. F_{st} is Wright's fixation index, a measure of genetic divergence among subgroups (Holsinger and Weir 2009), F_{st} =0.01 is a typical value for European populations. For the interacting SNP the frequency in the "low"-risk subpopulation was fixed at 0.1 and we vary this value in the risk subpopulation from 0.4 to 0.8 respectively. For the differentiated SNPs with no association to the disease, we assumed a large variation in frequencies by setting F_{st} values to be equal to 0.06. For the simulation of the interacting SNP we assumed a multiplicative trait model and fixed the relative risk at a value of 2 for the casual genotype. Exposure frequencies and baseline disease risk were fixed for all scenarios. In the "low-risk" ethnicity we set the prevalence of the environment to be equal to 0.1 and the background disease risk to be 0.02. In the high risk ethnicity corresponding values were set to 0.3, 0.5 and 0.1 or 0.05 respectively. For our study we considered a cohort with two underlying discrete subpopulations that could be present in cases and controls in different proportions as described above. Note that it is expected that there will be no

confounding for the 0.5 admixture sampling ratio. **Table 3.5** summarizes simulated scenarios. We replicated each scenario 1000 times and obtained the distribution of the population stratification bias and mean squared error defined as $MSE = bias^2 + variance$ and their average values for G×E methods before and after principal components adjustment.

Table 3.5 Summary of the simulated scenarios

Scenario	Parameters						
	ratio	**p_{01}**	**p_{02}**	**p_{e1}**	**p_{e2}**	**p_{g1}**	**p_{g2}**
1	0.5	0.02	0.1	0.1	0.3	0.1	0.4
2	0.5	0.02	0.1	0.1	0.3	0.1	0.8
3	0.4	0.02	0.1	0.1	0.3	0.1	0.4
4	0.4	0.02	0.1	0.1	0.3	0.1	0.8
5	0.4	0.02	0.05	0.1	0.5	0.1	0.4
6	0.2	0.02	0.1	0.1	0.3	0.1	0.8
7	0.2	0.02	0.1	0.1	0.3	0.1	0.4
8	0.2	0.02	0.05	0.1	0.5	0.1	0.8

ratio, proportion of cases sampling from "low-risk" ethnicity; p_{01}, baseline disease risk in "low-risk" ethnicity; p_{02}, baseline disease risk in "at-risk" ethnicity; p_{e1}, prevalence of environmental exposure in "low-risk" ethnicity; pe_2, prevalence of environmental exposure in "at-risk" ethnicity; p_{g1}, susceptible genotype frequency in "low-risk" ethnicity; p_{g2}, susceptible genotype frequency in "high-risk" ethnicity;

3.2.3. Simulation Study Results

Table 3.6 and **Table 3.7** summarize results of the simulation study. The classic case-control estimator of G×E interaction, the recently introduced EHB-GE_{CHI} estimator and MUK-EB demonstrate smaller bias in all of the scenarios compared to the case-only estimator. One should note that for the matched case-control design (0.5) including two admixed subpopulations the case-control estimator tends to give better results in terms of smaller population stratification bias than the other considered estimators. For the moderate admixture or matched case-control design population stratification bias of the case-control estimator is negligible, however this is not the case for the case-only estimator. The explanation can be found by considering the main idea behind those methods. The original case-control estimator compares odds in cases and controls. In contrast to this the case-only estimator based only on the odds in cases. Therefore the case-control estimator tends to overcome the lack of homogeneity in both groups by matching cases to controls. The CC estimator does not take

into account other SNPs and analyzing each SNP separately. The EHB-GE_{CHI} method takes into consideration all other SNPs and therefore may suffer from difficulties in a single SNP estimate when there is some hidden substructure in other SNPs, like G-E correlations or presence of differentiating SNPs for which frequencies vary significantly across subpopulations. EIGENSTRAT is population stratification correction method, which is based on the principal components analysis and eigenvalues analysis. These are integrated in the software called EIGENSOFT (Patterson, Price et al. 2006). PCA was performed to derive principal components to account for population stratification. We included the first two principal components as covariates in the "unadjusted" model to account for PS. Population stratification bias was reevaluated for each G×E interaction approach after principal components adjustment and appeared to be practically zero for all methods. As only two sub-populations were admixed it is sufficient to only include the first two principal components as covariate, accounting for the variation in the sample due to the stratification.

Table 3.6 Bias of G×E interaction estimators, calculated as observed difference of the estimates in two logistic regression models for G×E interaction methods

Scenario	bias CC	bias CO	bias MUK-EB	bias EHB-GE_{CHI}
1	0.124	0.418	0.100	0.183
2	0.351	0.234	0.260	0.249
3	0.361	0.636	0.181	0.192
4	0.101	0.905	0.247	0.267
5	0.009	0.489	0.083	0.176
6	0.015	0.969	0.300	0.332
7	0.179	0.659	0.157	0.178
8	0.518	0.383	0.418	0.301

Table 3.7 Mean Squared Error of G×E interaction estimators

Scenario	MSE CC	MSE CO	MSE MUK-EB	MSE EHB-GE_{CHI}
1	0.071	0.201	0.073	0.087
2	0.208	0.076	0.148	0.134
3	0.210	0.441	0.117	0.106
4	0.066	0.861	0.148	0.159
5	0.053	0.269	0.067	0.089
6	0.055	0.986	0.184	0.208
7	0.078	0.460	0.082	0.082
8	0.337	0.166	0.246	0.159

Scenarios employed in analysis (Table 3.5 for specification); CC, Case-Control estimator; CO, Case-Only estimator; MUR, MUK-EB, Mukherjee's Empirical Bayes type shrinkage estimator; EHB-GE_{CHI}, empirical hierarchical Bayes approach to G×E interaction;

Chapter 4

4. Extensions for the Empirical Hierarchical Bayes Approach to G×E Interaction EHB-GE_{CHI}

The general concept of the empirical hierarchical Bayes (EHB) modeling approach is described in the current Chapter, since the EHB analysis is the focus of this dissertation. Recently, Sohns proposed an empirical hierarchical Bayes approach to G×E interaction designated as EHB-GE_{CHI} (Sohns 2012, Sohns, Viktorova et al. 2013). The EHB-GE_{CHI} was developed as a compromise between the often underpowered case-control test and the case-only test, which has highly inflated type I error in the presence of G-E correlation. The proposed method is based on a two-level hierarchical model to estimate G-E correlation and employs the *chi* distribution on the first level and a mixture distribution with point mass at zero on the second level. The EHB-GE_{CHI} method is based on the Lewinger et al. (Lewinger, Conti et al. 2007) hierarchical Bayes prioritization approach, which was originally proposed for the genetic main effect. Sohns expanded this approach to studies of G×E interactions, as well as to using the available pathway information. The method first obtains estimated posterior G-E correlation effects, which are calculated for each marker by borrowing information across all SNPs over the sample of controls. These posterior effects are subtracted from the corresponding case-only G×E interaction estimates. A detailed description of the EHB-GE_{CHI} method is available in the dissertation of M. Sohns (Sohns 2012) as well as in our joint paper (Sohns, Viktorova et al. 2013). Here we only present the summarized derivation of the EHB-GE_{CHI} approach. The thesis by Sohns also includes the description and results of the extensive simulation study comparing EHB-GE_{CHI} with the methods mentioned above for G×E interaction in terms of rank power. Rank power is defined as the percentage of simulated replicates for which the true G×E interacting SNP is within the top ranking positions, according to the absolute value of the corresponding rank statistics. The EHB-GE_{CHI} method is characterized by a greater rank power

than the introduced rival methods (CC, CO, TWO, MUR, MUK-EB), while accounting for population-based G-E correlation. The original EHB-GE_{CHI} was introduced for binary trait, exposure, and genotype without covariates. We extend the EHB-GE_{CHI} approach in several ways.

We demonstrate how the method can be applied to handle multilevel or continuous exposure and genotype variables in contrast to the original binary setting. This is an important extension because in many cases exposure information is collected as a continuous variable. Such recoding may lead to information loss. In addition, use of the continuous probabilities of genotypes, obtained through available imputation techniques, is taken into consideration more often.

We considered the performance of EHB-GE_{CHI} under the assumption of the additive risk model in contrast to the dominant or recessive model discussed previously. It is well known that the additive risk model is preferred for most genetic scenarios, when the etiology of the disease is not known, and most probably is not recessive. It is therefore important that the approach be able to handle various risk models.

We show that the EHB-GE_{CHI} approach can be applied adjusting for important covariates separately in cases and in controls. The separate adjustment is required by the construction of the EHB-GE_{CHI} approach. We prove that separate adjustment is allowed when independence of the covariate distribution from the G×E interaction odds ratio is a reasonable assumption. We proposed using log-linear models in place of logistic regression models, when such an assumption is not valid. Originally, EHB-GE_{CHI} did not consider covariate adjustment. Certainly, adjustment for sex, age, ethnical background, and so on is usually performed during analyses. Therefore, it should be clear as to how the method can be applied under the

adjustment for important covariates. This chapter focuses on presenting the proposed extensions to the EHB-GE_{CHI} approach and the derivation of solutions.

4.1. Empirical Hierarchical Bayesian Models

Statistical science may be viewed as two main competing colleges of thought: the frequentist or classic approach to statistical inference and the Bayesian approach. In the following sections, we shall introduce the basics of empirical hierarchical Bayes data analysis. The major units of a Bayesian analysis are the likelihood function, which represents information on the parameters of the data, and the prior distribution, which quantifies what is known about the parameters before observing the data. To form the posterior distribution, the prior distribution and the likelihood are combined. The posterior distribution reflects the total knowledge on the parameters after observing the data. Simple summaries, such as mean or median, of the posterior distribution are used to express quantities of interest and eventually to draw conclusions. Most of the information given in the following sections was adopted from (Morris 1983, Robert 1994, Lee 1997, Sohns 2012).

4.1.1. The Bayes Model

The initial step to perform Bayes inferences is to specify a probability model for the data. Assume that we want to specify a sampling model of n data points $X=(X_1,\ldots,X_n)$ depending on the vector of unknown parameters $\theta=(\theta_1,\ldots,\theta_n)$ and that data points are independent, conditional on θ. This can be expressed in a functional term using the probability density function $f(X|\theta)$, where

$$f(X|\theta)=\prod_{i=1}^{n} f(X_i|\theta)$$

represents the probability of observing the data point *X* conditional on the values of parameters *θ*. In frequentist statistics, *f(X|θ)* is termed the likelihood function and is considered as a function of θ for fixed data points *X*. Parameter estimates in the frequentist inferences are derived by maximizing the likelihood and are termed maximum likelihood estimates (Robert 1994, Dehling and Haupt 2004). Bayesians also are interested in the estimation of *θ*. However, they prefer to obtain the parameter estimates that are most likely given the fixed data. In other words, in the Bayes inference framework we are interested in the conditional probability of θ given *X*, $\pi(\theta|X)$. To derive $\pi(\theta|X)$ we use Bayes' Theorem (Bayes 1991, Robert 1994).

$$\pi(\theta|X) = \frac{f(X|\theta)\pi(\theta)}{h(X)} = \frac{f(X|\theta)\pi(\theta)}{\int f(X|\theta)\pi(\theta)d\theta}$$

where $\pi(\theta)$ is the prior distribution function of θ, $f(X|\theta)$ is the likelihood function, $h(X)$ is the marginal distribution function of *X* and $\pi(\theta|X)$ is the posterior distribution function of the parameters. In many situations, it is computationally challenging to obtain $h(X)$ and therefore the posterior distribution often has no closed form. Thus $\pi(\theta|X) \propto f(X|\theta)\pi(\theta)$ is widely used (Robert 1994, Gelman, Carlin et al. 1995).

4.1.2. Empirical Hierarchical Bayes Models

In the Bayes analysis framework, defining the prior distribution always involves large uncertainty and sometimes subjectivism and is therefore often subject to criticism (Lee 1997). However, it is possible to model this uncertainty in a Bayesian manner by the decomposing the prior information into separate distributional levels. This is what is referred to as hierarchical Bayes (HB) modeling (Robert 1994). According to the HB approach, the prior $\pi(\theta)$ is separated into conditional distributions $\int \pi_1(\theta|\eta_1)\pi_2(\eta_1|\eta_2) \dots \pi_J(\eta_{J-1}|\eta_J)$ and a marginal distribution $\pi_{J+1}(\eta_J)$, so that

$$\pi(\theta) = \int \pi_1(\theta|\eta_1)\pi_2(\eta_1|\eta_2) \dots \pi_J(\eta_{J-1}|\eta_J)\pi_{J+1}(\eta_J)d\eta_1 d\eta_2 \dots \, d\eta_J.$$

The hyperparameter η_j is termed the level *j* hyperparameter to distinguish it from the parameter of interest θ. Therefore, given the data as in the previous section, we have the following hierarchical model.

$\theta_i \sim \pi(\theta_i|\eta)$, with *i=1 ...n, i.d.* independently distributed

$\eta \sim \pi(\eta)$

For such a two-level hierarchical model, the first stage represents the model hyperparameter θ relationship with level hyperparameter η. The second stage reflects our prior belief about η (Lee 1997).

The hierarchical model can be combined with empirical Bayes models, resulting in empirical hierarchical Bayes models (EHB). The empirical Bayes models form a special class of Bayes models, which differ from the fully Bayesian approach in the construction of the prior distribution. The prior distribution in empirical Bayes methods is usually given a frequency interpretation, in contrast to that of the true Bayes methods (Lee 1997). In the empirical Bayes context, hyperparameter estimation is performed in a similar approach to that of the frequentist, i.e. through the maximization of the marginal distribution function, $h(X|\eta)$ with respect to η. In this way, marginal maximum likelihood estimates (MMLE) of the parameters are obtained (Berger 1985, Heron, O'Dushlaine et al. 2011).

Consider a general EHB model:

Level 1 $\quad X_i|\theta_i \sim f(X_i|\theta_i)$, with *i=1 ...n, i.d.*

Level 2 $\quad \theta_i|\eta \sim \pi(\theta_i|\eta)$, *iid* independent identically distributed

Level 3 $\quad \eta \sim \pi(\eta).$

Since the likelihood function of the data is

$$\prod_{i=1}^{n} f(X_i|\theta_i)\pi(\theta_i|\eta)$$

and the marginal distribution function of the data is given by

$$h(X|\eta) = \int f(X_i|\theta_i)\pi(\theta_i|\eta)d\theta_i.$$

Thus, the marginal likelihood function of the data is obtained by marginalizing the likelihood over the parameters of interest

$$L = \prod_{i=1}^{n} h(X_i|\eta) = \prod_{i=1}^{n} \int f(X_i|\theta_i)\pi(\theta_i|\eta)d\theta_i.$$

When the marginal distribution has a relatively simple form, it is possible to obtain an exact solution of the MMLE performing standard iterative maximum likelihood methods, such as, for example, the expectation-maximization algorithm (EM), (Lee 1997).

EB analysis assumes that the prior distribution is known by using MLE estimates, $\hat{\eta}$ and modeling $\theta_i|\hat{\eta} \sim \pi(\theta_i|\hat{\eta}), iid.$ Based on this assumption the posterior of θ_i can be calculated by,

$$\pi(\theta_i|X_i, \hat{\eta}) = \frac{f(X_i|\theta_i)\pi(\theta_i|\hat{\eta})}{h(X_i|\hat{\eta})}.$$

Therefore the posterior depends on all data, summarized in $\hat{\eta}$.

4.2. The Empirical Hierarchical Bayes Approach to G×E Interaction (EHB-GE$_{CHI}$)

Let *D=1* denote that an individual has the disease (case), *D=0* otherwise (control). Let G=*1* denote carriers of the minor allele, *G=0* non-carriers, i.e. a dominant model for SNP. A binary environmental factor is assumed, so that *E=1* indicates exposed subjects and *E=0* otherwise.

Estimates of G-E correlation within cases and controls can be obtained from the following logistic regression models

$$logit\ (P(E = 1 \mid G, D = 1)) = \alpha_m^{cases} + \beta_m^{cases} G \quad \textbf{(4.1)}$$

$$logit\ (P(E = 1 \mid G, D = 0)) = \alpha_m^{controls} + \beta_m^{controls} G \quad \textbf{(4.2)}$$

Under the assumption of a rare disease and population-based G-E independence, $\beta_{controls}=0$. Then (4.1) corresponds to the valid model for the case-only test for G×E interaction (as illustrated in Chapter 2) and (4.2) provides estimates of the G-E correlation effects within controls. However, when such assumption is not true, $\beta_m^{controls}$coefficients should be properly estimated and consequently subtracted from β_m^{cases} coefficients in order to obtain an unbiased estimate of the G×E interaction effect. For example, with the classic case-control approach, one can estimate the G-E correlation within controls for each SNP using equation 4.2. These estimates are then subtracted from the coefficients within the cases (4.1).

In the context of GWAS, let M be the total number of genetic markers or SNPs m, $m=1 \ldots M$, and β_m^{cases}, $\beta_m^{controls}$ be the corresponding regression coefficients for G-E correlation among cases or controls, respectively (obtained by equations (4.1) and (4.2) for each SNP m) with corresponding standard deviations σ_m^{cases}, $\sigma_m^{controls}$. The remainder of Section 4.2 is based on (Sohns 2012, Sohns, Viktorova et al. 2013). The test statistics are $T_m^{cases} = \hat{\beta}_m^{cases} / \hat{\sigma}_m^{cases}$ and $T_m^{controls} = \hat{\beta}_m^{controls} / \hat{\sigma}_m^{controls}$, both normally distributed. A hierarchical Bayes framework is applied to model the $\hat{\beta}_m^{controls}$estimated effect and to calculate $\hat{\beta}_m^{cases} - sgn_m \hat{\lambda}_m$, with $\hat{\lambda}_m$ being *a posteriori* estimators of $|\beta_m^{controls}|$, and sgn_m denoting the sign of $\hat{\beta}_m^{controls}$. The corresponding hierarchical model is given by

$$\text{Level 1} \qquad \left|\hat{\beta}_m^{controls}\right| \mid \lambda_m \ \sim \ \hat{\sigma}_m^{controls} \chi_1(\lambda_m) \quad \textbf{(4.3)}$$

$$\text{Level 2} \qquad \lambda_m \mid \theta, \sigma, p \ \sim \ p\sigma\chi_1(\theta) + (1-p)\delta(0) \quad \textbf{(4.4)}$$

where λ_m are noncentrality parameters of the χ distribution with one degree of freedom ($\chi_1(\lambda_m)$) and p the estimated proportion of SNPs with G-E correlation. Assuming $\lambda_m > 0$, λ_m

is assumed to have a χ_1 distribution with noncentrality parameter θ as a measure of correlation and a scaling parameter $\sigma > 0$. Given $\lambda_m=0$, $\delta(0)$ denotes a point mass at zero.

Next, the probability density function, the prior probability, marginal distribution, and posterior expected values can be derived. What results is the following form for the posterior expectation of the non-centrality parameter

$$\hat{\lambda}_m = E[\lambda_m| \, |\hat{\beta}_m^{controls}| \, \hat{\theta}, \hat{\sigma}, \hat{p}],$$

based on the MLE estimates of the hyperparameters $\hat{\Theta} = (\hat{\theta}, \hat{\sigma}, \hat{p})$. The EHB-$GE_{CHI}$ rank statistic is given in (4.5).

$$T_m^{EHB-GE} = \frac{\hat{\beta}_m^{cases} - \text{sgn}_m posterior(|\beta_m^{controls}|)}{\sqrt{(\hat{\sigma}_m^{cases})^2 + Var\left(posterior(|\beta_m^{controls}|)\right)}} = \frac{\hat{\beta}_m^{cases} - \text{sgn}_m \hat{\lambda}_m}{\sqrt{(\hat{\sigma}_m^{cases})^2 + Var[\lambda_m| \, |\hat{\beta}_m^{controls}|]}} \quad \textbf{(4.5)}$$

Notice that the case and the control part of the differences $\left(\hat{\beta}_m^{cases} - sgn_m \hat{\lambda}_m\right)$ are independent of each other. It is known that $Var(\hat{\beta}_m^{cases}) = (\sigma_m^{cases})^2$ is estimated by $(\hat{\sigma}_m^{cases})^2$. Therefore one only needs to obtain the variance for the control part, which is $Var\left[\lambda_m|\hat{\beta}_m^{controls}\right]$. To estimate this variance Sohns proposed using an approximation by (Kass and Steffey 1989). To derive the rank statistic T_m^{EHB-GE} the following steps are undertaken:

a. Obtain the marginal likelihood of the hierarchical model $L = \prod_m h\left(\left|\hat{\beta}_m^{controls}\right| \left|\theta, \sigma, p\right.\right)$.

The density functions for the hierarchical model and the hyperparameters $\Theta = (\theta, \sigma, p)$ are

$$f(\left|\hat{\beta}_m^{controls}\right| \, | \, \lambda_m) = \left(\varphi\left(\frac{|\hat{\beta}_m^{controls}| - \lambda_m}{\hat{\sigma}_m^{controls}}\right) + \varphi\left(\frac{|\hat{\beta}_m^{controls}| + \lambda_m}{\hat{\sigma}_m^{controls}}\right)\right) / \hat{\sigma}_m^{controls},$$

$$g(\lambda_m|\theta, \sigma, p) = p\left(\varphi\left(\frac{\lambda_m - \theta}{\sigma}\right) + \varphi\left(\frac{\lambda_m + \theta}{\sigma}\right)\right)/\sigma + (1-p)\delta(0)$$

where $\varphi(.)$ is the standard normal density. The marginal distribution is given by

$$h\left(\left|\hat{\beta}_m^{controls}\right| \left|\theta,\sigma,p\right.\right) = \int f\left(\left|\hat{\beta}_m^{controls}\right| \mid \lambda_m\right) g\left(\lambda_m \middle| \theta,\sigma,p\right) d\lambda_m$$
$$= p\frac{\varphi(D_{+m})+\varphi(D_{-m})}{\sqrt{\left(\hat{\sigma}_m^{controls}\right)^2+\sigma^2}} + (1-p)2\varphi\left(\frac{\left|\hat{\beta}_m^{controls}\right|}{\hat{\sigma}_m^{controls}}\right)/\hat{\sigma}_m^{controls},$$

where $D_{+m} = \frac{\left|\hat{\beta}_m^{controls}\right|+\theta}{\sqrt{\left(\hat{\sigma}_m^{controls}\right)^2+\sigma^2}}$, $D_{-m} = \frac{\left|\hat{\beta}_m^{controls}\right|-\theta}{\sqrt{\left(\hat{\sigma}_m^{controls}\right)^2+\sigma^2}}$

b. Obtain the MLE $\widehat{\Theta} = (\hat{\theta},\hat{\sigma},\hat{p})$ from the marginal log likelihood maximizing with respect to Θ. $(\hat{\theta},\hat{\sigma},\hat{p})$ are common hyperparameters estimates.

c. Obtain the posterior expectation of λ_m as $\hat{\lambda}_m = E[\lambda_m \mid \left|\hat{\beta}_m^{controls}\right| \hat{\theta},\hat{\sigma},\hat{p}]$ based on $(\hat{\theta},\hat{\sigma},\hat{p})$.

d. Obtain the inverse negative Hessian of the marginal log-likelihood evaluated at the MLE,

$$\tilde{\Sigma} = (-D^2\log(L)(\widehat{\Theta}))^{-1} = \begin{pmatrix} \widetilde{\tau_{\theta\theta}} & \widetilde{\tau_{\theta\sigma}} & \widetilde{\tau_{\theta p}} \\ \widetilde{\tau_{\sigma\theta}} & \widetilde{\tau_{\sigma\sigma}} & \widetilde{\tau_{\sigma p}} \\ \widetilde{\tau_{p\theta}} & \widetilde{\tau_{p\sigma}} & \widetilde{\tau_{pp}} \end{pmatrix}.$$

e. Obtain the Jacobian of the posterior expectation $\tilde{\delta}_{mk} = (\frac{\partial}{\partial\Theta_k})E\left[\lambda_m \middle| \left|\hat{\beta}_m^{controls}\right|,\Theta\right]|_{\Theta=\widehat{\Theta}}$.

f. Obtain the first order approximation to the posterior variance using an approximation by Kass and Steffey (1989).

$$Var\left[\lambda_m \middle| \left|\hat{\beta}_m^{controls}\right|\right] = E\left[Var\left[\lambda_m \middle| \left|\hat{\beta}_m^{controls}\right|,\Theta\right]\right] + Var\left[E\left[\lambda_m \middle| \left|\hat{\beta}_m^{controls}\right|,\Theta\right]\right]$$
$$\approx Var\left[\lambda_m \middle| \left|\hat{\beta}_m^{controls}\right|,\widehat{\Theta}\right] + \sum\nolimits_{j,i} \tilde{\tau}_{ji}\tilde{\delta}_{mj}\,\tilde{\delta}_{mi}$$
$$= E\left[\lambda_m^2 \middle| \left|\hat{\beta}_m^{controls}\right|,\widehat{\Theta}\right] - \left(E\left[\lambda_m \middle| \left|\hat{\beta}_m^{controls}\right|,\widehat{\Theta}\right]\right)^2 + \sum\nolimits_{j,i} \tilde{\tau}_{ji}\tilde{\delta}_{mj}\,\tilde{\delta}_{mi}$$

where

$$E\left[\lambda_m^2 \middle| \left|\hat{\beta}_m^{controls}\right|,\widehat{\Theta}\right] = E\left[\lambda_m^2 \middle| \lambda_m > 0, \left|\hat{\beta}_m^{controls}\right|,\widehat{\Theta}\right]P\left[\lambda_m > 0 \middle| \left|\hat{\beta}_m^{controls}\right|,\widehat{\Theta}\right]$$
$$+ E\left[\lambda_m^2 \middle| \lambda_m = 0, \left|\hat{\beta}_m^{controls}\right|,\widehat{\Theta}\right]P\left[\lambda_m = 0 \middle| \left|\hat{\beta}_m^{controls}\right|,\widehat{\Theta}\right]$$

where the first summand in the above expression can be calculated and

$$\mathrm{E}\left[\lambda_m^2 | \lambda_m = 0, \hat{\beta}_m^{controls}, \widehat{\Theta}\right] = 0$$

g. Insert the posterior expectation and the posterior variance into the final test statistic

$$T_m^{EHB-GE} = \frac{\hat{\beta}_m^{cases} - \mathrm{sgn}_m \hat{\lambda}_m}{\sqrt{(\hat{\sigma}_m^{cases})^2 + Var\left[\lambda_m | |\hat{\beta}_m^{controls}|\right]}}$$ (Sohns 2012, Sohns, Viktorova et al. 2013)

Detailed equations for the posterior variance derivation can be found in (Sohns 2012).

4.3. General Exposure Variable and Genotype Variable

The EHB-GE_{CHI} method requires estimation of G-E correlations separately within cases and controls. This can be achieved employing equations (4.1) and (4.2). Therefore, the question arises as to if EHB-GE_{CHI} can be extended to work with continuous or multi-level exposures and categorical genotypes. This is certainly the case. In the regression models (4.1) and (4.2), the exposure variable can be represented by multiple levels or by a continuous variable in a general linear models framework with a link function appropriate to the format of the *E* data. For example, in the case of normally distributed *E* the following model can be applied to the data, conditional on the disease.

$$E(E|G,D) = \alpha_{G\times E} + \beta_{G\times E} G. \qquad \textbf{(4.6)}$$

According to equation (4.6), the relationship between genotype and environment is modeled as a simple linear regression. This approach to data modeling (formula (4.6)) was evaluated by Clarke and Morris (Clarke and Morris 2010). If *E* and *G* are coded as categorical or are categorical by nature, then proportional, multinomial, or ordinal regression techniques can be performed to model the G-E relationship (Kraft, Yen et al. 2007).

As an alternative, we propose modeling *P(G|E, D)* instead of the original *P(E|G, D)* (note that we follow the approach of Piegorsch et al. to construct the necessary proof (Piegorsch and Casella 1996) of equality in the approaches). Assume that *G*, *E*, and *D* are all binary. Then we

can estimate G-E correlation effects separately in cases and controls. We treat the binary genotype variable as an outcome and estimate the main effects for the exposure categories.

To demonstrate the equality of the data modeling approaches, first consider the ratio

$$\frac{P(G = 0|E = 1, D = 0)P(G = 1|E = 0, D = 0)}{P(G = 1|E = 1, D = 0)P(G = 0|E = 0, D = 0)}$$

The odds ratio associated with $G\times E$ interaction on the multiplicative scale can be defined by $OR_{G\times E} = \frac{OR_{GE}}{OR_G OR_E}$ as discussed in Chapter 2, where OR_{GE} is the odds ratio relating risk at the G=1, E=1 combination to the G=0, E=0 'baseline' genotype-exposure combination, OR_G is the odds ratio relating risk at the G=1, E=0 to the G=0, E=0 'baseline' combination and OR_E is the odds ratio relating risk at the G=0, E=1 gene-exposure combination to the 'baseline' combination. Consider the following data model

$$logit\ (P(D = 1|\ G, E)) = \alpha_{CC} + \beta_G G + \beta_E E + \beta_{CC} G\times E \qquad \textbf{(4.7)}$$

From the logistic regression formula (4.7),

$$\beta_{CC} = log(OR_{G\times E}) = log(\frac{OR_{GE}}{OR_G OR_E}),$$

which can be written applying Bayes' rule twice as

$$OR_{G\times E} = \frac{OR_{GE}}{OR_G OR_E}$$

$$= \frac{P(G = 1|E = 1, D = 1)P(G = 0|E = 0, D = 1)P(G = 0|E = 1, D = 0)P(G = 1|E = 0, D = 0)}{P(G = 1|E = 1, D = 0)P(G = 0|E = 0, D = 0)P(G = 0|E = 1, D = 1)P(G = 1|E = 0, D = 1)}$$

$$= \frac{P(G = 1|E = 1, D = 1)P(G = 0|E = 0, D = 1)}{P(G = 0|E = 1, D = 1)P(G = 1|E = 0, D = 1)} \frac{P(G = 0|E = 1, D = 0)P(G = 1|E = 0, D = 0)}{P(G = 1|E = 1, D = 0)P(G = 0|E = 0, D = 0)}$$

Applying natural logarithm to both sides of the equation, we obtain the following result

$$\beta_{CC} = \log(\mathrm{OR}_{\mathrm{G\times E}}) = \log\left(\frac{\mathrm{OR}_{\mathrm{GE}}}{\mathrm{OR}_{\mathrm{G}}\mathrm{OR}_{\mathrm{E}}}\right)$$

$$= \log\left(\frac{\mathrm{P(G=1|E=1,D=1)P(G=0|E=0,D=1)}}{\mathrm{P(G=0|E=1,D=1)P(G=1|E=0,D=1)}}\right)$$

$$- \log\left(\frac{\mathrm{P(G=1|E=1,D=0)P(G=0|E=0,D=0)}}{\mathrm{P(G=0|E=1,D=0)P(G=1|E=0,D=0)}}\right)$$

$$= \log(OR_{cases}) - \log(OR_{controls}) = \beta_{cases} - \beta_{controls}$$

Therefore, it is possible to model *P(G|E,D)* instead of *P(E|G,D)*, the interpretation of the beta coefficients from the logistic regression models (4.1) and (4.2) is the same. These calculations are easily extended for a categorical or a continuous *E* variable. The proportional odds regression analysis techniques to model $P(G \leq k|E, D)$ or multinomial for a genetic response model to design $P(G=k|E,D)$, $k = 0, 1, 2$ can be performed in situations when the genotype variable is coded as 0, 1, and 2. We proved the equality of modeling the probability of the exposure, conditional on the genotype or otherwise, within the logistic regression framework. This provides us with a simple way to extend the EHB-GE_{CHI} approach to application with a continuous or categorical exposure variable.

4.4. Additive Risk Model

We discussed characteristics of case-control studies and possibilities for different types of genotype and exposure coding within the regression analysis framework. However, we have not discussed the possibility of assuming a particular mode of inheritance so far. Logistic regression has an advantage over the chi-squared test (which is not discussed in this dissertation), in that it is easier to test different genetic models (co-dominant, log-additive, dominant, recessive) or account for the covariates. The assumption of a specific mode of inheritance in many scenarios can bring power gain of the test. Three genetic models are

commonly used for the inheritance mode: the additive model refers to an "additive" risk on the log-scale, where each allele carries an equal risk; the dominant model, where the relative risk associated with one at-risk allele is the same as the risk for carriers of two allele copies; and the recessive model, where the only individuals at risk are those carrying two copies of the risk allele (Zeggini and Morris 2010). The co-dominant model formulation is the one, assuming no particular mode of inheritance, in which the relative risk of a disease associated with the heterozygous genotype and with the homozygous genotype do not relate to one another from a statistical point of view (Sasieni 1997). It is easy to implement these various models in a logistic regression framework. To do so, the genotypes at each locus, for example AA/AT/TT, where T is the risk allele, are coded as categorical for a genotypic association test, 0/1/2 for an additive risk test, and 0/1/1 or 0/0/1 for the dominant or recessive test.

In real data settings, it is clear that we do not know the true mode of inheritance for the disease and can therefore only assume a genetic risk model for the analyses. Previously, it was shown that the additive risk model has an advantage over the other models. It is a more powerful model to test for when the true inheritance mode is additive or close to additive. Furthermore, it has comparable power to the dominant model when the true mode of inheritance is dominant or similar to dominant (Lettre, Lange et al. 2007). However the disadvantage of the additive risk model implementation within the logistic regression analysis framework is that analytical estimates of the corresponding *beta* coefficients (*log*s of the odds ratios) are not available (Eiichiro 2004), but iterative estimates are available. Therefore we cannot prove analytically that the relationship $\beta_{CC} = \beta_{cases}$-$\beta_{controls}$ holds under the assumption of the additive risk model. It is clear that both estimators $\hat{\beta}_{CC}$ and $\hat{\beta}_{cases} - \hat{\beta}_{controls}$ are testing the same null hypothesis of no G×E interaction effect. However, model fit to the simulated data demonstrates that the equality does not hold exactly any longer. Therefore, to derive the properties of the two estimators of G×E interaction, we relied on the simulations instead of analytical solution and

asymptotic theory to generalize our conclusions for limited-sample estimates. Simulation results based on a limited sample are presented in this chapter.

4.5. Simulation Study Set-up

A simulation study was conducted to evaluate sample properties of $\hat{\beta}_{CC}$ and $\hat{\beta}_{cases} - \hat{\beta}_{controls}$. For a total of 1500 cases and 1500 controls, the genotypes of 5000 SNPs were generated for each replicate. Power, type I error, and MSE of both $\hat{\beta}_{CC}$ and $\hat{\beta}_{cases} - \hat{\beta}_{controls}$ were assessed based on 100 replicates. The phenotype variable *D* and the exposure variable *E* were generated as binary variables, where 1 stands for cases or exposed and 0 for controls or non-exposed, respectively. The genotype *G* at the locus was coded as 0/1/2 according to the minor allele count and was generated to satisfy *Hardy-Weinberg equilibrium*. As one of the 5000 SNPs, a single SNP with G×E interaction effect was created, according to the model

$$logit\,(P(D = 1|\,G, E)) = p_d + \beta_{G\times E}\, G\times E,$$

with disease prevalence p_d = 0.05 and exposure frequency p_e = 0.1, 0.3, odds ratio of the interaction effect $exp(\beta_{G\times E}) = OR_{G\times E} = 1.5, 2, 2.5$ and the interacting SNP MAF $p_a = 0.1, 0.3$. All the other SNPs were sampled with a MAF from a beta distribution *B(1,3)* truncated to [0.01, 0.5]. The genotypes of these SNPs were generated independently of the disease or exposure (dummy SNPs). Dummy SNP genotypes were generated to evaluate the type I error of both estimators after multiple testing Bonferroni correction.

We fit three regression models to the data, equation (4.8), (4.9), (4.10), to estimate β_{CC}, β_{cases} and $\beta_{controls}$ for each of the 5000 SNPs and each of the 100 replicates.

$$logit\,(P(D = 1|\,G, E)) = \alpha_{CC} + \beta_G G + \beta_E E + \beta_{CC} G\times E \qquad \textbf{(4.8)}$$

$$logit\,(P(E = 1|\,G, D = 1)) = \alpha_{cases} + \beta_{cases} G \qquad \textbf{(4.9)}$$

$$logit\ (P(E = 1|\ G, D = 0)) = \alpha_{controls} + \beta_{controls}G \quad \textbf{(4.10)}$$

We compared the characteristics of $\hat{\beta}_{CC}$ and $\hat{\beta}_{cases} - \hat{\beta}_{controls}$ estimators visualizing the difference and evaluating their corresponding type I error, power, and MSE for the simulated scenarios. The goal of the simulation study was to compare $\hat{\beta}_{CC}$ and $\hat{\beta}_{cases} - \hat{\beta}_{controls}$ based on the listed characteristics. It is therefore sufficient that we simulate a single G×E interacting SNP to assess the power and dummy SNPs to assess the type I error.

4.6. Simulation Results

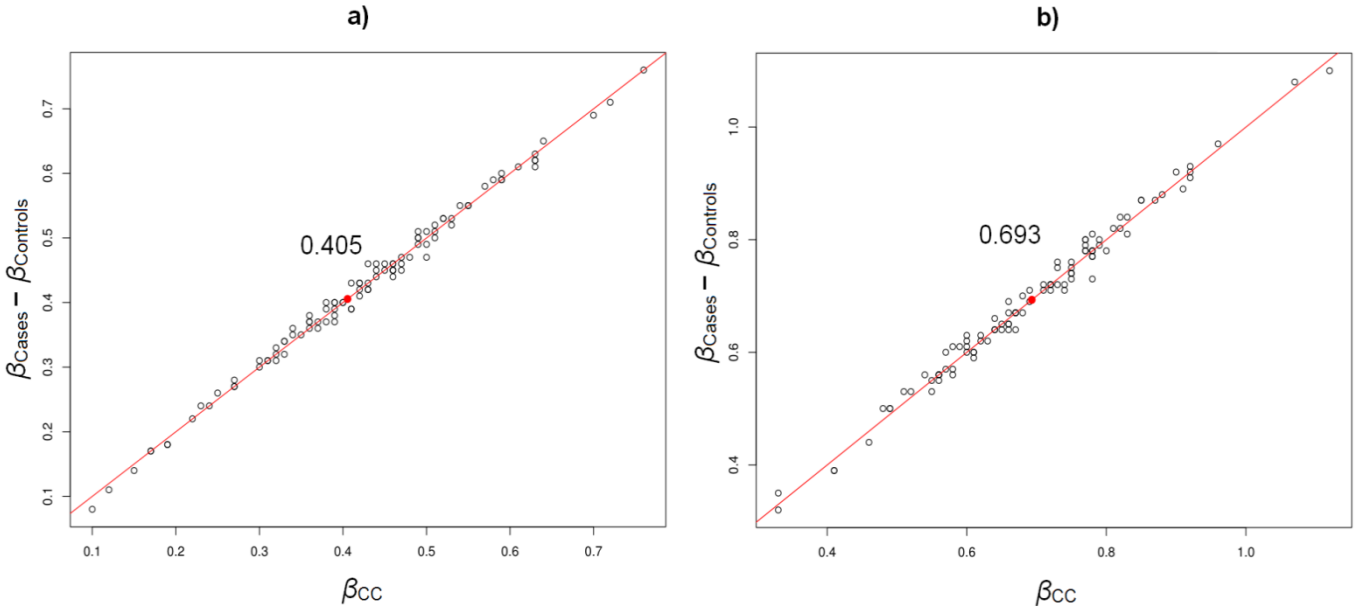
To demonstrate that the equality does not hold exactly, we plotted $\hat{\beta}_{CC}$ versus $\hat{\beta}_{cases} - \hat{\beta}_{controls}$. **Figure 4.1** demonstrates exemplarily differences in corresponding estimators with p_e = 0.3, p_a = 0.3, OR_{GxE} = *1.5, 2, 2.5*. We obtained similar results for other combinations of the parameter values (plots not shown). As shown on **Figure 4.1** and **Figure 4.2**, $\hat{\beta}_{CC}$ and $\hat{\beta}_{cases} - \hat{\beta}_{controls}$ are close to each other, however they are not exactly the same. Asymptotic theory suggests that for an infinitely large sample size, both estimators will converge to the simulated true values, in our example 0.405, 0.693, or 0.916, respectively, and are therefore asymptotically equivalent. Deviations from equality decrease with decreasing size of the estimated effect. For example, $OR_{G\times E}$ = 1.5 (a) compared to $OR_{G\times E}$ = 2.5 (c) in **Figure 4.1**. Deviations slightly increase with decreasing frequency of *E* (**Figure 4.2**c) and slightly decrease with decreasing frequency of *G* (**Figure 4.2**b), both compared to **Figure 4.2**a.

As MSE, type I error, power, and rank power are typical characteristics used to compare estimators, we considered them for both estimators. **Table 4.1** presents a comparison of MSE, type I error, power, and rank power for the following settings p_e = 0.1 or 0.3, p_a = 0.1 or 0.3 and exp(0.693) = $OR_{G\times E}$ = 2 versus exp(0.405) = $OR_{G\times E}$ = 1.5. Clearly in terms of these major characteristics, both estimators performed equally well. Since they also test the same

hypothesis, we can conclude that $\hat{\beta}_{CC}$ and $\hat{\beta}_{cases} - \hat{\beta}_{controls}$ are practically equivalent estimators of the G×E interaction effect even for samples that are moderate in size (1500 cases and 1500 controls) and are asymptotically equivalent.

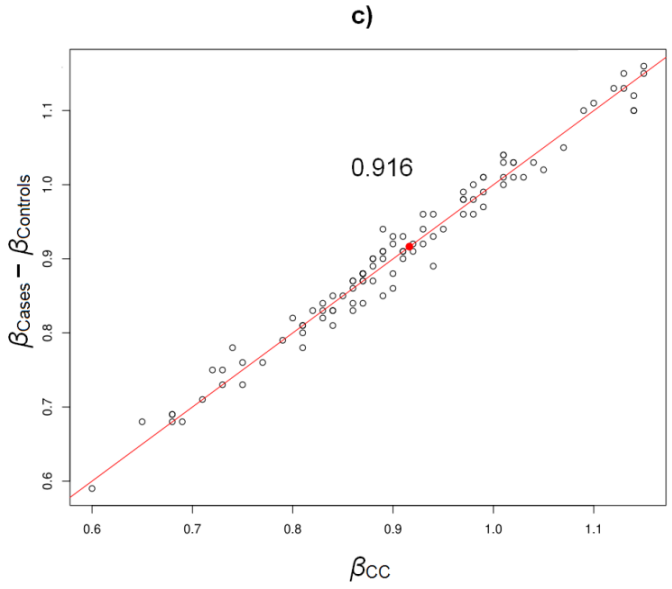
Figure 4.1 Comparison of β_{cases}-$\beta_{controls}$ vs. β_{cc} as estimators of G×E interaction for different $OR_{G\times E}$

a) pe = 0.3, pa = 0.3, exp(0.405) = ORGxE = 1.5



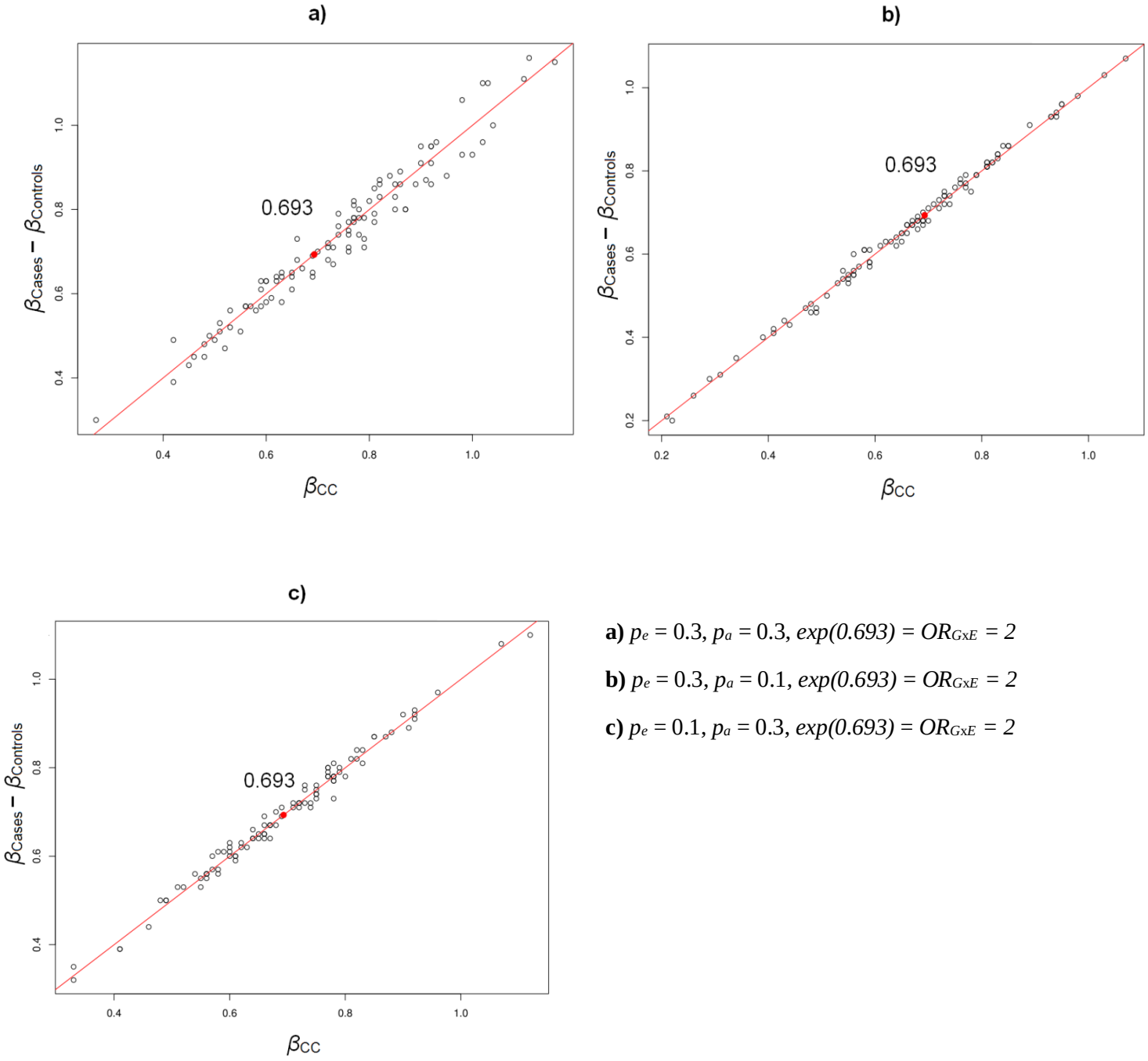
b) p_e = 0.3, p_a = 0.3, *exp(0.693)* = OR_{GxE} = *2*

c) p_e = 0.3, p_a = 0.3, *exp(0.916)* = OR_{GxE} = *2.5*



In bold are simulated true log($OR_{G\times E}$). Points located on the diagonal line represent the equality of two estimators $\hat{\beta}_{CC}$ and $\hat{\beta}_{cases} - \hat{\beta}_{controls}$, deviations represent violations of the exact equality.

Figure 4.2 Comparison of β_{cases}-$\beta_{controls}$ vs. β_{cc} as estimators of G×E interaction for different exposure frequency and allele frequency



a) $p_e = 0.3$, $p_a = 0.3$, $exp(0.693) = OR_{GxE} = 2$

b) $p_e = 0.3$, $p_a = 0.1$, $exp(0.693) = OR_{GxE} = 2$

c) $p_e = 0.1$, $p_a = 0.3$, $exp(0.693) = OR_{GxE} = 2$

In bold are simulated true log($OR_{G\times E}$). Points located on the diagonal line represent the equality of two estimators $\hat{\beta}_{CC}$ and $\hat{\beta}_{cases} - \hat{\beta}_{controls}$, deviations represent violation of the exact equality.

We conclude that the EHB-GE_{CHI} approach can be applied under the assumption of the "log-additive" mode of disease inheritance, as both estimators of G×E interaction are asymptotically equivalent and perform similarly when applied to moderately sized studies.

Table 4.1 Properties of two estimators for $OR_{G\times E}$

Estimator	$\beta_{G\times E}$	βcases-βcontrols	$\beta_{G\times E}$	βcases-βcontrols
Scenario	**$OR_{G\times E}$ = 1.5**		**$OR_{G\times E}$ = 2**	
	p_e = 0.3 & p_a = 0.3			
MSE	0.03	0.03	0.03	0.03
power	0.17	0.17	0.91	0.92
type I error	0.05	0.06	0.05	0.05
rank power top 1	0.38	0.38	0.95	0.95
rank power top 25	0.76	0.76	1.00	0.99
Scenario	**p_e = 0.3 & p_a = 0.1**			
MSE	0.06	0.06	0.06	0.06
power	0.04	0.04	0.36	0.34
type I error	0.04	0.05	0.04	0.04
rank power top 1	0.06	0.06	0.45	0.44
rank power top 25	0.35	0.35	0.86	0.83
Scenario	**p_e = 0.1 & p_a = 0.3**			
MSE	0.07	0.07	0.06	0.06
power	0.02	0.02	0.36	0.38
type I error	0.03	0.04	0.01	0.01
rank power top 1	0.07	0.07	0.67	0.67
rank power top 25	0.32	0.31	0.92	0.92

4.7. Covariate Adjustment

It is well recognized that in genetic association studies, including G×E interaction studies, great care should be taken to account for covariates to avoid bias. As defined in Chapter 2, a covariable is a variable in the analysis that relates to both the genotype of interest and to the phenotype (disease status), but is not an intermediate factor of the cause of disease (Rothman, Greenland et al. 1980). This is of key importance in case-control studies, as uncontrolled differences between disease carriers and healthy individuals in the study sample may lead to

spurious associations or mask true association signals (Zeggini and Morris 2010). In genetic association studies, it is generally recommended to adjust for age and sex of the individuals if the study design does not include matching for, these variables (Zeggini and Morris 2010). Other important covariables are those accounting for the ancestry of the individuals, such as principal components. As was discussed in detail in Chapter 3, failure to control for population stratification leads to bias in estimates.

EHB-GE_{CHI} was originally proposed without discussing covariate adjustment. Since EHB-GE_{CHI} requires estimates of the G-E correlation obtained within cases and controls, a proof is needed to illustrate that separate adjustment for covariates within cases and controls in a logistic regression framework would lead to the same estimates as an adjustment on the whole case-control sample.

Let Z denote any covariate, and G, E, D all binary. To perform association analysis adjusted for the covariate Z one can fit the following three logistic regression models to the data.

$$logit\,(P(D = 1|\,G, E)) = \alpha_{CC\_Z} + \beta_{G\_Z}G + \beta_{E\_Z}E + \beta_{CC\_Z}G\times E + \beta_{Z\_CC}Z^T \quad \textbf{(4.11)}$$

$$logit\,(P(E = 1|\,G, D = 1)) = \alpha_{cases\_Z} + \beta_{cases\_Z}G + \beta_{Z\_cases}Z^T \quad \textbf{(4.12)}$$

$$logit\,(P(E = 1|\,G, D = 0)) = \alpha_{controls\_Z} + \beta_{controls\_Z}G + \beta_{Z\_controls}Z^T \quad \textbf{(4.13)}$$

In the equations above $\beta_{CC\_Z} = log(OR_{G\times E|Z})$, where $OR_{G\times E|Z}$ is the population G×E interaction odds ratio adjusted for Z, likewise corresponding coefficients from (4.12) and (4.13) have the same interpretation.

To prove that the population interaction odds ratio, $OR_{G\times E|Z,}$ can indeed be written as the ratio of the population odds ratios in cases and controls respectively, or that the relationship $\beta_{CC\_Z} = \beta_{cases\_Z}\text{-}\beta_{controls\_Z}$ holds, assume that $OR_{G\times E}$ is independent of the covariate Z (in the case of a categorical Z, we assume that the G×E interaction effect is the same across the strata of Z). This assumption is commonly made in analyses that adjust for covariates Z as $G\times E\times Z$ terms are

rarely included in the regression model. In finite samples, the estimated interaction odds ratio will not generally be identical to the ratio of estimated odds ratios, but the result holds asymptotically. This follows from the consistency (provided the model is not misspecified) of the estimated odds ratios in cases and controls and the equality holding at the population level. Denote $f_0(G,E,Z)=f(G,E,Z|D=0)$ and $f_1(G,E,Z)=f(G,E,Z|D=1)$ as the conditional probability of (G, E, Z) given $D=0$ or $D=1$. Then

$$\begin{aligned} f(G=g, E=e, Z=z \mid D=d)=&\mu_0+\mu_1 D+\mu_2 E+\mu_3 G+\mu_4 Z \\ &+\beta_1 DE+\beta_2 DG+\beta_3 DZ \\ &+\gamma DGE \\ &+\alpha_1 GE+\alpha_2 EZ+\alpha_3 GZ+\alpha_4 GEZ \\ &+\delta_1 DEZ+\delta_2 DGZ+\delta_3 DGEZ \end{aligned}$$

Then, $P(D=1|G=g, E=e, Z=z)=\frac{e^{f1(G,E,Z)}}{\left(e^{f1(G,E,Z)}+e^{f0(G,E,Z)}\right)}$

so $Odds_{G=1,E=1|Z} = P(D=1|G=1\ E=1, Z=z)/(1- P(D=1|G=1\ E=1, Z=z))$

$Odds_{G=0,E=0|Z} = P(D=1|G=0\ E=0, Z=z)/(1- P(D=1|G=0\ E=0, Z=z))$

$Odds_{G|E=0,Z} = P(D=1|G=1\ E=0, Z=z)/(1- P(D=1|G=1\ E=0, Z=z))$

$Odds_{E|G=0,Z} = P(D=1|G=0\ E=1, Z=z)/(1- P(D=1|G=0\ E=1, Z=z))$

Then, the case-control estimate of an adjusted G×E interaction odds ratio

$$OR_{G\times E|Z}=(Odds_{G=1,E=1|Z} Odds_{G=0,E=0|Z})/(Odds_{G|E=0,Z} Odds_{E|G=0,Z})$$

reduces to $e^{(\gamma+\delta_3 Z)}$.

Now consider the estimator based on the comparison of adjusted G-E correlations in cases and controls, which EHB-GE_{CHI} proposes. Proceeding as above, starting with the probability of exposure given genotype, covariate, and disease status:

$$P(E=1|G=g, Z=z, D=d)=\frac{e^{f(G,E=1,Z|D)}}{\left(e^{f(G,E=1,Z|D)}+e^{f(G,E=0,Z|D)}\right)}$$

$OR_{G\text{-}E\_cases}$ reduces to $e^{(\gamma+\alpha_1+\alpha_4 Z+\delta_3 Z)}$ and $OR_{G\text{-}E\_controls}$ reduces to $e^{(\alpha_1+\alpha_4 Z)}$, so

$$OR_{G\times E|Z}= OR_{G\text{-}E\_cases|Z}/OR_{G\text{-}E\_controls|Z}=e^{(\gamma+\delta_3 Z)},$$

that is, the equivalent to that for the case-control estimator.

Both are equal to the G×E interaction parameter, $exp(\gamma)$, provided only that $\delta_3 = 0$, i.e., that there is no 4-way interaction in the log-linear model (3-way interaction of $G\times E\times Z$ in logistic regression framework), or equivalently that Z does not modify the magnitude of the G×E interaction. There is no need to assume Z to be independent of G, E, or Z or any of its lower-order interactions for this result to hold.

Another possible solution to the covariates problem relies on the simultaneous estimation of G-E correlation in cases and controls adjusted for the covariate Z in a single model. This can be performed with log-linear modeling. It was demonstrated by Umbach and Weinberg and later by Eiichiro (Umbach and Weinberg 1997, Eiichiro 2004) that the estimated coefficients of G-E correlation obtained employing a log-linear model are exactly equivalent to those calculated based on the logistic model. Therefore, one can fit the following single log-linear model to the data to obtain β_{CC_Z} and $\beta_{controls\_Z}$ and consequently β_{cases_Z} as $\beta_{CC\_Z} +\beta_{controls\_Z}=\beta_{cases\_Z}$. We used notations as in (Umbach and Weinberg 1997).

Let n_{ijk} denote the number of subjects having $D = i$, $G = j$, and $E = k$ and N is the total number of individuals. For the sake of simplicity, assume that i, j, and k all binary. Replacing any subscript with a dot (.) denotes summation over the subscript. Our data are summarized in **Table 4.2**.

Table 4.2 Data representation for log-linear model

	E=1		***E=0***		**Total**
	G=1	***G=0***	***G=1***	***G=0***	
D=1	n_{111}	n_{101}	n_{110}	n_{100}	$\mathbf{n_{1..}}$
D=0	n_{011}	n_{001}	n_{010}	n_{000}	$\mathbf{n_{0..}}$
Total	$\mathbf{n_{.11}}$	$\mathbf{n_{.01}}$	$\mathbf{n_{.10}}$	$\mathbf{n_{.00}}$	**N**

Log-linear models assume a multiplicative relationship between categorical variables, that is, the expected value of any cell counts n_{ijk} can be modeled as a product of the overall number of observations (N) and the main effect of each variable and their respective interaction.

$$log(n_{ijk})=\alpha_0+\beta_{G\_controls}G+\beta_{E\_controls}E+\beta_{controls}G\times E+\alpha D+\beta_G G\times D+\beta_E E\times D+\beta_{CC}G\times E\times D \quad \mathbf{(4.14)}$$

Equation (4.15) links the logistic model given in (4.8) to the log-linear model in (4.14) (Bishop, Fienberg et al. 2007))

$$log(n_{1jk})-log(n_{0jk})=logit(P(D=1|G,E,Z))=\alpha+\beta_G G+\beta_E E+\beta_{CC}G\times E \quad \mathbf{(4.15)}$$

Log-linear models allow us to model cell counts for eight cells in the table above explicitly and simultaneously. Clearly from (4.14), adjustment for the categorical covariate Z can be handled in the log-linear models framework to obtain estimates β_{CC_Z} and $\beta_{controls\_Z}$ simultaneously adjusted for Z. For additional discussion on covariates adjustment in log-linear data analysis please refer to (Umbach and Weinberg 1997). Based on our proof for the logistic regression framework and discussion on log-linear models, we conclude that the analysis of the data adjusted for the covariates can be applied to the EHB-GE_{CHI} approach to G×E interactions.

Chapter 5

5. Modified Empirical Hierarchical Bayes Approach for G×E Interaction

In Chapter 4, we introduced the EHB-GE_{CHI} approach for G×E interaction. It combines estimates of G×E interaction from the case-only test and posterior estimates of population-based G-E correlations made among controls to construct a powerful rank statistic. As discussed in Chapter 4, this statistic can only be applied to perform ranking and is not recommended for significance testing, since the test has inflated rate of type I error.

In the current chapter, we introduce an improved, computationally faster, and more stable alternative approach we have developed, which obtains the posterior estimates of G-E correlation in controls basing on a normal-normal hierarchical model. The normal-normal model is a classic example of the empirical Bayes inferences framework (Chapter 4). This model allows us to reduce the variance of G-E control-based estimates. Thus, we gain power over the case-control statistic while keeping a tighter control on type I error than a pure case-only test. We name this new approach the empirical hierarchical Bayes approach for G×E interaction (EHB-GE_{NN}). With EHB-GE_{NN}, only a single hyperparameter τ^2, the common variance of G-E effects, has to be estimated in contrast to three hyperparameters in the previous EHB-GE_{CHI} method (Sohns, Viktorova et al. 2013), (Chapter 4). This leads to greater estimation stability and simplicity. Thanks to the improved type I error rate control, EHB-GE_{NN} can be utilized for significance testing and not just for ranking of G×E interactions, a notable and important improvement compared to EHB-GE_{CHI}.

We conducted an extensive simulation study to evaluate our approach. We compared EHB-GE_{NN} with EHB-GE_{CHI} and Mukherjee's empirical Bayes approaches in terms of the achieved power and inflation of type I error, since these two methods were altogether favorable to other

tests (CC, CO, MUR) in most situations (Sohns, Viktorova et al. 2013). We compared rank power of EHB-GE_{NN} against all other G×E methods, EHB-GE_{CHI}, CC, CO, MUK-EB, and MUR.

As already mentioned in Chapter 1, joint discovery of G×E interactions and genetic main effects may aid the detection of genetic variants potentially missed by the initial GWASs focusing on the marginal marker-trait association or pure interaction analysis (Kraft, Yen et al. 2007, Dai, Logsdon et al. 2012, Vanderweele, Ko et al. 2013). Thus, methods capable of considering genetic main and G×E interaction effects simultaneously are important.

In 2007, Kraft and colleagues presented a joint likelihood ratio test of SNP main and G×E interaction for case-control data (Kraft, Yen et al. 2007). In 2011, Dai and colleagues exploited Kraft's approach and proposed three joint tests different from Kraft et al. in two respects (Dai, Logsdon et al. 2012). They proposed using a marginal genetic association component instead of the main effect estimate in a model for G×E interaction. Secondly, they demonstrated that not only the CC, but also the CO or the MUK-EB estimators can be used to test the G×E interaction component (Dai, Logsdon et al. 2012). Dai's joint tests are more flexible than and at least as powerful as Kraft's likelihood test. Therefore, we outline three joint *2-degree-of-freedom* tests (CC^J, CO^J, MUK-EB^J), proposed by Dai (Dai, Logsdon et al. 2012). Similarly to Dai, we construct the joint EHB-GE_{NN}^J test, proposing combining both estimators of the genetic marginal effect and the EHB-GE_{NN} G×E interaction in a single statistic. Both EHB-GE_{NN} and joint EHB-GE_{NN}^J do not assume G-E independence, which makes them favorable in the genome-wide testing context.

5.1. The Normal-Normal Model

(Morris 1983)

Assume that p parameters need to be estimated $\theta=(\theta_1,\ldots,\theta_p)$ and that we have p independent unbiased estimates $X=(X_1,\ldots,X_p)$, $E(X_i)=\theta_i$, $i=1\ldots p$. Assuming X_i's are independently normal, we can write

$$\text{Level 1} \qquad X_i|\theta_i \sim N(\theta_i, V),$$
$$\text{Level 2} \qquad \theta_i|\mu, A \sim N(\mu, A),$$

with $V=var(X_i)$, $A=var(\theta_i)$ (equal variances case), $i=1\ldots p$, and $\eta=(\mu, A)$ are level hyperparameters. The estimates X_i are usually a statistic of the original data, for example sample means. Assume that V, common variance is known and do not need to be estimated in a Bayesian manner. We concentrate on the estimation of hyperparameters θ.

For this model, the marginal distribution of X_i is given by

$$X_i|\mu, A \sim N(\mu, V + A),$$

and estimates of η can be obtained by maximizing the *log* of the marginal likelihood, given as

$$\log(L(X|\mu)) = \log\left(\frac{1}{\sqrt{2\pi(V+A)}} e^{\left(-\frac{\sum(X_i-\mu)^2}{2(V+A)}\right)}\right)$$

Maximization yields the MMLE $\hat{\mu} = \frac{1}{n}\sum X_i$. Plugging in this estimate, the posterior distribution $\pi(\theta_i|X_i, \hat{\mu}, A)$ is obtained as

$$\theta_i|X_i \sim N(B\hat{\mu} + (1-B)X_i, BA),$$

Where $B = \frac{B}{B+A}$.

The posterior estimate of θ_i is given as

$$\hat{\theta}_i = B\hat{\mu} + (1 - B)X_i.$$

In the case of *A* being an unknown hyperparameter, it can also be estimated from the data. The overall data variance is estimated by $s^2 = \frac{1}{n}\sum(X_i - \hat{\mu})^2$ and the known *V*, then

$$\hat{A} = \max(0,\ s^2 - V).$$

Any desirable number of hierarchical levels can be implemented within EHB models.

5.2. Construction of the EHB-GE$_{NN}$ Statistics

As discussed earlier, G×E interaction is challenging to detect, particularly on the genome-wide scale, mainly due to the lack of power of common interaction tests. Addressing this issue, the CO test can be performed to increase power, however a large inflation in type I error associated with the test in the presence of G-E correlation must be taken into consideration. It is essential to account for population-level G-E correlation in studies of G×E interaction when using case-only-related methods. The method we propose aims to estimate this correlation efficiently based on the information in controls as a sample from the general population, such that an increase in power over CC is achieved, while keeping type I error inflation low.

Consider a case-control study with a total of *N* individuals, $N_1 = ccr \times N_0$ cases and N_0 controls, where *ccr* stands for case-control ratio. Let *m* denote a SNP, *m = 1…M*, where *M* is the total number of SNPs in a GWAS analysis. Let *G* denote a genotype and G_m denote a genotype at a specific SNP *m*. Let *E* denote the exposure variable and *D* the disease outcome variable. Let us also assume that all three variables *D*, *G*, and *E* are binary.

We assume that G-E effects in controls, i.e. parameters $\beta_m^{controls}$, *m=1…M*, can be estimated, yielding independent and unbiased estimates $\hat{\beta}_m^{controls}$. Standard logistic regression models

(5.1) and (5.2) below can be applied per SNP to derive the needed effect estimates of G-E within cases and within controls ($\hat{\beta}_m^{cases}$ and $\hat{\beta}_m^{controls}$) together with their variances ($(\hat{\sigma}_m^{cases})^2$,$(\hat{\sigma}_m^{controls})^2$).

$$logit(P(E=1|G, D=1))=\alpha_{cases}+\beta_{cases}G_m \quad \textbf{(5.1)}$$

$$logit(P(E=1|G, D=0))=\alpha_{controls}+\beta_{controls}G_m \quad \textbf{(5.2)}$$

Remember that the G×E interaction estimate on the multiplicative scale can be represented as the difference of β_{cases}-$\beta_{controls}$ per SNP, (Chapter 2). We propose estimating β_{cases}-$posterior(\beta_{controls})$ for each SNP to reduce the variance of the control-based G-E correlation.

Adopting empirical hierarchical Bayes inference theory (Chapter 4), we propose the hierarchical Bayes model (5.3)-(5.4) to estimate the posterior mean of G-E correlation in controls and its variance.

Level 1 $\hat{\beta}_m^{controls} \mid \beta_m^{controls} \sim N(\beta_m^{controls}, (\sigma_m^{controls})^2)$ **(5.3)**

Level 2 $\beta_m^{controls} \mid \tau^2 \sim N(0, \tau^2)$ **(5.4)**

If $\hat{\beta}_m^{controls}$ are maximum likelihood estimates (*MLE*) of the true parameters, obtained from equation (5.2), then, referring to the asymptotic theory, we can assume that they are normally distributed and can construct a valid two-stage model (5.3)-(5.4).

Here $\sigma_m^{controls}$ is the standard error of $\hat{\beta}_m^{controls}$. Each $\sigma_m^{controls}$ can be substituted by the corresponding *MLE*. The prior mean of G-E effects in controls ($\beta_m^{controls}$) is centered at zero, since we expect no association for the vast majority of SNPs and the prior variance is τ^2. The hyperparameter τ^2 is estimated borrowing information across all SNPs from the marginal distribution of $\hat{\beta}_m^{controls}$, given in (5.5) by maximizing the *log* of the marginal likelihood (*L*), given by (5.6) with respect to τ^2

$$\hat{\beta}_m^{controls} \mid \tau^2 \sim N(0, (\sigma_m^{controls})^2 + \tau^2) \quad \textbf{(5.5)}$$

$$L = \prod_{m=1}^{M} f_m(\hat{\beta}_m^{controls} \mid \tau^2) \propto \prod_{m=1}^{M} \frac{\exp\left(-\frac{1}{2} \sum_{m=1}^{M} \frac{(\hat{\beta}_m^{controls})^2}{(\sigma_m^{controls})^2 + \tau^2} \right)}{\sqrt{(\sigma_m^{controls})^2 + \tau^2}} \qquad \textbf{(5.6)}$$

where $f_m(\hat{\beta}_m^{controls} \mid \tau^2)$ is the marginal density function. From (5.3) and (5.4), we derive the posterior distribution of the unknown parameters $\beta_m^{controls}$, $m = 1 \ldots M$, see (5.7) below.

$$\beta_m^{controls} \mid \hat{\beta}_m^{controls}, \tau^2 \sim N((\beta_m^{controls})^*, (\sigma_m^{2\,controls})^*) \qquad \textbf{(5.7)}$$

$(\beta_m^{controls})^* = posterior(\beta_m^{controls}) = (1\text{-}B_m) \times \hat{\beta}_m^{controls}$,

$(\sigma_m^{2\,controls})^* = (1\text{-}B_m) \times (\hat{\sigma}_m^{controls})^2$,

$B_m = (\sigma_m^{controls})^2 / ((\sigma_m^{controls})^2 + \tau^2)$ and

$\hat{B}_m = (\hat{\sigma}_m^{controls})^2 / ((\hat{\sigma}_m^{controls})^2 + \hat{\tau}^2)$

Here, B_m is the SNP specific shrinkage factor, $0 \leq B_m \leq 1$. The amount of shrinkage depends on τ^2, with virtually no shrinkage in $\hat{\beta}_m^{controls}$ when $B_m \approx 0$ ($\tau^2 \to \infty$) and complete shrinkage to zero when $B_m = 1$ ($\tau^2 = 0$).

The corresponding variance of $posterior(\beta_m^{controls})$ is given by

$$Var(posterior(\beta_m^{controls})) = (1 - B_m)^2 \times (\hat{\sigma}_m^{controls})^2$$

We propose the following test statistic for G×E per SNP *m* accounting for population G-E.

$$Z_m^{EHB-GE_{NN}} = \frac{\hat{\beta}_m^{cases} - (\beta_m^{controls})^*}{\sqrt{(\hat{\sigma}_m^{cases})^2 + Var\left((\beta_m^{controls})^*\right)}}$$

Alternatively, this can be done simultaneously, using a *log-linear* model framework to derive $\hat{\beta}_m^{cases}$, $(\hat{\sigma}_m^{cases})^2$, $\hat{\beta}_m^{controls}$, $(\hat{\sigma}_m^{controls})^2$ (Umbach and Weinberg 1997). The possibility for

covariate adjustment separately within cases and within controls as well as effect estimation based on a multilevel or continuous genotype or exposure variable is discussed in Chapter 4. All extensions we derived for EHB-GE_{CHI} are applicable to the EHB-GE_{NN} approach.

5.3. Simulation Study Set-up

We simulated genotypes (*G*) at 10,000 SNPs (*m*), one environmental factor (*E*) and disease outcome (*D*) for a total of 3000 individuals, with three different case-control ratios (*ccr*), 1:1 (1500 : 1500), 1:2 (1000 : 2000) and 2:1 (2000 : 1000) to represent balanced and unbalanced study designs. Linkage disequilibrium (LD) between SNPs was not modeled. Presence of the correlation structure between SNPs should not affect the validity of our approach but only its efficiency. All three variables *D*, *G*, and *E* were generated as binary, where *1* codes for *cases*, *carriers*, and *exposed* and *0* stands for *controls*, *non-carriers*, and *unexposed*. **Table 5.1** summarizes the simulation scenarios generated. The simulation setup remains the same as in (Sohns 2012) for a valid comparison across the G×E interaction methods.

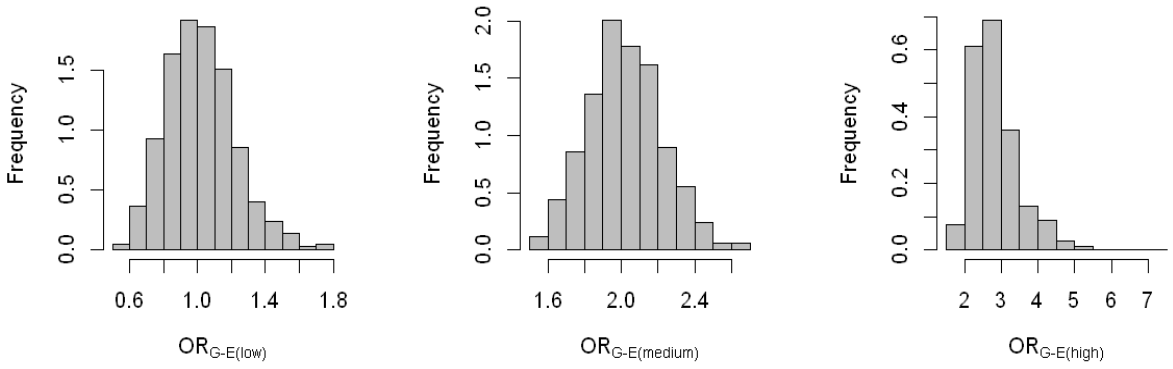
A single SNP with G×E interaction effect was generated based on the following disease model

$$logit(P(D=1|G_m, E)) = p_d + \beta_{G\times E} G_m \times E$$

with baseline disease risk $p_d = 0.01$ or 0.05, exposure frequency $p_e = 0.1, 0.3,$ or 0.5, genotype carrier frequency $p_g = 0.1, 0.3,$ or 0.5 and odds ratio associated with G×E interaction $exp(\beta_{G\times E}) = OR_{G\times E} = 1.2, 1.5, 2, 2.5, 3$. Note that *E* and *G* are not modeled as main effects. However, the frequencies influence the number of individuals in each G×E stratum. Among the total number of SNPs, we designed a number of signals, $N_{G\text{-}E}$ =0, 1000, 2500, or 5000, with population-based G-E correlation. The strength of these correlations was classified in three groups, *low*, *medium*, and *high*, based on the sampling distribution of the corresponding

coefficients *N(0, log(1.5)/2)*, *N(0.7,0.1)*, and *exp(N(0,log(1.5)/2))*, respectively. **Figure 5.1** demonstrates the distribution of these three groups of G-E correlation effects in controls. For the remaining SNPs, without any G×E interaction or G-E correlation effect, the frequency of the at-risk genotype carriers was sampled from a beta distribution *B(1, 3)* truncated to [0.01, 0.5]. Analysis was conducted for each of 1000 replicated datasets.

Figure 5.1 Distribution of G-E correlation effects in controls



$OR_{G\text{-}E(low)} \sim N(0,log(1.5)/2)$, $OR_{G\text{-}E(medium)} \sim N(0.7,0.1)$, and $OR_{G\text{-}E(high)} \sim exp(N(0,log(1.5)/2))$

Table 5.1 Simulation study settings, 3240 scenarios

Scenario	Frequency	Number	OR
Prevalence of D	p_d = 0.01, 0.05		
E	p_e = 0.1, 0.3, 0.5	1	OR_E=1
G×E interaction SNP	p_g = 0.1, 0.3, 0.5	1	$OR_{G\times E}$=1.2, 1.5, 2, 2.5, 3
G-E correlated SNPs	0.01 - 0.5	0, 1000, 2500, 5000	$OR_{G\text{-}E}$=(low, medium, high)

5.4. Simulation Study Results

We evaluated the performance of EHB-GE_{NN} with regard to type I error, power, and rank power and compared those to EHB-GE_{CHI} and MUK-EB. As expected, the power of EHB-GE_{NN} increases with increasing G×E effect size, increasing p_g, p_e frequencies, and decreasing p_d. The

power of EHB-GE_{NN} depends on *ccr*, generally favoring a balanced design (**Figure 5.2**). This behavior persists independently of the number of G-E correlations. **Figure 5.2** illustrates different *ccr*s, different G-Es, and effect sizes for $OR_{G\times E}$=2.5 and p_g=0.3, p_e=0.3, p_d=0.05 (upper row) or $OR_{G\times E}$=2.5, p_g=0.5, p_e=0.5, p_d=0.05 (lower row). As can be seen from **Figure 5.2**, having twice as many controls as cases is the most unfavorable situation. **Figure 5.2** also indicates an increase in power with increasing genotype frequency and exposure. The power depends on the number of SNPs with G-E. The EHB-GE_{NN} test achieves higher power in the presence of low-effect G-E correlation when compared to medium or high effects. The decline in power depending on the number of G-E correlations is not dramatic and even less so in high effect scenarios.

Subsequently, we compared the type I error rate and the power of EHB-GE_{NN} with EHB-GE_{CHI} and MUK-EB. To account for the possible inflation of the family-wise type I error rate due to the multiple testing of 10,000 SNPs we used Bonferroni correction, by setting up the significance level for the p-value of each test to 5×10^{-5}. For clarification note that, under the null hypothesis the percentage among 1000 replications is given, where any one of 10,000 SNPs is significant. Evaluation shows that relative loss in absolute power of the EHB-GE_{NN} method compared to the EHB-GE_{CHI} or the MUK-EB approaches are smaller on average than inflation of the type I error of the later methods in comparison to EHB-GE_{NN}. This conclusion is depicted in **Tables 5.2** and **5.3**. **Table 5.2** portrays type I error and power of EHB-GE_{NN}, EHB-GE_{CHI}, and MUK-EB for 1500 cases and 1500 controls, $OR_{G\times E}$*=2.5,* p_d*=0.05,* with p_g*=0.3,* p_e*=0.3* on the left and p_g*=0.5,* p_e*=0.5* on the right in the absence of or presence of a large number of G-E correlations with either medium or high effect. **Table 5.3** portrays the results for p_d*=0.01*, all other conditions remaining the same as in **Table 5.2**. If the G-E independence assumption is valid (OR_G-E = 1 in **Tables 5.2, 5.3**), all three approaches maintain type I error at a nominal 5% level or lower (see $OR_{G\times E}$ =1, *in italics*). **Tables 5.2** and

5.3 also reflect that under the assumption of G-E independence (OR_G-E=1), the type I error rate of EHB-GE_{NN} is less conservative than that of EHB-GE_{CHI} or MUK-EB. This leads to a power gain for our new approach in such settings. We therefore conclude that EHB-GE_{NN} is more powerful than EHB-GE_{CHI} and MUK-EB, under the assumption of G-E independence (**Table 5.2** and **Table 5.3**, upper blocks). In the presence of a large number of G-Es of medium to high effect size, EHB-GE_{NN} always holds type I error less than or equal to 10%, except for situations with an infrequent environmental factor, i.e. in our simulation p_e=0.1, when type I error can rise to 20%. In this case, the responsible sub-stratum is not large enough to estimate the variance of the correlation signals properly. Clearly, EHB-GE_{NN} has much lower type I error compared to EHB-GE_{CHI} and MUK-EB in the presence of G-E correlations of medium to high effect size. Controlling type I error is a critical issue when performing significance testing in contrast to a ranking of test statistics for follow-up. Hence, the ability of EHB-GE_{NN} to maintain it at a reasonable 10% or lower level compared to over 50% for EHB-GE_{CHI} or around 20% for MUK-EB (**Table 5.2** and **Table 5.3**) is a clear advantage of the EHB-GE_{NN} approach. **Table 5.2** and **Table 5.3** also present results on the power of the three approaches. In general, EHB-GE_{NN} appears to have lower power compared to the competitors; this should be seen as a compromise with the type I error control. The EHB-GE_{NN} approach always has greater or equal power compared to the classic CC test (data not shown). To evaluate the relative loss of power combined with a decrease in type I error towards an acceptable level for EHB-GE_{NN} versus the other two considered methods, we plotted the differences in type I error of EHB-GE_{CHI} (**Figure 5.3**, upper row) or MUK-EB (**Figure 5.3**, lower row), respectively, minus type I error of EHB-GE_{NN} on the *y-axis* and the corresponding differences in power on the *x-axis*. Please note the difference in the scales between the upper (EHB-GE_{NN} versus EHB-GE_{CHI}) and lower (EHB-GE_{NN} versus MUK-EB) rows of **Figure 5.3**. Each of the points on the graphs represents one simulated scenario. Points above the diagonal reflect the situation in which the

increase in type I error level is greater than the power gain of the rival method compared to EHB-GE_{NN}. In most situations, we observe a much larger type I error level increase compared to the corresponding power increase of the competitors. In many situations, both EHB-GE_{CHI} and MUK-EB do not even gain in power at all at the expense of a large inflation in type I error.

To summarize the results on power and type I error, EHB-GE_{NN} is an improvement on the earlier EHB-GE_{CHI} approach, because it maintains type I error rate reasonably well in the presence of G-E correlations, whereas the loss in power is not very critical. The EHB-GE_{NN} test is more powerful than the case-control test. In contrast to MUK-EB, it does not assume G-E independence in the general population. Moreover, EHB-GE_{NN} is computationally much faster than EHB-GE_{CHI} and more stable in terms of parameter estimation.

We also evaluated the rank power of our approach. EHB-GE_{NN} always has equal or greater rank power than the CC or CO methods in the presence of G-E correlations. The rank power gain of EHB-GE_{NN} compared to CO is extreme, reaching almost 100% in the presence of a large number of G-E correlations with high effect size, because the CO test has almost no power in these scenarios. EHB-GE_{NN} nearly always has greater rank power than MUR except in some situations of low G-E correlations, when for a rare exposure variable EHB-GE_{NN} lacks in power. On average over the 1000 replicates, EHB-GE_{NN} has about 5% lower rank power than EHB-GE_{CHI} for the top 100 ranks. The rank power of EHB-GE_{NN} was often lower than that for MUK-EB but not dramatically so. Nevertheless, in scenarios with low G-E correlation signals and some scenarios with high effect signals, the rank power of EHB-GE_{NN} was larger than that of MUK-EB. Rank power of EHB-GE_{NN} vs. CC, MUR, CO, MUK-EB, and EHB-GE_{CHI} is demonstrated in **Figures 5.4-5.6** for different ccr (1:1, 1:2, 2:1).

Figure 5.2 Power of EHB-GE$_{NN}$ to detect a SNP with GxE interaction for *ccr = 1:1, 1:2, 2:1* and different numbers of G-E correlations (# of G-E correlation) with different effect sizes OR_G-E low, medium and high, $OR_{G\times E}$ = 2.5, p_g = 0.3, p_e = 0.3, p_d = 0.05 (upper row) and $OR_{G\times E}$ = 2.5, p_g = 0.5, p_e = 0.5, p_d = 0.05 (lower row).

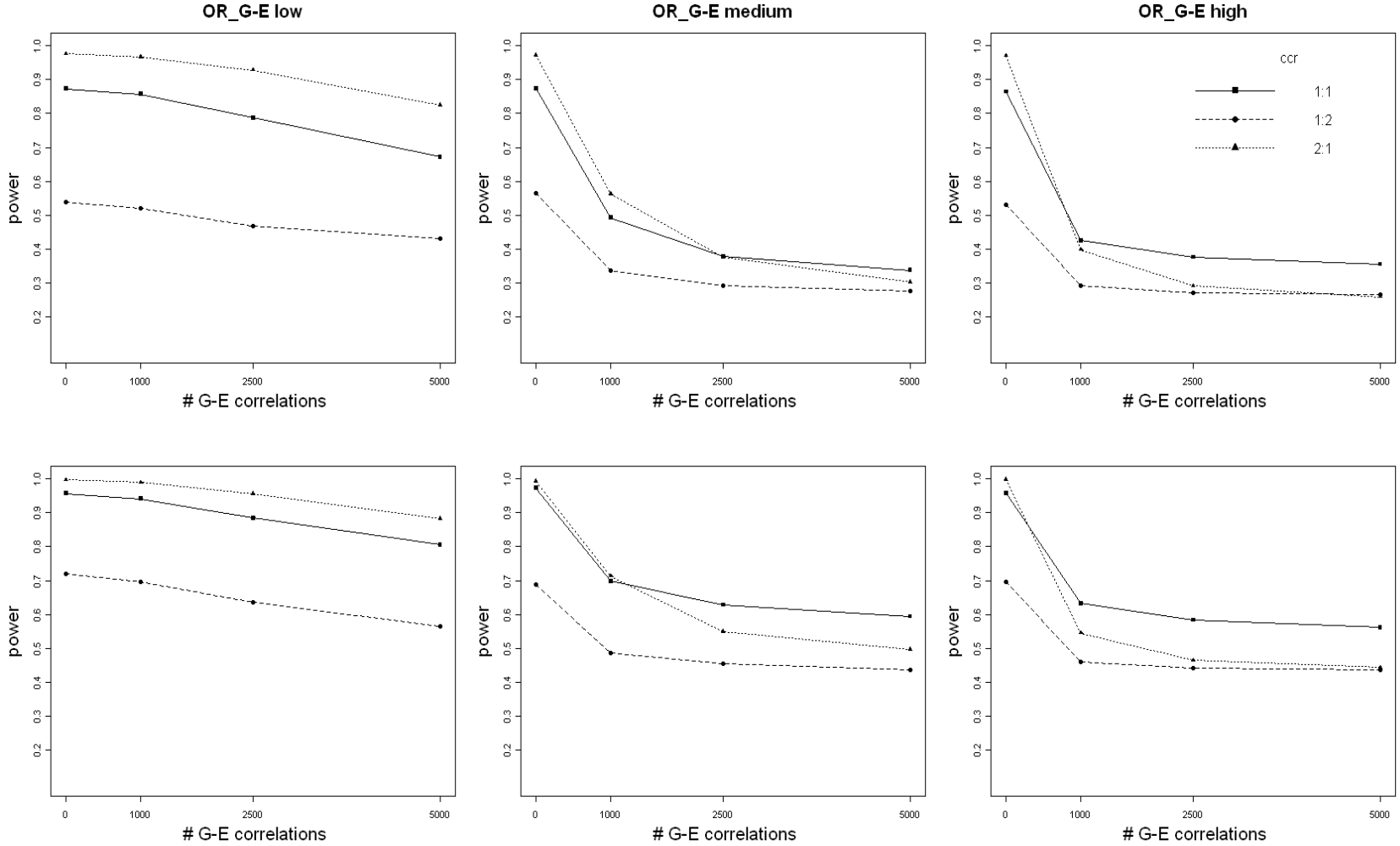


Table 5.2 Type I error (in *italic*) and Power of EHB-GE_{NN}, EHB-GE_{CHI}, MUK-EB, p_d=0.05

***ccr* =1:1**			p_d **= 0.05,** p_e **= 0.3,** p_g **= 0.3**			p_d **= 0.05,** p_e **= 0.5,** p_g **= 0.5**		
OR_G-E	***N_G-E***	$OR_{G\times E}$	***EHB-GE_{NN}***	***EHB-GE_{CHI}***	***MUK-EB***	***EHB-GE_{NN}***	***EHB-GE_{CHI}***	***MUK-EB***
1	*0*	*1*	***0.05***	***0.03***	***0.01***	***0.04***	***0.03***	***0.02***
		1.2	0.00	0.00	0.00	0.00	0.00	0.00
		1.5	0.01	0.01	0.01	0.20	0.19	0.09
		2	0.22	0.23	0.11	0.97	0.94	0.80
		2.5	0.74	0.72	0.42	1.00	1.00	0.99
		3	0.96	0.95	0.73	1.00	1.00	1.00
medium	*2500*	1	***0.10***	***0.92***	***0.18***	***0.06***	***0.82***	***0.06***
		1.2	0.00	0.00	0.00	0.00	0.00	0.00
		1.5	0.02	0.09	0.05	0.05	0.18	0.09
		2	0.38	0.83	0.62	0.63	0.93	0.80
		2.5	0.87	0.98	0.95	0.97	0.99	0.98
		3	0.99	0.99	1.00	1.00	0.99	1.00
	5000	1	***0.08***	***0.93***	***0.33***	***0.05***	***0.82***	***0.11***
		1.2	0.00	0.00	0.00	0.00	0.00	0.00
		1.5	0.02	0.09	0.05	0.04	0.17	0.09
		2	0.34	0.81	0.62	0.60	0.92	0.80
		2.5	0.84	0.96	0.95	0.97	0.99	0.98
		3	0.99	0.98	0.99	1.00	1.00	1.00
high	*2500*	*1*	***0.05***	***0.54***	***0.08***	***0.05***	***0.32***	***0.30***
		1.2	0.00	0.00	0.00	0.00	0.00	0.00
		1.5	0.01	0.08	0.04	0.04	0.18	0.10
		2	0.38	0.84	0.62	0.58	0.95	0.83
		2.5	0.85	0.99	0.95	0.97	0.99	0.99
		3	0.99	1.00	1.00	1.00	1.00	1.00
	5000	1	***0.04***	***0.52***	***0.15***	***0.05***	***0.30***	***0.04***
		1.2	0.00	0.00	0.00	0.00	0.00	0.00
		1.5	0.01	0.08	0.04	0.04	0.18	0.10
		2	0.36	0.83	0.62	0.56	0.95	0.82
		2.5	0.84	0.99	0.95	0.97	0.99	0.99
		3	0.98	1.00	1.00	1.00	1.00	1.00

Abbreviations: *ccr*, case-control ratio; p_d, baseline disease risk; p_e, exposure frequency; p_g, genotype carrier frequency; *OR_G-E*, odds ratio associated with G-E correlation; *N_G-E*, number of SNPs with population based G-E correlation (strength of G-E correlations); $OR_{G\times E}$, odds ratio associated with G×E interaction; *EHB-GE_{NN}*, parametric empirical hierarchical Bayes approach for G×E interaction, based on normal-normal model; *EHB-GE_{CHI}*, empirical hierarchical Bayes approach to G×E interaction, based on chi-distribution; *MUK-EB*, empirical Bayes shrinkage estimator;

Notes: $OR_{G\times E}$ *=1* corresponds to the null hypothesis;

N_G-E=0 corresponds to absence of G-E correlation;

Significance level per test is set to 5×10^{-5}, as 10,000 SNPs were simulated;

Table 5.3 Type I error (in *italic*) and Power of EHB-GE_{NN}, EHB-GE_{CHI}, MUK-EB p_d=0.01

***ccr* =1:1**			p_d **= 0.01,** p_e **= 0.3,** p_g **= 0.3**			p_d **= 0.01,** p_e **= 0.5,** p_g **= 0.5**		
OR_G-E	***N_G-E***	***OR_G×E***	***EHB-GE_{NN}***	***EHB-GE_{CHI}***	***MUK-EB***	***EHB-GE_{NN}***	***EHB-GE_{CHI}***	***MUK-EB***
1	*0*	1	***0.05***	***0.03***	***0.05***	***0.04***	***0.04***	***0.02***
		1.2	0.00	0.00	0.00	0.00	0.00	0.00
		1.5	0.13	0.14	0.06	0.23	0.21	0.12
		2	0.92	0.90	0.67	0.97	0.97	0.83
		2.5	1.00	1.00	0.93	1.00	1.00	1.00
		3	1.00	1.00	0.99	1.00	1.00	1.00
medium	*2500*	1	***0.09***	***0.92***	***0.20***	***0.07***	***0.76***	***0.07***
		1.2	0.00	0.00	0.00	0.00	0.00	0.00
		1.5	0.09	0.11	0.05	0.13	0.22	0.13
		2	0.84	0.90	0.68	0.90	0.96	0.83
		2.5	1.00	0.99	0.95	1.00	0.99	1.00
		3	1.00	0.99	0.99	1.00	1.00	1.00
	5000	1	***0.09***	***0.93***	***0.35***	***0.08***	***0.82***	***0.07***
		1.2	0.00	0.00	0.00	0.00	0.00	0.00
		1.5	0.06	0.10	0.05	0.08	0.22	0.13
		2	0.70	0.87	0.68	0.83	0.96	0.83
		2.5	0.99	0.97	0.95	0.99	0.99	0.98
		3	1.00	0.99	0.99	1.00	1.00	1.00
high	*2500*	*1*	***0.08***	***0.53***	***0.08***	***0.07***	***0.34***	***0.04***
		1.2	0.00	0.00	0.00	0.00	0.00	0.00
		1.5	0.01	0.13	0.07	0.04	0.22	0.12
		2	0.42	0.90	0.70	0.60	0.97	0.83
		2.5	0.87	0.99	0.93	0.96	1.00	0.98
		3	1.00	1.00	0.99	1.00	1.00	1.00
	5000	1	***0.06***	***0.49***	***0.13***	***0.05***	***0.27***	***0.05***
		1.2	0.00	0.00	0.00	0.00	0.00	0.00
		1.5	0.01	0.14	0.07	0.04	0.21	0.12
		2	0.36	0.90	0.70	0.57	0.96	0.83
		2.5	0.84	0.99	0.93	0.96	1.00	0.98
		3	0.99	0.99	0.99	1.00	1.00	1.00

Abbreviations: *ccr*, case-control ratio; p_d, baseline disease risk; p_e, exposure frequency; p_g, genotype carrier frequency; *OR_G-E,* odds ratio associated with G-E correlation; *N_G-E,* number of SNPs with population based G-E correlation (strength of G-E correlations); $OR_{G\times E}$, odds ratio associated with G×E interaction; *EHB-GE_{NN},* parametric empirical hierarchical Bayes approach for G×E interaction, based on normal-normal model; *EHB-GE_{CHI},* empirical hierarchical Bayes approach to G×E interaction, based on chi-distribution; *MUK-EB*, empirical Bayes shrinkage estimator;

Notes: $OR_{G\times E}$ *=1*corresponds to the null hypothesis;

N_G-E=0 corresponds to absence of G-E correlation;

Significance level per test is set to 5×10^{-5}, as 10,000 SNPs were simulated;

Figure 5.3 Evaluation of relative changes in power and type I error. The difference in power (on x-axis) and the difference in type I error (on y-axis) for EHB-GE_{NN} vs. EHB-GE_{CHI} (upper row) and for EHB-GE_{NN} vs. MUK-EB (lower row)

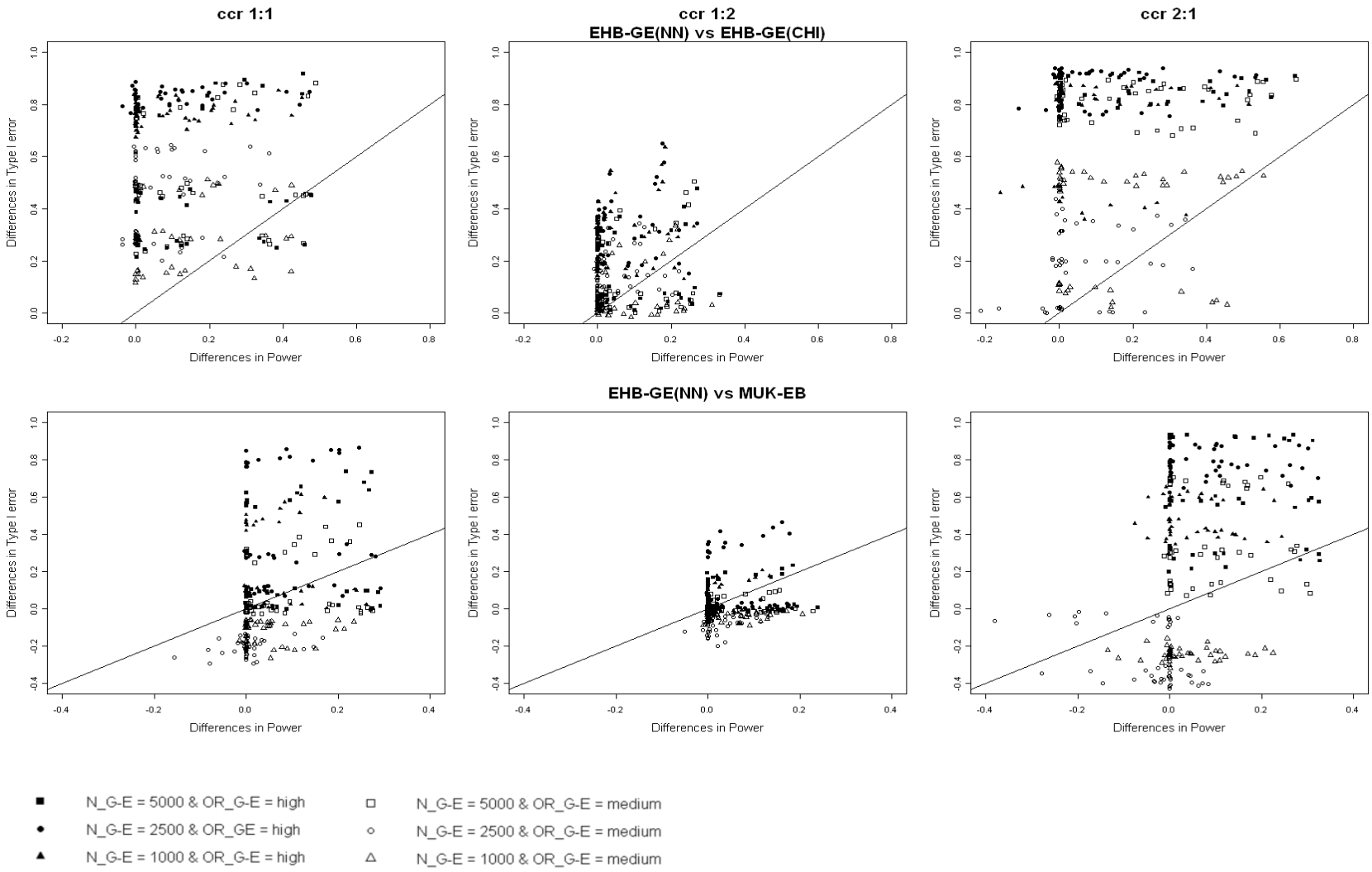


Figure 5.4 Rank power comparison to detect a G×E interaction in the top 100 SNPs between EHB-GE_{NN} and competing methods (CC, MUR, CO, MUK-EB, EHB-GE_{CHI}) for parameter combinations (OR_{GxE} = 1.2, 1.5, 2, 2.5, 3; p_g = 0.1, 0.3, 0.5; p_e = 0.1, 0.3, 0.5, and p_d = 0.05) given 1500 cases and 1500 control, and 1000 replicates.

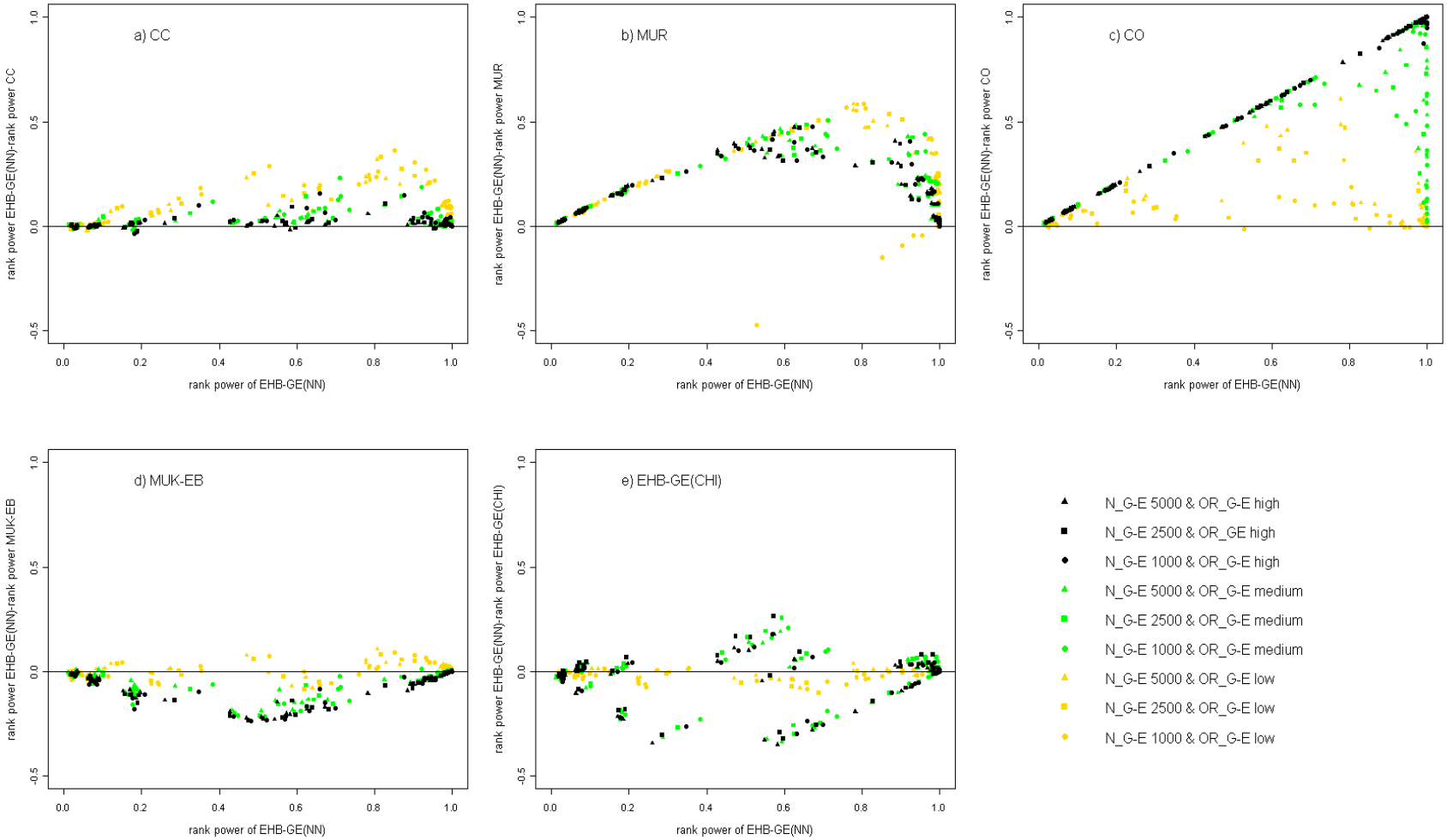


Figure 5.5 Rank power comparison to detect a G×E interaction in the top 100 SNPs between EHB-GE_{NN} and competing methods (CC, MUR, CO, MUK-EB, EHB-GE_{CHI}) for parameter combinations (OR_{GxE} = 1.2, 1.5, 2, 2.5, 3; p_g = 0.1, 0.3, 0.5; p_e = 0.1, 0.3, 0.5, and p_d = 0.05) given 1000 cases and 2000 control, and 1000 replicates.

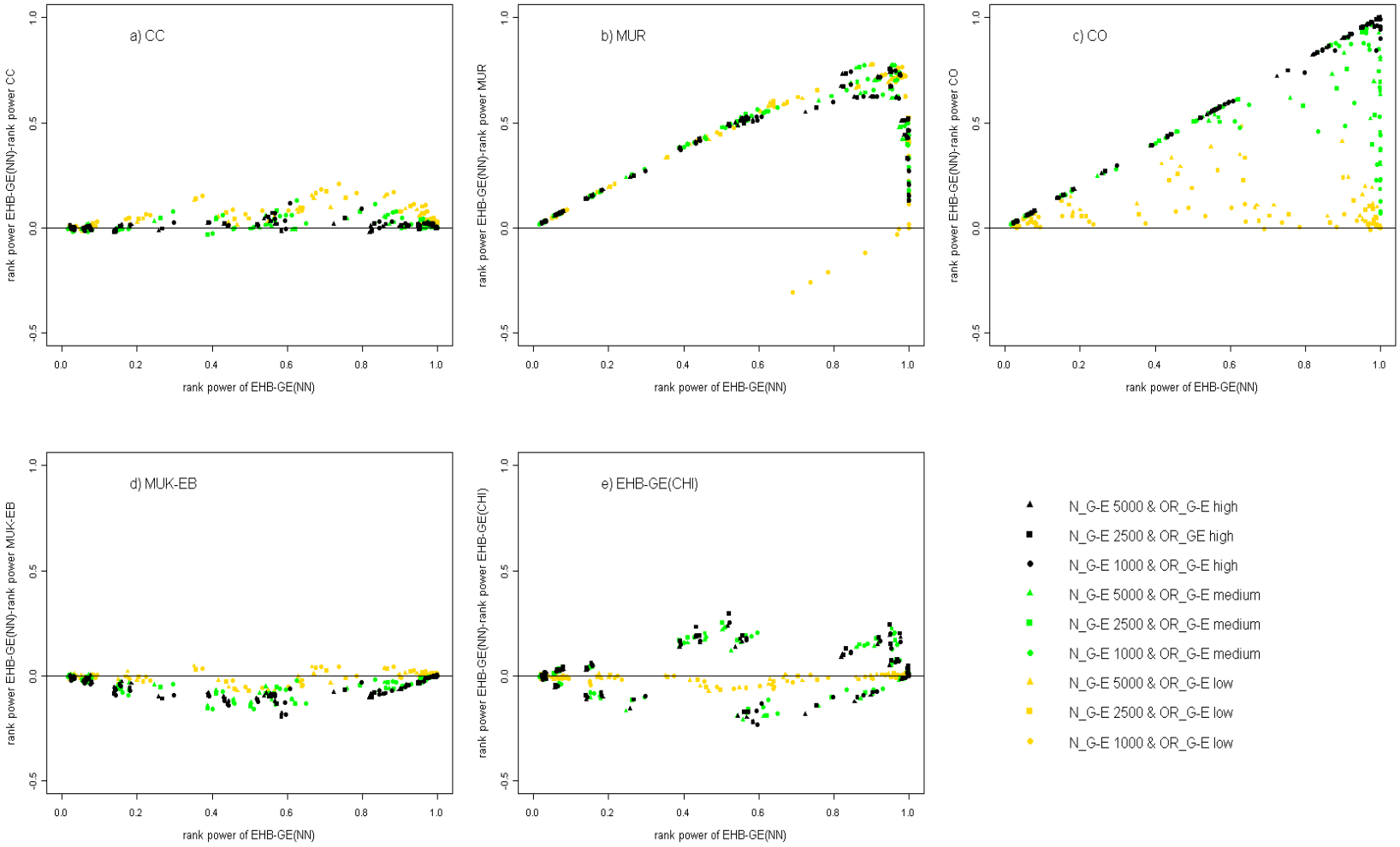


Figure 5.6 Rank power comparison to detect a G×E interaction in the top 100 SNPs between EHB-GE_{NN} and competing methods (CC, MUR, CO, MUK-EB, EHB-GE_{CHI}) for parameter combinations (OR_{GxE} = 1.2, 1.5, 2, 2.5, 3; p_g = 0.1, 0.3, 0.5; p_e = 0.1, 0.3, 0.5, and p_d = 0.05) given 2000 cases and 1000 control, and 1000 replicates.

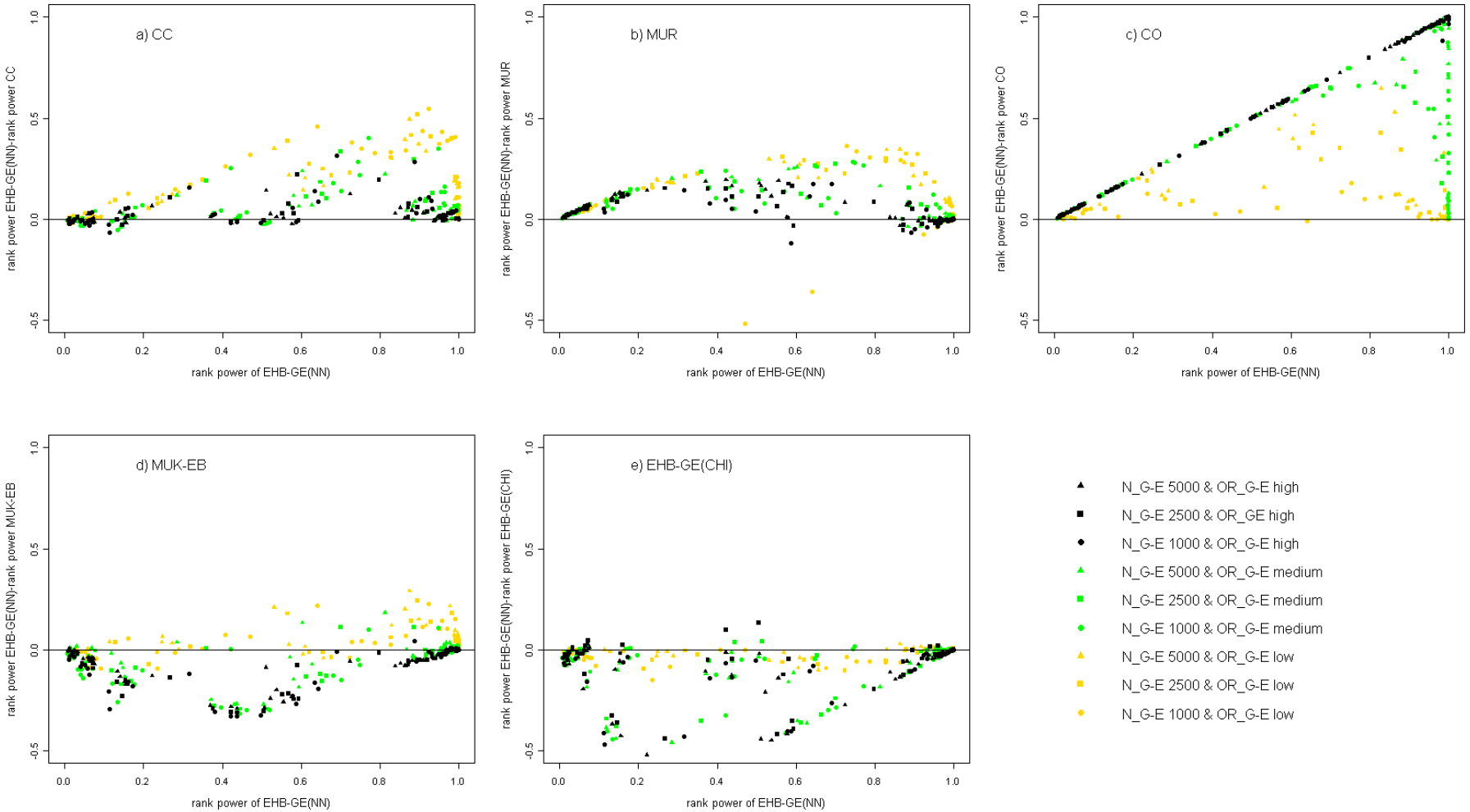


Figure 5.4-**Figure 5.6**, plots a) demonstrate that EHB-GE_{NN} has much greater rank power in almost all simulation scenarios compared to the case-control test. Comparing EHB-GE_{NN} versus MUR (**Figure 5.4**-**Figure 5.6**, plots b)), we conclude that in presence of a larger number of G-E correlations with high effect size. EHB-GE_{NN} reaches remarkably increased rank power. The rank power of MUR is higher compared to EHB-GE_{NN} when there are only G-E correlations with low effect size. From **Figure 5.4**-**Figure 5.6**, plots c) it can be seen that EHB-GE_{NN} outperforms the case-only test too. A clear triangular structure above the *x-axis* suggests that EHB-GE_{NN} has considerably larger rank power than CO in presence of medium to high G-E correlation signals. Irrespective of other parameters, the CO test has very low rank power in these situations. The EHB-GE_{NN} approach performed better, achieving greater rank power than MUK-EB (**Figure 5.4**-**Figure 5.6**, plots d)) in presence of G-E correlations with low effect size. Generally, EHB-GE_{NN} has lower rank power, thank MUK-EB; however, the loss in rank power is not dramatic. EHB-GE_{CHI} has a higher rank power than EHB-GE_{NN} (**Figure 5.4**-**Figure 5.6**, plots e)) for almost all simulated scenarios. The average loss in rank power for all situations over 1000 replicates is $\leq 5\%$ for EHB-GE_{NN}. It is important to notice that the rank power of MUK-EB and EHB-GE_{CHI} compared to EHB-GE_{NN} should be discussed in the context of the type I error inflation for each of the methods.

Generally, EHB-GE_{NN} can be applied for significance testing in GWAS to search for G×E signals without assuming gene-environment independence. It maintains adequate power and in this respect nearly always performs better, in terms of reaching higher rank power than CC or MUR tests, those methods requiring no assumption of genotype-environment independence. Based on the results from the simulation study, we recommend performing EHB-GE_{NN} to test for interaction when a large number of G-Es with moderate to high effect size is expected to occur in the study sample and preferably with a frequent exposure variable, so that the strata are large enough for the hyperparameter estimation.

5.5. Joint Tests for Genetic Marginal Effect and G×E Interaction Effects

Originally, a joint likelihood ratio test for the genetic main effect combined with the G×E interaction effect was introduced by Kraft and colleagues (Kraft, Yen et al. 2007). They compared power and sample size requirement of this joint test to all: the marginal case-control test, case-control and case-only tests for G×E interaction. It was revealed that even though for many penetrance models the joint test is not the most powerful, it is nearly optimal across all genetic risk models considered (Kraft, Yen et al. 2007). Dai and colleagues presented a novel approach to joint testing (Dai, Logsdon et al. 2012). They proposed identifying the involvement of a genetic factor in terms of its marginal association with the trait and/or involvement in gene-environment interaction. This allows them to use CC, CO and MUK-EB estimators to estimate the G×E interaction component (Dai, Logsdon et al. 2012). Dai's joint tests are more flexible than and at least as powerful as the likelihood test by Kraft et al. We present below three joint *2 degrees of freedom* tests (CC^J, CO^J, MUK-EB^J), proposed by Dai (Dai, Logsdon et al. 2012).

In accordance with Dai's proposal, let us consider the four logistic regression models presented in equations (5.9) – (5.12). Let *D* denote the binary disease outcome, *E* denote the exposure variable, and *G* denote genotype. Let Z be a vector of potential covariates.

$$logit(P(D=1|G))=\alpha_0+\alpha_G G_m+\alpha_Z^T Z^T \quad \textbf{(5.9)}$$

$$logit(P(D=1|G, E))=\beta_0+\beta_G G_m+\beta_E E+\beta_{CC} G_m\times E+\beta_Z Z^T \quad \textbf{(5.10)}$$

$$logit(P(E=1|G, D=1)) = \lambda_0 + \beta_{cases} G_m+\lambda_Z Z^T \quad \textbf{(5.11)}$$

$$logit(P(E=1|G, D=0))=\gamma_0+\beta_{controls} G_m+\gamma_Z Z^T \quad \textbf{(5.12)}$$

Equation (5.9) models the association between trait *D* and SNP effect *Gm*, therefore $\alpha_G = 0$ tests for the presence of the genetic marginal effect of SNP *m*. The hypothesis $\beta_{CC} = 0$ in equation (5.10) is a basis for the classic case-control test for G×E interaction in the presence of

G and E main effects. In order to combine two regression estimators $\widehat{\alpha_G}$ and $\widehat{\beta_{CC}}$ of two different models (5.9) and (5.10) into a single test statistic, these estimators should be asymptotically independent. As suggested in Dai et al. (2011), the independence follows, since for any two nested generalized linear models with a canonical link function, the parameter estimates of a smaller model are asymptotically independent of the estimates in a larger model (Dai, Logsdon et al. 2012). Because of the independence, Dai's tests statistics have a simple form of the sums of two squared *Z* scores. Under the two null hypotheses H_{01}: $\alpha_G=0$ and H_{02}: $\beta_{CC}=0$, or $\beta_{CO}=0$, or $\beta_{MUK\text{-}EB}=0$, all three test statistics follow a χ^2 distribution with 2 *degrees of freedom*.

The corollary to the theorem concludes the statement of asymptotic independence: We present the corollary as in (Dai, Kooperberg et al. 2012).

Corollary (Dai, Kooperberg et al. 2012): *Let Y be an outcome variable in a generalized linear model with a canonical link function g, and let X be the genetic variable, E the environmental variable and Z the additional covariates. Consider two nested generalized linear models*

$$g\{E(Y|E)\} = \beta_0 + \beta_1 X + \beta_2 Z$$

$$g\{E(Y|X,E)\} = \gamma_0 + \gamma_1 X + \gamma_2 E + \gamma_3 X \times E + \gamma_4 Z$$

Then the maximum likelihood estimators, $\widehat{\beta_1}$ *and* $\widehat{\gamma_3}$ *are asymptotically independent.*

In the logistic regression models framework, a simulation study performed by Dai and colleagues demonstrated almost zero empirical correlation when the sample size was at least a few hundred individuals (Dai, Kooperberg et al. 2012).

Based on the corollary above, one can conclude that $\hat{\alpha}_G$ and $\hat{\beta}_{CC}$ estimated on the basis of models (5.9) and (5.10) are independent and therefore can be pooled to construct joint statistics

distributed as *2 df* chi square (χ^2_{2df}) for joint testing of genetic marginal and G×E interaction by adding the appropriate terms as follows in equation (5.13)

$$T^J_{CC} = \frac{\widehat{a_G}^2}{\widehat{Var(a_G)}} + \frac{\widehat{\beta_{CC}}^2}{\widehat{Var(\beta_{CC})}} \quad \textbf{(5.13)}$$

The following joint test statistics for CO and MUK-EB can be obtained in the same way (Dai, Logsdon et al. 2012).

$$T^J_{CO} = \frac{\widehat{a_G}^2}{\widehat{Var(a_G)}} + \frac{\widehat{\beta_{CO}}^2}{\widehat{Var(\beta_{CO})}} \quad \textbf{(5.14)}$$

$$T^J_{MUK-EB} = \frac{\widehat{a_G}^2}{\widehat{Var(a_G)}} + \frac{\widehat{\beta_{MUK-EB}}^2}{\widehat{Var(\beta_{MUK-EB})}} \quad \textbf{(5.15)}$$

The estimators $\hat{\beta}_{MUK-EB}$ and $\widehat{Var}(\hat{\beta}_{MUK-EB})$ correspond to Mukherjee's et al. empirical Bayes shrinkage estimator for G×E interaction that can be derived based on models (5.10), (5.11), and (5.12) as shown in (Mukherjee and Chatterjee 2008) and summarized in Chapter 2.

5.6. Joint EHB-GE$_{NN}^{J}$ Test

Based on the same reasoning as above, we constructed a joint test for the EHB-GE$_{NN}$ approach. We showed in Chapter 2 that $\beta_{CC}=\beta_{cases}-\beta_{controls}$ (5.10-5.12). As proven in Chapter 4, the latter relationship also holds after separate covariate adjustment within cases and controls, when the covariate distribution is independent of the G×E interaction. Thus β_{CC} is a linear combination of β_{cases} and $\beta_{controls}$. Independence of $\hat{\alpha}_G$ and $\hat{\beta}_{cases}$ is stated in (Dai, Logsdon et al. 2012) and follows from the examination of covariance of respective estimating functions for the two estimators. Therefore, because of the linearity, in the relationship $\hat{\alpha}_G$ and $\hat{\beta}_{controls}$ are independent.

Since $\hat{\beta}_{EHB-GE_{NN}} = \widehat{\beta}_{cases} - posterior(\widehat{\beta}_{controls})$, independence of $\hat{\alpha}_G$ and $\hat{\beta}_{EHB-GE_{NN}}$ follows immediately. We propose the joint test statistic EHB-GE_{NN}^{J} for simultaneous testing of genetic marginal and G×E interaction effects as follows (5.16).

$$T^{J}_{EHB-GE_{NN}} = \frac{\widehat{a_G}^2}{\widehat{Var(a_G)}} + \frac{\widehat{\beta_{EHB-GE_{NN}}}^2}{\widehat{Var(\hat{\beta}_{EHB-GE_{NN}})}} \qquad \textbf{(5.16)}$$

where $\widehat{Var(\hat{\beta}_{EHB-GE_{NN}})} = \left(\sqrt{(\hat{\sigma}_m^{cases})^2 + Var\left(posterior(\beta_{controls})\right)}\right)^2$, (see Section 5.2).

In contrast to the CO test that was employed in Dai's 2 *df* test construction (Section 5.5), our EHB-GE_{NN} approach does not require the assumption of G-E independence (Section 5.2). Thus, this also holds for the EHB-GE_{NN}^{J} approach by the construction.

We did not perform an additional simulation study to compare the power and type I error of T^{J}_{CC}, T^{J}_{CO} T^{J}_{MUK-EB}, and $T^{J}_{EHB-GE_{NN}}$. All four tests have the same contributor for estimating the genetic marginal effect and differ only in the G×E interaction component; we therefore expect to see the same behavior as seen in the simulation study described in Section 5.4 in terms of the comparative performance of these tests. For the general comparison between main effect association tests and joint tests, we refer to the simulation studies previously conducted and published (Kraft, Yen et al. 2007, Dai, Logsdon et al. 2012, Vanderweele, Ko et al. 2013). Dai's joint tests were evaluated in terms of type I error, power, and robustness to G-E correlations (Dai, Logsdon et al. 2012, Vanderweele, Ko et al. 2013). These previous studies suggest that in the presence of main genetic effects only the classic case-control main effect test is more powerful than joint tests; however, the power loss of the joint test is only moderate to small. In the situation in which both a genetic main effect and a G×E interaction effect are present for a SNP, joint tests have substantially more power than pure main effect or G×E interaction tests.

Chapter 6

6. Applications to Lung Cancer Data from the ILCCO/TRICL Consortium

Lung cancer remains to be the leading cause of the cancer mortality in the world (Jemal, Bray et al. 2011). Every year, nearly 1.35 million newly diagnosed cases occur worldwide (Herbst, Heymach et al. 2008). A substantial proportion of individuals newly diagnosed with lung cancer dies within two years of diagnosis (Ferlay, Autier et al. 2007). Tobacco smoking is the major risk factor in lung cancer, accounting for nearly 85% of cases in men and 50% in women worldwide (Jemal, Bray et al. 2011). However, a fairly large proportion of patients develop the disease without any history of smoking (Bryant and Cerfolio 2007, Sun, Schiller et al. 2007, Couraud, Zalcman et al. 2012). Furthermore, there are many reports suggesting that a positive family history of lung cancer is an important risk factor, especially in the young (Coté, Liu et al. 2012, Krebsregister and (GEKID) 2013). It is widely accepted that lung cancer is a complex disease, attributed to the complex interaction of genetic and environmental factors (Osann 1991, Catelinois, Rogel et al. 2006, Chiu, Cheng et al. 2006, Kabir, Bennett et al. 2007, O'Reilly, McLaughlin et al. 2007, Brüske-Hohlfeld 2009).

Despite many studies devoted to identifying genetic factors that modify lung cancer risk, the majority of genetic markers and genes responsible for the development of lung cancer remain undiscovered. In recent years, quite a number of GWASs and meta-analyses were conducted. These have identified some risk variants for lung cancer. SNPs on chromosome *5p15* (Landi, Chatterjee et al. 2009, Truong, Hung et al. 2010, Brenner, Boffetta et al. 2012, Fehringer, Liu et al. 2012, Timofeeva, Hung et al. 2012, Li, Yin et al. 2013, Myneni, Chang et al. 2013), *6q23-25* (Bailey-Wilson, Amos et al. 2004, Hung, McKay et al. 2008, Amos, Pinney et al. 2010), *15q24-25* (Amos, Wu et al. 2008, Hung, McKay et al. 2008, Amos, Gorlov et al. 2010, Brenner,

Boffetta et al. 2012, Fehringer, Liu et al. 2012, Timofeeva, Hung et al. 2012) were discovered to be in association with lung cancer overall or with a specific histological subtype, such as adenocarcinoma, non-small-cell and small-cell lung cancer (NSCLC, SCLC) in European and Asian populations, some of them in African Americans. Several GWASs on smoking behavior have identified loci associated with the number of cigarettes per day as well as other measures of tobacco addiction/consumption (Lee, Jeon et al. 2005, Bierut, Madden et al. 2007, Thorgeirsson, Geller et al. 2008, Heller, Zielinski et al. 2010, Liu, Tozzi et al. 2010, Thorgeirsson, Gudbjartsson et al. 2010).

We had access to four lung cancer case-control GWASs from the International Lung Cancer Consortium/Transdisciplinary Research in Cancer of the Lung (ILCCO/TRICL) consortia. The GWASs are described below. In our investigation into genetic variants influencing the risk of lung cancer, we performed five statistical tests on each of the four GWASs, assuming a dominant mode of inheritance for all analyses. We applied EHB-GE_{NN} to investigate G×E interaction. We also applied joint CC^{J}, CO^{J}, and MUK-EB^{J} proposed by Dai and colleagues (Dai, Logsdon et al. 2012), all described in Chapter 5, to test simultaneously for the genetic marginal and gene-environment interaction effects. Moreover and in a similar way to those, we combined both estimators of the genetic main effect (G) and G×E interaction, later obtained by EHB-GE_{NN}, into a single joint test statistic EHB-GE_{NN}^{J} (see Chapter 5) and applied it to the respective GWAS datasets. Several authors have suggested that G×E interaction might help detect genetic variants potentially missed by standard tests for association of main effects (Kraft, Yen et al. 2007, Dai, Logsdon et al. 2012, Vanderweele, Ko et al. 2013). Specifically, some SNPs may exercise a small to moderate genetic main as well as a G×E interaction effect. Therefore, joint tests for the marginal association combined with the test for G×E interaction have been developed to gain additional power over tests of main effects only (Kraft, Yen et al.

2007, Dai, Logsdon et al. 2012). Thus, joint testing can help identify such signals in lung cancer.

6.1. ILCCO/TRICL GWAS Study Description

We analyzed four GWASs from the ILCCO/TRICL consortia. The German Lung Cancer GWAS (GLC) (Holle, Happich et al. 2005, Wichmann, Gieger et al. 2005, Sauter, Rosenberger et al. 2008), the Central Europe Lung Cancer GWAS (Central Europe IARC, CE-IARC) (Amos, Wu et al. 2008, Hung, McKay et al. 2008), the Toronto Lung Cancer GWAS (Samuel Lunenfeld Research Institute, SLRI) (Hung, McKay et al. 2008), and the Texas Lung Cancer GWAS (MD Anderson Cancer Center, MDACC) (Amos, Wu et al. 2008, Hung, Christiani et al. 2008, Wang, Broderick et al. 2008) were included in the analysis.

The **German Lung Cancer Study** (GLC, Bickeböller, Heinrich, Risch) is a population-based, case-control study comprising of 514 cases and 488 controls. It is a genome-wide study that includes cases diagnosed with lung cancer before the age of 51 and controls matched to them by sex and age. All subjects in the GLC Study were genotyped on the HumanHap 550K genome-wide SNP chip (Landi, Chatterjee et al. 2009). The GLC GWAS consists of three independent studies conducted in Germany, namely the Heidelberg Lung Cancer Study numbering 201 cases, the LUCY Study numbering 305 cases and the KORA Study with 488 controls (Sauter, Rosenberger et al. 2008).

The Heidelberg Lung Cancer Study is an ongoing hospital-based case-control genome-wide study initiated and conducted by the German Cancer Research Center (DKFZ, PD Risch). Initially started in 1997, more than 2,000 lung cancer cases were recruited in collaboration with the Thoraxklinik Heidelberg, 300 of which were cases with an age of onset $\leq$ 50 years. Data

on occupational exposure, tobacco smoking, and educational status, as well as family history of lung cancer for a subgroup of participants is available (Sauter, Rosenberger et al. 2008).

Lung Cancer in the Young (LUCY) is a multicenter study with 31 participating hospitals in Germany, organized and conducted by the Helmholtz Zentrum Munich in collaboration with the University Medical Center, Göttingen. Patients with histologically or cytologically confirmed primary lung cancer were recruited. The data comprise information on family history, smoking exposure, occupational exposure, and blood samples (Sauter, Rosenberger et al. 2008). Recruitment ended in 2011, with a total of 847 lung cancer cases and 5,524 family members being recruited.

Cooperative Health Research in the Augsburg Region (KORA) is a population-based, genome-wide study in the area of Augsburg in southern Germany, conducted by the Helmholtz Zentrum Munich. In total, 18,000 participants were recruited between 1984 and 2001. The data include information on multiple phenotypes, medical and laboratory data, as well as blood samples to uncover genetic information (Sauter, Rosenberger et al. 2008). KORA is a representative sample of Caucasians in Germany (Steffens, Lamina et al. 2006).

The **Central Europe Lung Cancer Study** of the IARC (CE-IARC, Brennan) is a hospital-based case-control genome-wide study conducted in 15 centers in 6 central and eastern European countries (Czech Republic, Hungary, Poland, Romania, Russia, Slovakia) and in Liverpool (United Kingdom) between 1998 and 2002. The study collected lifestyle risk factors, occupational, medical, and family history information on a total of 2633 newly diagnosed lung cancer cases and 2884 controls frequency matched by age, sex, geographical area, and period of recruitment (Scelo, Constantinescu et al. 2004). All study individuals were genotyped on Illumina HumanHap 300K platforms (Hung, McKay et al. 2008).

The **Toronto Lung Cancer Study** (SLRI, Hung), is a hospital-based, genome-wide, case-control study that was conducted by the University of Toronto and the Samuel Lunenfeld

Research Institute in the greater Toronto area between 1997 and 2002. The study contained genetic, lifestyle risk factors, occupational, medical, and family history information on 332 lung cancer patients and 505 controls of European origin (Hung, McKay et al. 2008).

Table 6.1 demonstrates major characteristics of the GWASs and individuals, as well as information on the respective genotyping platform.

Table 6.1 Characteristics of the four lung cancer GWASs, QC is quality control

	GLC		**CE-IARC**		**SLRI**		**MDACC**	
Data collection area	Germany		Central Europe: Czech Republic, Hungary, Poland, Romania, Russia, Slovakia		greater Toronto area, Canada		Houston, Texas, USA	
Origin of control	Population-based		Hospital-based		Hospital-based		Hospital-based	
Frequency matching factors	Ages, sex, location		Age, sex, location		Age, sex, ethnic origin		Age, sex, ethnic origin, smoking habits	
Genotyping	HumanHap 550K		HumanHap 300K		HumanHap 300K		HumanHap 300K	
Cases/Controls genotyped	514/488		1989/2625		332/505		1154/1137	
Cases/Controls after QC	467/468		1901/2503		331/499		1150/1134	
# SNPs after QC	529,730		312,452		310,045		314,072	
	cases	**controls**	**cases**	**controls**	**cases**	**controls**	**cases**	**controls**
Gender (Male/Female)	286/181	237/231	1493/408	1821/682	159/172	190/309	655/495	644/490
Age (years)								
< 45	169	112	246	415	41	233	176	120
45-49	239	283						
50-54	50	73	272	378	32	62	109	107
55-59	-	-	329	394	28	46	158	210
60-64	-	-	386	435	46	32	186	278
65-69	-	-	353	430	62	35	202	236
70-74	-	-	286	368	69	41	184	134
≥ 75	-	-	29	83	52	49	135	49
Missing	9	0	-	-	1	1	-	-
Smoking status								
Never	35	214	144	884	91	215	-	-
Former	45	121	373	656	95	143	601	655
Current	377	133	1380	954	90	90	549	479
Any	10	-	4	7	55	46	-	-
Missing	-	-	-	2	-	5	-	-
Smoking quantity (for smokers in pack/years)								
Moderate (≤ 20 pack/years)	83	152	248	619	38	122	160	230
Heavy (> 20 pack/years)	328	101	1504	988	145	106	990	904
Missing	21	1	5	10	58	51	-	-

6.2. GWAS Data Quality Control

Quality control (QC) of the data in the genome-wide context is essential. We performed standard, systematic quality control on all four ILCCO/TRICL datasets prior to the association and interaction analyses and after QC carried out by the data providers. QC was mainly conducted in PLINK (Purcell, Neale et al. 2007) and EIGENSOFT (Price, Zaitlen et al. 2010). Comparable quality criteria were applied for each of the four GWAS.

QC was performed on the subject level as well as on the SNP level (see **Table 6.2**). Standard filters on the subject level included checks for genotype call rate, cryptic relatedness as measured by identity by state (IBS) between pairs of subjects (if the IBS is too high, subjects might be closely related and should be excluded from further analysis), ancestral origin that can be determined for example by principal component (PC) (study populations should be as homogeneous as possible and subjects with a different ethnic background should be excluded from analysis), excessive number of heterozygous individuals (if the heterozygosity for a subject is too high, the DNA is suspected of being contaminated, low heterozygosity suggests that hybridization might have failed). Standard filters on the SNP level include checks for SNP call rate, minor allele frequency (MAF) (many genotype-calling algorithms tend to perform poorly for SNPs with low MAF, and the power of a study is low in detecting associations for SNPs with a low MAF, usually lower than 0.01), Hardy-Weinberg equilibrium (HWE) (SNPs are excluded if significantly more or fewer individuals are heterozygous at a SNP than expected, HWE is performed on unrelated control subjects with relatively homogeneous ancestry).

For the GLC study, principal component analyses on a subset of around 100,000 independent markers was performed to access the population structure and identify ethnic outliers. As an outcome, we obtained 20 principal components (PC) with corresponding p-values < 0.05. Seventeen of them had p-values less than 10^{-7}. To remove population outliers from the analyses

and restrict the sample only to individuals of Caucasian origin, we performed an iterative procedure integrated in EIGENSOFT to remove outliers automatically. A similar procedure was performed for the SLRI study. For the CE-IARC study, STRUCTURE software was used to identify population outliers as individuals with an ancestry probability rate of being Caucasians < 80%. The MD Anderson Cancer Center (MDACC) used a procedure in PLINK (absolute value of the nearest neighbor > 4) to identify outliers. **Table 6.2** summarizes the quality control criteria that we used to preprocess our data.

Table 6.2 Filters for standard quality control of ILCCO/TRICL GWASs

Level	Filter	Standard value for filter
Subject	Call rate	≥ 90%
	Cryptic relatedness	proportion of IBD < 0.20
	Sex mismatch	female F < 0.2 and male F > 0.8
	Heterozygosity	mean F +/- 6 SD F, over all samples
	Ethnic origin	Caucasian ancestry PLINK nearest neighbor Z score| < 4
SNP	Call rate	≥ 95%
	MAF	≥ 1%
	HWE	p-values HWE in controls ≥ 10^{-7}

F, estimate for homozygosity, Wright's inbreeding coefficient; SD, standard deviation; Recommended values to assign the sex are <0.2 for females and >0.8 for males;

6.3. Covariates

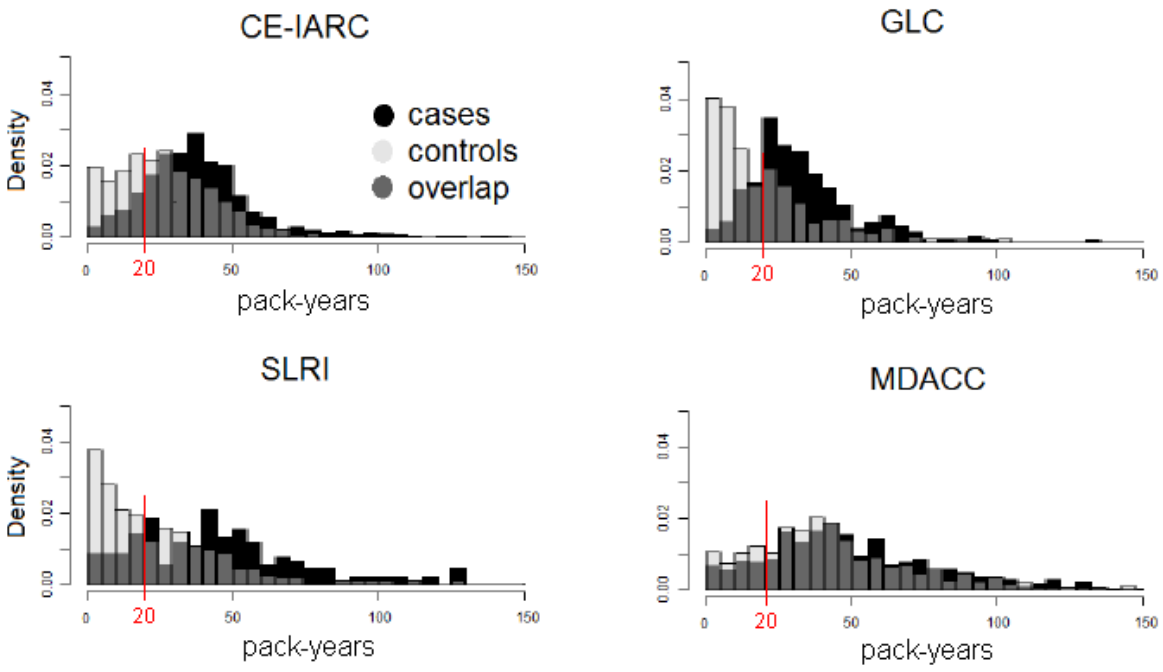
All four GWASs collected covariate information on individuals, such as sex, age, smoking status, smoking amount, and ethnicity (See **Table 6.1**).

An individual's sex was coded as 1 for female and 0 for male. Age was originally presented in years, later being coded as a categorical variable grouping age into blocks of five years, after an initial group encompassing everyone under the age of 45 years (**Table 6.1**). As described in Section 6.1, the GLC Study only contains subjects younger than 51 years. As a result, the GLC study only has three age groups.

Tobacco smoking as the major risk factor in lung cancer development was considered as the major environmental risk factor possibly interacting with the genetic factors of the individuals to influence occurrence of the disease. In all studies, smoking information was collected as pack-years per individual, defined as the number of cigarette packs smoked by the subject in one day multiplied by the duration of smoking in years (Amos, Wu et al. 2008). Hence pack-years combines the amount smoked and the duration. **Figure 6.1** demonstrates the distribution of pack-years in the four studies within cases and controls. Clearly, cases tend to have consumed more pack-years than controls. For our analysis, the smoking status of each subject was coded in two different ways: NE describes *ever=1* and *never=0* smokers and MH denotes *heavy=1* and *moderate=0* smokers. We defined n*ever* smokers as those individuals having consumed no more than 100 cigarettes over their life span and *ever* smokers as those having consumed more. Generally, there were few never smokers in the GWASs. The MDACC study considered only ever smokers. For the MH coding, we defined *moderate* smokers as those with a consumption ≤ 20 pack-years and *heavy* smokers with a consumption > 20 pack-years. Never smokers were excluded from the consideration in this model to ensure comparison across the GWASs, as the MDACC study did not include never smokers.

To account for possible population stratification, PCs should be included as covariates in the analysis. For the GLC study, the first four significant PCs, for CE-IARC the first six significant PCs, representing the six countries of the data collection, for MDACC the first two PCs and for SLRI the first three PCs were included in the analyses.

Figure 6.1 Distribution of pack-years in each GWAS within cases and within controls



6.4. Data Analysis Strategies

To discover signals that may influence the risk of lung cancer we performed five statistical tests, the new EHB-GE$_{NN}$ approach and four joint tests as described in Chapter 5, on the GWASs described in Section 6.2.

After the quality control procedures and prior to the interaction analysis, we estimated G-E correlation effects within cases and within controls and their corresponding standard errors by the appropriate PLINK functions (Purcell, Neale et al. 2007). We also estimated the genetic main effects. We assumed a dominant mode of inheritance for all analyses, which were performed for both environment models never vs. ever and moderate vs. heavy (Section 6.4). For the joint testing, we included sex, age, and principal components as covariates and additionally controlled for smoking as main effect.

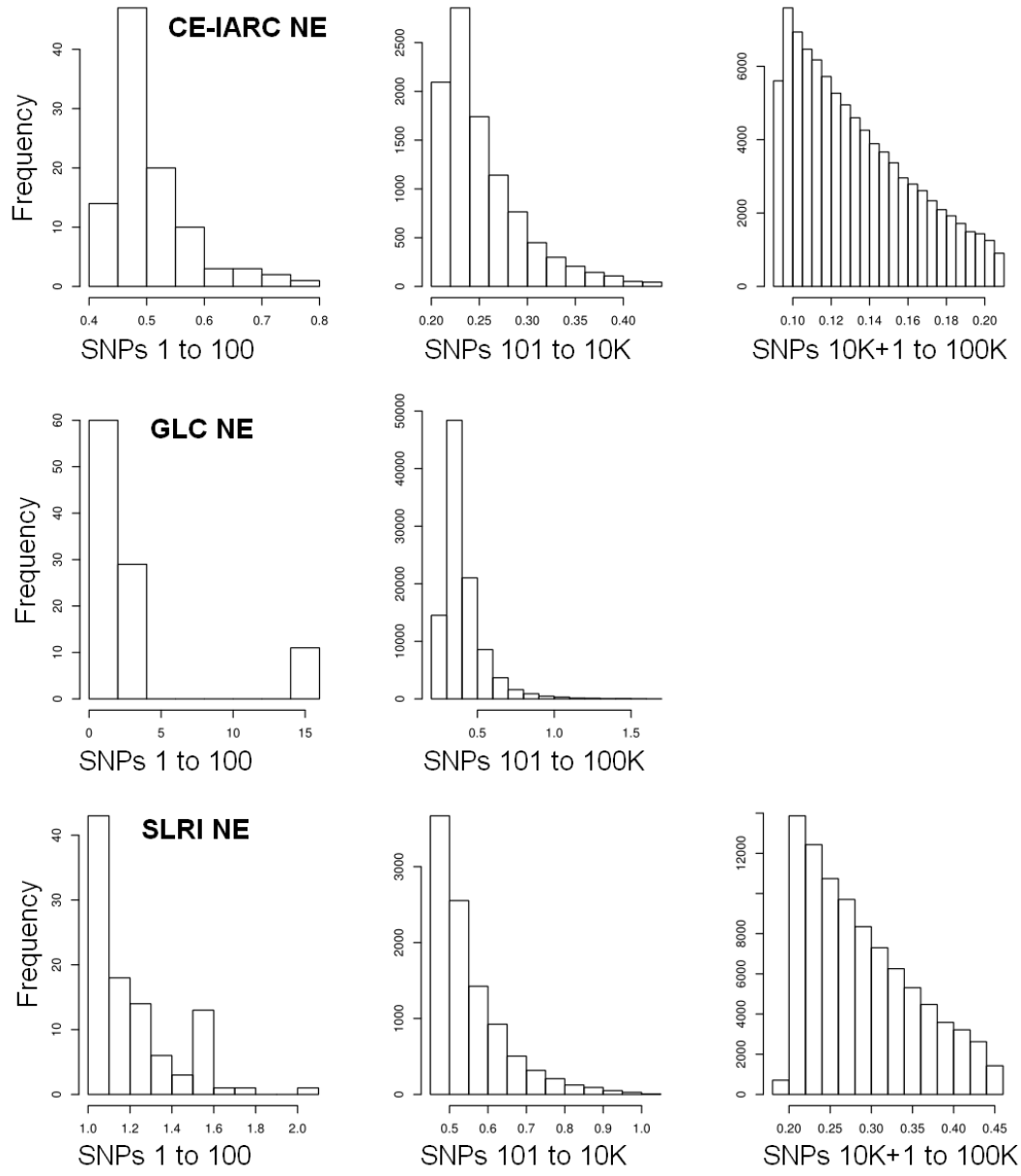
To ensure that the application of EHB-GE$_{NN}$ on the lung cancer GWASs is appropriate, we first evaluated the frequency of smoking exposure for both NE and MH coding in each of the four

GWASs. The frequency of smokers ranges from 54% to 65% with the frequency of heavy smokers ranging from 40% to 63%. Taking into account the sample sizes of the GWASs, we concluded that the exposed sub-strata are big enough to estimate the necessary parameters. We then looked at the distribution of the G-E correlation signals in controls for each study. There was evidence for the presence of a relatively large number of G-E correlation signals with medium and sometimes high effect size in all studies. This suggests that the application of EHB-GE_{NN} on these data is appropriate and can be advantageous. **Figure 6.2** and **Figure 6.3** present histograms of the beta coefficients estimating G-E correlation effects in controls for all GWASs studies. We displayed the beta coefficients estimating G-E correlation in controls giving the ordered absolute values for the largest 100,000 coefficients. For the GLC study with HumanHap 550K, 529,730 SNPs passed quality control thresholds, here approximately 19% of the data are shown. For the CE-IARC, MDACC, and SLRI studies with HumanHap 300K, approximately 30% of the data are demonstrated. We split the data into two to three histograms for visualization purposes, owing to a large difference in the scales of G-E correlation effects.

First, we applied the EHB-GE_{NN} approach to test for G×E interaction as described in Chapter 5. To construct the EHB-GE_{NN} test statistics posterior estimates of G-E correlation effects were derived based on their prior estimates obtained with PLINK. Then the $Z^{EHB\text{-}GE_{NN}}$ statistic was constructed per SNP to test for significance. Results of this analysis are presented in Section 6.7.

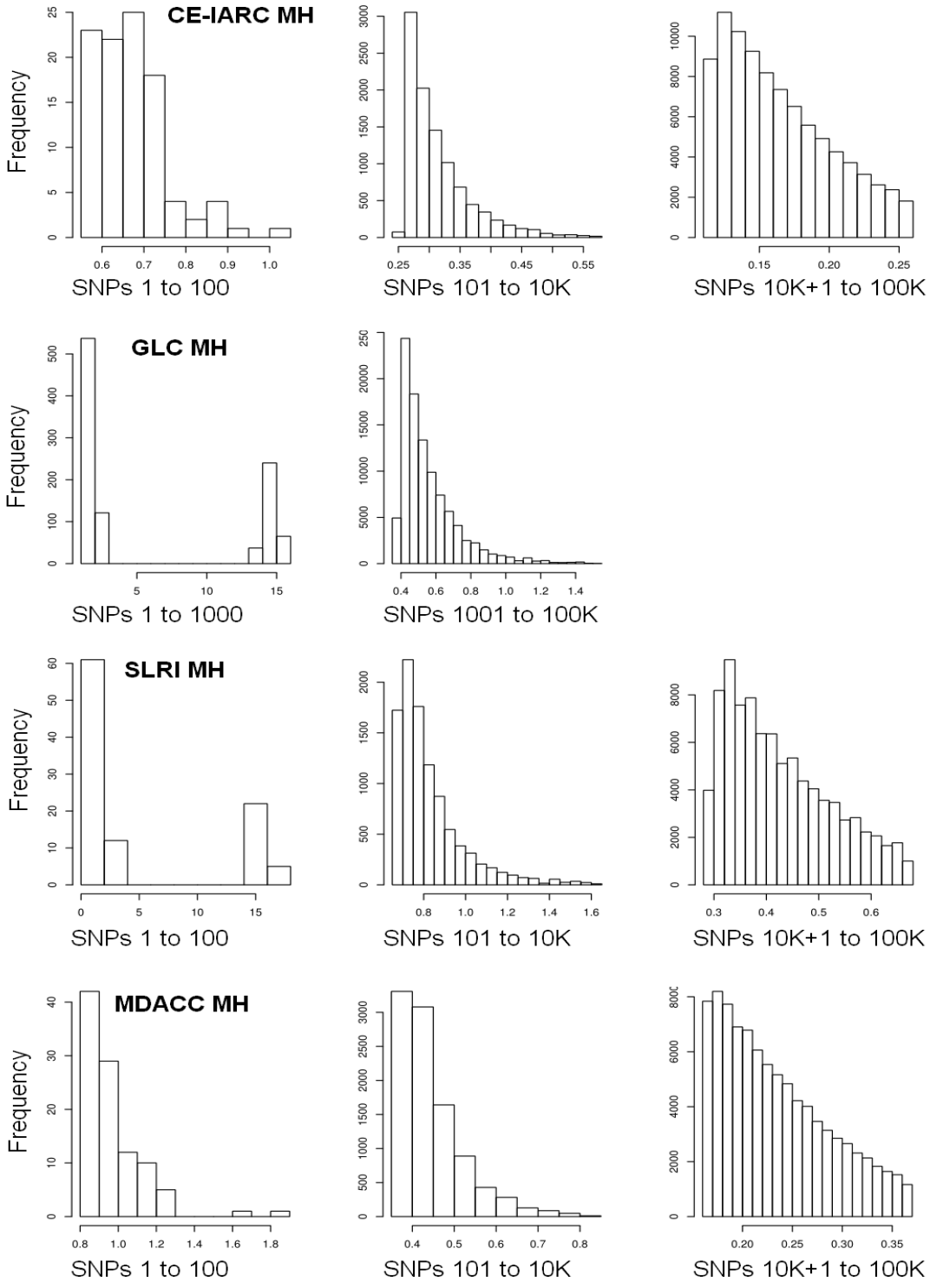
Subsequently, we applied the *2 df* tests proposed by Dai (Dai, Logsdon et al. 2012) and our EHB-GE_{NN}^{J} to test simultaneously for the genetic main and gene-environment interaction effects as described in Chapter 5. Results of the joint test are presented in Section 6.8. **Table 6.3** summarizes the tests performed.

Figure 6.2 Frequency histograms of the beta coefficients estimating G-E correlation effects in controls for each GWAS for never vs. ever smokers. Shown are the 100,000 largest coefficients in absolute value.



NE = never versus ever smokers coding;

Figure 6.3 Frequency histograms of the beta coefficients estimating G-E correlation effects in controls for each GWAS for moderate vs. heavy smokers. Shown are the 100,000 largest coefficients in absolute value.



MH = moderate versus heavy smokers coding;

Table 6.3 Summary of methods applied to ILLCO/TRICL GWASs.

Method	Test Statistics
EHB-GE$_{NN}$	$Z^{EHB-GE_{NN}} = \frac{\hat{\beta}^{cases} - (\beta^{controls})^*}{\sqrt{(\hat{\sigma}^{cases})^2 + Var\,((\beta^{controls})^*)}}$
EHB-GE$_{NN}^J$	$T^J_{EHB-GE_{NN}} = \frac{\widehat{\beta_G}^2}{\widehat{Var}(\widehat{\beta_G})} + \frac{\widehat{\beta_{EHB-GE_{NN}}}^2}{\widehat{Var}(\widehat{\beta_{EHB-GE_{NN}}})}$
CC^J (Dai, Logsdon et al. 2012)	$T^J_{CC} = \frac{\widehat{\beta_G}^2}{\widehat{Var}(\widehat{\beta_G})} + \frac{\widehat{\beta_{CC}}^2}{\widehat{Var}(\widehat{\beta_{CC}})}$
CO^J (Dai, Logsdon et al. 2012)	$T^J_{CO} = \frac{\widehat{\beta_G}^2}{\widehat{Var}(\widehat{\beta_G})} + \frac{\widehat{\beta_{CO}}^2}{\widehat{Var}(\widehat{\beta_{CO}})}$
MUK-EB^J (Dai, Logsdon et al. 2012)	$T^J_{MUK-EB} = \frac{\widehat{\beta_G}^2}{\widehat{Var}(\widehat{\beta_G})} + \frac{\widehat{\beta_{MUK-EB}}^2}{\widehat{Var}(\widehat{\beta_{MUK-EB}})}$

Abbreviations: *EHB-GE$_{NN}$*, parametric empirical hierarchical Bayes approach for G×E interaction; *EHB-GE$_{NN}^J$/ CC^J/ CO^J/ MUK-EB^J*, joint method for genetic main (*G*) and G×E interaction effect based on original G×E test indicated; *CC*, classical case-control interaction estimator; *CO*, case-only interaction estimator; *MUK-EB*, empirical Bayes shrinkage estimator;

6.5. Review and Replication of Results of Genetic Main Effect Analysis

We reviewed the literature, investigating SNPs' genetic main effect on the risk of developing lung cancer. Here, we first summarize some previously found results, i.e. significant association signals for lung cancer in Caucasian, Asian, and African American populations. We performed an extensive search of the PubMed database for articles concerning GWAS involving lung cancer.

A number of genome regions were identified. SNPs on chromosome *15q25.1* (Amos, Wu et al. 2008, Hung, McKay et al. 2008, Thorgeirsson, Geller et al. 2008), on chromosome *5p13.3* (McKay, Hung et al. 2008, Wang, Broderick et al. 2008, Landi, Chatterjee et al. 2009), on chromosome *6p21.33* (Wang, Broderick et al. 2008), on *12p13.3* (Shi, Chatterjee et al. 2012), and on chromosome *9p21.3* (Timofeeva, Hung et al. 2012). The most famous signal for the

association with lung cancer consists of two SNPs on chromosome *15q24-25*, namely *rs1051730* mapped to the *CHRNA3* gene and *rs8034191,* mapped to the *AGPHD1* gene. Originally this signal was reported by three independent research groups in different datasets on the same day. Our search identified 40 potentially relevant articles for this association. Amos and colleagues (Amos, Wu et al. 2008) reported the association of these two variants and lung cancer based on the case-control sample of 1,154 current and former (ever) smoking cases of European ancestry and 1,137 frequency-matched, ever-smoking controls from Houston, Texas. They also replicated the findings on an additional cohort of 711 cases and 632 controls from Texas and 2,013 cases and 3,062 controls from the UK. They reported an odds ratio of 1.32 and p-value ($p < 10^{-17}$) of the combined analysis for both SNPs (Amos, Wu et al. 2008). Many other studies replicated this signal in case-control and meta-analyses with p-values significant for association in Caucasians (Hung, McKay et al. 2008, Thorgeirsson, Geller et al. 2008, Broderick, Wang et al. 2009, Lips, Gaborieau et al. 2010, Truong, Hung et al. 2010, Fehringer, Liu et al. 2012, Timofeeva, Hung et al. 2012). The same two SNPs were also identified as influencing the risk of lung cancer in Asians (Truong, Hung et al. 2010, Gu, Dong et al. 2012) and African Americans (Amos, Pinney et al. 2010). Additionally, a single SNP, *rs16969968* on chromosome *15q* mapped to the *CHRNA5* gene, was identified as being associated with lung cancer risk in Europeans (Lips, Gaborieau et al. 2010) and African Americans (Walsh, Amos et al. 2012).

Another interesting signal for the association with lung cancer was formed by SNPs on chromosome *5p15*, namely *rs2736100*, which belongs to the *TERT* gene and *rs402710*, located on the *CLPTM1L* gene. These two SNPs were replicated in many large meta-analytic GWASs, including over 10,000 individuals and as many as 21 different GWASs to confirm signals (Wang, Broderick et al. 2008, Truong, Hung et al. 2010, Timofeeva, Hung et al. 2012). For example, Timofeeva and colleagues reported the association of *rs2736100* (OR = 1.14, p =

5.00×10^{-8}) and the association of *rs402710* (OR = 0.87, p = 1.70×10^{-7}) with the risk of lung cancer based on a meta-analysis of 14,900 cases and 29,485 controls of Caucasian origin (Timofeeva, Hung et al. 2012). A candidate SNP meta-analysis study for the variant *rs2736100* confirmed the association of this marker with lung cancer risk. The SNP *rs2736100* was associated with the risk of lung cancer in a dominant model (OR = 1.14, 95% CI: 1.01-1.28; p = 0.03) based on 14,492 subjects (Wang, Zhang et al. 2013). A few meta-analyses reported the association of *rs2736100* with the risk of adenocarcinoma (OR = 1.23, p = 3.02×10^{-7}) e.g. (Landi, Chatterjee et al. 2009). These signals were also confirmed in the Chinese (Hsiung, Lan et al. 2010, Li, Yin et al. 2013).

The third interesting signal comprises markers on chromosome *6p21*. Hung and colleagues (Hung, McKay et al. 2008) reported that a signal of ten SNPs clustered in a segment of approximately one mega-base on chromosome *6p* with *rs432479* was the strongest in their data (Hung, McKay et al. 2008). However Hung et al. (Hung, McKay et al. 2008) mentioned that this association is not confirmed and needs to be studied further and replicated. In the same year, Wang et al. (Wang, Broderick et al. 2008) published a study describing the association of another SNP, *rs3117582*, mapped to *BAT3-MSH5*, with p = 4.97×10^{-10} in the same chromosomal region, based on their pooled analyses of 5,095 cases, and 5,200 controls.

Many other SNPs across the genome have been reported to influence the risk of lung cancer generally or to be associated with the specific histology, such as NSCLC, SCLC, or adenocarcinoma risk. However, not all of these findings have been replicated in other studies or populations; therefore we only briefly mention some of them. For example, Timofeeva et al. mentioned that SNPs on *12p13* demonstrate an association with lung cancer. Furthermore, it was shown in the same study that the *9p21.3* variation is a determinant of squamous cell lung cancer risk (Timofeeva, Hung et al. 2012) in Caucasians. Novel SNPs on chromosome *15q*,

rs2036534, *rs667282*, *rs12910984*, and *rs6495309* were reported to be in association with lung cancer risk in the Chinese, but this was not confirmed in Caucasians (Wu, Hu et al. 2009).

Since not all of the published studies controlled for smoking in their analyses, which is important from our point of view, we obtained the main effect analysis results for each of four GWASs from the research groups of CE-IARC, GLC (our group), SLRI, and MDACC, controlled for sex, age, PC, and main effect of smoking. None of the SNPs passed the genome-wide significant level after including smoking as covariate in any GWAS.

In the CE-IARC Study, two SNPs on chromosome *15q25 rs8034191* and *rs1051730* had p-values $< 10^{-5}$. These two SNPs were reported to be in association with lung cancer in the CE-IARC GWAS by Hung and colleagues (Hung, McKay et al. 2008). They reported corresponding SNP p-values, 8.8×10^{-10} (*rs8034191*) and 5.4×10^{-9} (*rs1051730*) in the model not adjusted for smoking. Another two markers had p-values lower than 10^{-5}, *rs10516367* (*KCNIP4* gene) and *rs1407503* (*GALNT12* gene).

For the GLC Study, there were some SNPs with corresponding p-values $\leq 10^{-5}$ spread along the genome, however no clear signal for the presence of a genetic main effect for any of the SNPs. Among those, three SNPs belonged to genes, namely *rs2866908* (*DKK2*), *rs2916508* (*CTNNA2*), *rs9643575* (*TRIM55*).

For the SLRI Study, six markers had p-values $\leq 10^{-5}$, however only two among them were mapped to genes. These two SNPs are *rs12112953* (*ADCY1*), *rs266508* (*RGSL1*).

In the MDACC Study, no SNPs in coding regions or the close neighborhoods of genes with p-values $\leq 10^{-5}$ were identified after adjusting for the main effect of smoking.

6.6. Results for G×E Interaction Analysis

Generally, we do not discuss or present the results of G×E interaction analysis on those SNPs located outside of known genes or further than ±500Kb from protein coding regions. For such signals it is hard to argue for any association with the specific trait. We therefore omitted them from the discussion. For all the studies and both exposure models (NE and MH) in the analysis applying EHB-GE_{NN} to test for G×E interaction, we only describe SNPs with p-values of interaction $p \leq 10^{-4}$ here.

In the CE-IARC study, we did not identify any SNPs with genome-wide significant interaction effect for either exposure models, NE or MH. In the NE model, there was only one SNP located in the coding region mapped to the *C1orf21* gene on chromosome *1* with $p=7.02\times10^{-6}$. Other SNPs investigated for this study and NE coding with p-values $\leq 10^{-4}$ were located in non-coding regions and therefore are not discussed. For the MH coding, two SNPs located on chromosome *14*, *rs2302591* and *rs175891*, mapping to the *TTLL5* gene, form an interesting signal. This tumor suppressor candidate gene encodes a member of the tubulin tyrosine ligase-like protein family. This protein may function as a co-regulator of glucocorticoid-receptor-mediated gene induction and repression. This protein may also function as an alpha tubulin polyglutamylase (Uhlen, Oksvold et al. 2010). Another signal here consists of three SNPs on chromosome *16*, *rs200528 ($p = 3.05\times10^{-5}$)*, *rs3803716 ($p = 2.94\times10^{-5}$)*, and *rs2112783 ($p = 2.79\times10^{-5}$)* that belong to the *TNRC6A* gene.

In the GLC study with NE exposure coding, two SNPs, *rs13244987* and *rs13438768*, reached genome-wide significance with p-values of 3.33×10^{-8} and 9.12×10^{-8}, respectively. Both markers belong to the human protein coding locus *LOC645249*, known to be expressed differently in tumor and normal cell tissues. Marker *rs7308621* on chromosome *12* in the *RERG* gene is worth mentioning, as this gene participates in tumor formation. Another interesting signal for this analysis comprises three SNPs on chromosome *13*, mapping to the *ENOX1* gene,

namely *rs7982922* ($p = 7.53\times10^{-7}$), *rs10492572* ($p = 8.20\times10^{-6}$), and *rs10492573* ($p = 5.63\times10^{-6}$). For the GLC, MH coding, two SNPs in the *ARHGEF38* gene with p-values < 10^{-5} appeared. Another signal for this analysis references three SNPs on chromosome *9* located in the *TRPM3* gene. These SNPs are *rs1421156* with $p = 1.72\times10^{-6}$, *rs656875* with $p = 1.30\times10^{-6}$, and *rs672801* with $p = 4.43\times10^{-6}$.

In the SLRI study and NE exposure model there are two signals of interest. The first signal consists of three SNPs (*rs1337862*, *rs1337863*, *rs945949*) with p-values of the order of 10^{-4} on chromosome *6*, mapping to the *NKAIN2* gene. It is known that the chromosomal translocation involving this gene is a cause of lymphoma. The second suggestion comprises five SNPs (*rs12956176*, *rs4486983*, *rs9646509*, *rs1880113*, *rs1403762*) with p-values of the order of 10^{-5} on chromosome *18* in the *KLHL14* gene. For the MH coding, a single SNP is an interesting signal; rs6872156 with $p = 8.31\times10^{-6}$ on chromosome *5* in the *DUSP1* gene. The role of this gene is increasingly recognized in tumor biology (Moncho-Amor, Ibanez de Caceres et al. 2011). Furthermore, two SNPs in *ADAMTSL1* and two SNPs in the closure of the *WWOX* gene were identified.

In the MDACC Study, MH analysis, there was only one SNP; *rs9323666* ($p = 3.79\times10^{-5}$) located on chromosome *14* in the *NRXN3* gene. The rest of the SNPs with relatively small p-values in this analysis were spread along the genome and were not located in any genes or their surroundings. **Table 6.4** summarizes our findings and gives further description of gene characteristics. **Figure 6.3** portrays Manhattan plots for GWASs for the G×E interaction analysis based on EHB-GE_{NN}.

Table 6.4 SNPs discovered by EHB-GE_{NN} in G×E Interaction Analysis of the ILCCO/TRICL GWASs

GWAS	E	SNP	p $Z^{EHB\text{-}GE}{}_{NN}$	CHR	Mapping*	characteristics**
CE-IARC	**NE**	rs2779286	7.02×10^{-6}	1	*C1orf21*	Human protein coding gene
		rs1455701	4.31×10^{-5}	1	+/- 500kb of the *ERRFI1*	Rare mutations in *MIG-6* have been identified in human lung cancer(Zhang 2008).
		rs7620618	2.12×10^{-6}	3	*±GOLGA4*	Human protein coding gene, postulated to play a role in Rab6-regulated membrane-tethering in the Golgi apparatus (Meyer, Brieger et al. 2009).
		rs4563628	2.42×10^{-4}	5	+/- 500kb of the *TAG*	tumor antigen gene, miscellaneous RNA
	MH	rs2302591 rs175891	4.35×10^{-6} 7.85×10^{-6}	14	*TTLL5*	This protein may function as a coregulator of glucocorticoid receptor mediated gene induction and repression (Uhlen, Oksvold et al. 2010). Tumor-suppressor candidate (Liang, Wang et al. 2005).
		rs1126289	4.31×10^{-5}	16	*PRKCB*	Protein kinase C (PKC) family members are known to be involved in diverse cellular signaling pathways. PKC family members also serve as major receptors for phorbol esters, a class of tumor promoters (Uhlen, Oksvold et al. 2010).
		rs200528 rs3803716 rs2112783	3.05×10^{-5} 2.94×10^{-5} 2.79×10^{-5}	16	*TNRC6A*	Expression *TNRC6A* delocalizes other *GW*-body proteins and impairs *RNA*i and mRNA-induced gene silencing (Uhlen, Oksvold et al. 2010).
GLC	**NE**	rs13244987 rs13438768	3.33×10^{-8} 9.12×10^{-8}	7	*LOC645249*	*LOC645249* is a human protein coding gene, expressing differently between tumor and normal samples. The expression profile of the gene has been previously studied in human non-small-cell lung cancer (Takahashi, Forrest et al. 2009).
		rs7308621	4.31×10^{-4}	12	*RERG*	RERG, a member of the RAS family of GTPases, inhibits cell proliferation and tumor formation (Finlin, Gau et al. 2001).

		rs7982922 rs10492573 rs10492572	7.53×10^{-7} 5.63×10^{-6} 8.20×10^{-6}	13	*ENOX1*	Plasma membrane electron transport pathways are involved in functions such as cellular defense, intracellular redox homeostasis, and control of cell growth and survival (Uhlen, Oksvold et al. 2010). Candidate growth-related constitutive hydroquinone
		rs4939642	4.65×10^{-5}	18	*MAPK4*	Mitogen-activated protein (MAP) kinase 4 is a member of the mitogen-activated protein kinase family (Uhlen, Oksvold et al. 2010). MAP kinase pathways constitute one of the hallmarks of many cancers (Kostenko, Dumitriu et al. 2012).
	MH	rs17035917 rs17035960	4.65×10^{-6} 4.65×10^{-6}	4	*ARHGEF38*	Pancreatic islets, lung macrophages, breast and myocytes as well as basal cells in prostate, squamous and respiratory epithelium and showed strong staining (Uhlen, Oksvold et al. 2010).
		rs1421156 rs656875 rs672801	1.72×10^{-6} 1.30×10^{-6} 4.43×10^{-6}	9	*TRPM3*	The product of *TRPM3* belongs to the family of transient receptor potential (TRP) channels. TRP channels are cation-selective and are important for cellular calcium signaling and homeostasis (Uhlen, Oksvold et al. 2010).
SLRI	**NE**	rs1337862 rs1337863 rs945949	1.23×10^{-4} 1.11×10^{-4} 1.10×10^{-4}	6	*NKAIN2*	The protein encoded by this gene is a transmembrane protein that interacts with the beta subunit of Na, K-ATPase (ATP1B1). A chromosomal translocation involving this gene is a cause of lymphoma.
		rs12956176 rs4486983 rs9646509 rs1880113 rs1403762	1.55×10^{-5} 1.75×10^{-5} 3.37×10^{-5} 2.98×10^{-5} 8.14×10^{-5}	18	*KLHL14*	protein coding, interacts with Torsin A
	MH	rs6872156	8.31×10^{-6}	5	*DUSP1*	Candidate cancer biomarker. DUSP1/MKP1 is a dual-specific phosphatase that regulates MAPKs activity, with an increasingly recognized role in tumor biology (Moncho-Amor, Ibanez de Caceres et al. 2011).

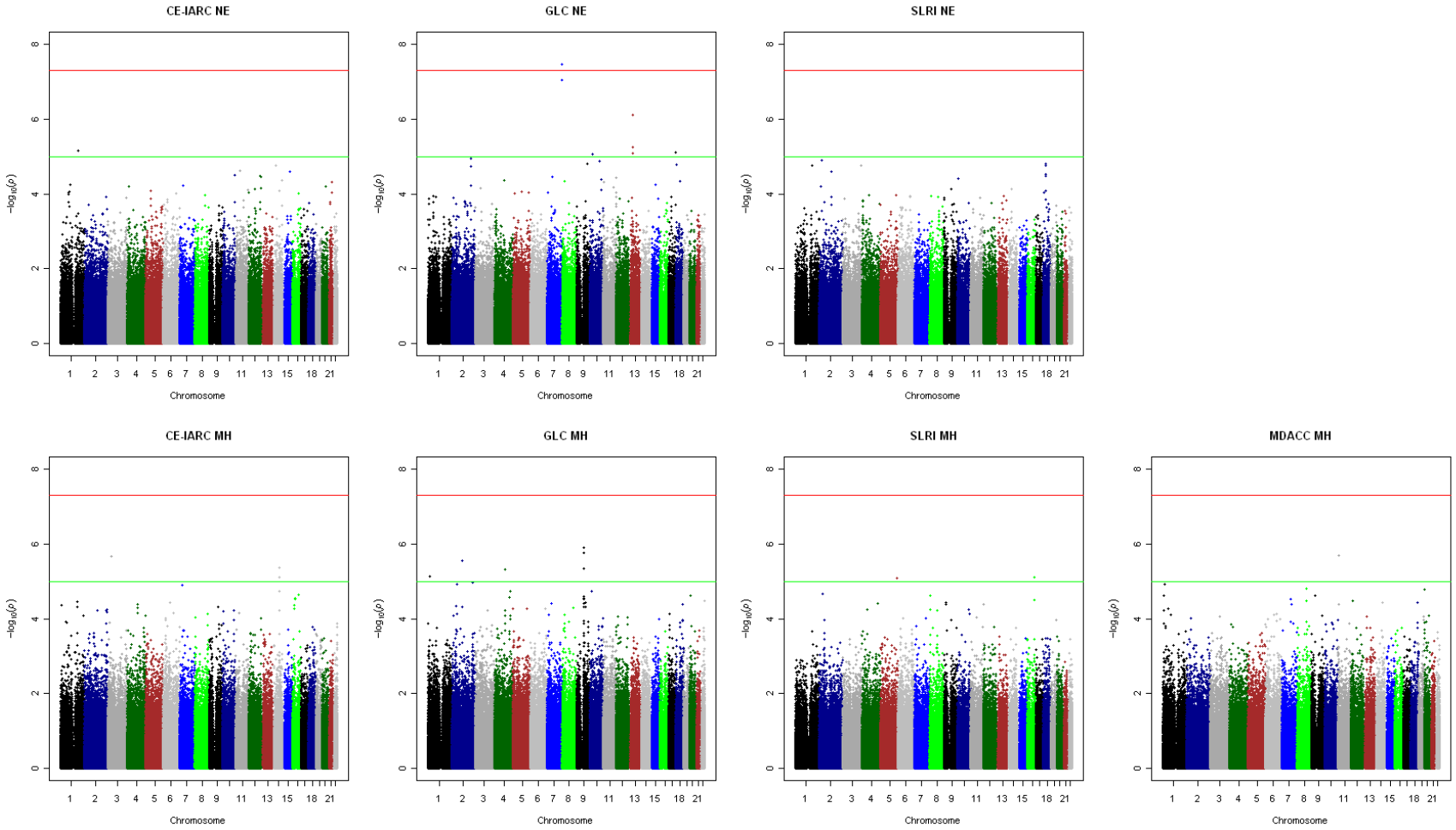
		rs6475227 rs7863071	4.13×10^{-5} 4.62×10^{-5}	9	*ADAMTSL1*	*ADAMTSL1* encodes a secreted protein and member of the ADAMTS family. This protein may have important functions in the extracellular matrix" (Uhlen, Oksvold et al. 2010).
		rs1876761 rs9927953	3.11×10^{-5} 7.66×10^{-6}	16	+/- 500kb of the *WWOX*	WW domain-containing proteins play an important role in the regulation of a wide variety of cellular functions such as protein degradation, transcription, and RNA splicing (Uhlen, Oksvold et al. 2010).
MDACC						
	MH	rs9323666	3.79×10^{-5}	14	*NRXN3*	Neurexins are a family of proteins that function in the vertebrate nervous system as cell adhesion molecules and receptors (Uhlen, Oksvold et al. 2010). Polymorphic site of NRXN3 gene was significantly associated with risk of breast cancer (Kusinska, Górniak et al. 2012). NRXN3 polymorphisms are associated with alcohol dependence (Hishimoto, Liu et al. 2007).

Abbreviations: genome-wide association study, GWAS; single nucleotide polymorphism, SNP; a parametric empirical hierarchical Bayes approach for G×E interaction, EHB-GE_{NN}; *E* = environmental coding (NE = *never* vs. *ever*, MH = *moderate* vs. *heavy*); p, p-value;

*Listed are only SNPs located within ±*500kb* of coding regions and with *p-values* < 10^{-4}

**Characteristics or function of the gene or function of the nearest gene to the SNP

Figure 6.4 Manhattan plots of p-values for EHB-GE_{NN}. Depicted are p-values for each SNP



Solid red line specifies 10^{-8} level of significance; Solid green line specifies 10^{-5} level of significance;

6.7. Results of Joint Tests for Genetic Main and G×E Interaction Effects

We performed four joint tests, namely EHB-GE_{NN}^{J}, as well as CC^{J}, CO^{J}, MUK-EB^{J}, on our GWAS data, to search for the association signals possibly missed by classic main effect or pure interaction genome-wide association tests. For all the studies and both exposure models, NE and MH, we only address SNPs with p-values of interaction effect $\leq 10^{-5}$ located within ±*500kb* of coding regions.

In the CE-IARC Study with NE coding, two SNPs on chromosome *15q24-25* rs*1051730* and *rs8034191* mapped to the nicotine acetylcholine acceptor subunit *CHRNA3* and *AGPHD1* genes had genome-wide significant p-values applying $T^{J}_{EHB-GE_{NN}}$ ($p = 6.0\times10^{-10}$, $p = 2.4\times10^{-9}$). These two markers were previously reported as being in association with lung cancer, identified by the classic main genetic effect test (Amos, Wu et al. 2008, Hung, McKay et al. 2008, Amos, Gorlov et al. 2010, Fehringer, Liu et al. 2012, Timofeeva, Hung et al. 2012) and discussed in Section 6.4. For the same GWAS with MH coding, the same two SNPs *rs1051730*, and *rs8034191* had greater p-values; however, they remained significant on the genome-wide level ($p = 5.8\times10^{-9}$, $p = 6.7\times10^{-8}$). Another signal of two SNPs, *rs13106574* ($p = 8.5\times10^{-6}$), *rs13149938* ($p = 1.8\times10^{-6}$) on chromosome *4* that both belong to the gene *SLC10A6* was discovered in this analysis. The *SLC10A6* locus is an important human sodium-dependent organic anion transporter gene, member 6 of the solute carrier family 10 (sodium/bile acid cotransporter family).

Joint T^{J}_{CC} identified a SNP on chromosome *5p15* that mapped to the *TERT* gene: *rs2736100* with corresponding p-value of 8.5×10^{-6}. The telomerase reverse transcriptase (*TERT*) gene is a candidate lung cancer biomarker. Recently, a number of studies reported *TERT* variant *rs2736100* in association with lung cancer impacting differently on lung cancer histology in European populations (Landi, Chatterjee et al. 2009, Truong, Hung et al. 2010, Brenner,

Boffetta et al. 2012, Timofeeva, Hung et al. 2012). SNPs of the *TERT* gene including *rs2736100* were also found to be associated with the risk of lung cancer in the Chinese population (Li, Yin et al. 2013, Myneni, Chang et al. 2013). This variant is discussed in Section 6.5. However, it is important to mention that it was previously only identified in a large scale meta-analysis study and that no single study prior to this has reported this signal. We also identified signals on chromosomes *16* and *14* with $T^{J}_{EHB-GE_{NN}}$ described in **Table 6.5** for CE-IARC, MH.

In the GLC study with never vs. ever coding, five SNPs with p-values lower than 10^{-5} on chromosome *13* that belong to different genes including *ENOX1* were identified by joint $T^{J}_{EHB-GE_{NN}}$ analysis (*rs1014744, rs10492572, rs10492573, rs10507886,* and *rs7982922),* see **Table 6.4**. The *ENOX1* protein is the constitutive *ENOX* family protein with an essential role in the enlargement phase of cell growth (Jiang, Gorenstein et al. 2008). It belongs to the same protein family and is very similar to the *ENOX2* gene that expresses on the cell surface of malignancies and is detectable in the serum of patients with cancer (Cho, Chueh et al. 2002, Hostetler, Weston et al. 2009). Three regions on chromosome *13* including *13q14,* where *ENOX1* is mapped, were reported to influence non-small-cell lung cancer (NSCLC) development (Tamura, Zhang et al. 1997). Another signal in the same analysis (GLC, NE) comprises four SNPs on chromosome *7 rs13244987* ($p = 5.1\times10^{-8}$), *rs13438768* ($p = 4.2\times10^{-8}$), *rs847916* ($P = 7.9\times10^{-6}$), *rs847918* ($P = 6.3\times10^{-8}$). With the MH model, we found the signals on chromosome *9* as described in **Table 6.5** and some additional individual association signals spread along the genome.

In SLRI NE, we identified the following SNPs with $T^{J}_{EHB-GE_{NN}}$. Two SNPs *rs10517026* ($p = 1.98\times10^{-6}$) and *rs10517026* ($p = 1.61\times10^{-6}$) on chromosome *4* mapped to the protein coding region. One marker, namely *rs12956176* located in the *KLHL14* gene on chromosome *18,* had

a p-value $\leq 10^{-6}$ with $T^{J}_{EHB-GE_{NN}}$ and $\leq 10^{-7}$ with T^{J}_{CC}. For the SLRI Study and the MH model, the $T^{J}_{EHB-GE_{NN}}$ test did not identify any SNPs with p-values $\leq 10^{-5}$. The T^{J}_{CC}test revealed five SNPs with p-values $\leq 10^{-5}$ that belong to genes. Data are in **Table 6.5**.

In MDACC, MH analysis, SNPs in the *SLC24A3* gene (*rs1555852*, *rs2876537*, *rs4239730*) on chromosome *20* form possible association signal T^{J}_{CC} ($p = 4.6\times10^{-6}$, $p = 3.3\times10^{-6}$, $p = 2.9\times10^{-6}$). The *SLC24A3* product is known as prostate cancer-associated protein 6.

The most prominent findings of simultaneous testing with each coding (NE, MH), each joint test statistic (EHB-GE_{NN}^{J}, CC^{J}, CO^{J}, MUK-EB^{J}) and for each GWAS (CE-IARC, GLC, SLRI, MDACC) are summarized in **Table 6.5**. In **Table 6.5**, only SNPs with corresponding p-values $\leq 10^{-5}$ for at least one of the joint tests and those located within the known genes or maximum ±500Kb away from the gene were included. Manhattan plots in **Figure 6.5** visualize the results for each GWAS. Generally EHB-GE_{NN}^{J} has similar power and as the consequence similar p-values as CO^{J} and MUK-EB^{J} and smaller p-values compared to CC^{J}. However when SNP has protective effect against the outcome (negative estimated coefficient of the association with the trait) then CC^{J} has greater power and as consequence lower p-values for such signals. An example of the later statement are SNPs *rs2736100* (*TERT*, CE-IARC GWAS, MH) and *rs9347645* (*PARK2*, SLRI GWAS, MH) in **Table 6.5**. Even though the simulation study in Chapter 5 reflected slight power loss of the EHB-GE_{NN} compared to both CO^{J} and MUK-EB^{J} in real data we observed only minor increase in p-values for the important signals. For example, SNP r*s1051730* (*CHRNA3*, CE-IARC GWAS, NE) has p-value 5.97×10^{-10} testing with EHB-GE_{NN}^{J} and 5.93×10^{-10} and 6.42×10^{-10} for CO^{J}, MUK-EB^{J} respectively.

Table 6.5 Markers indicated by joint tests in ILCCO/TRICL data with p-values ≤ 10^{-5} for at least one of the joint tests

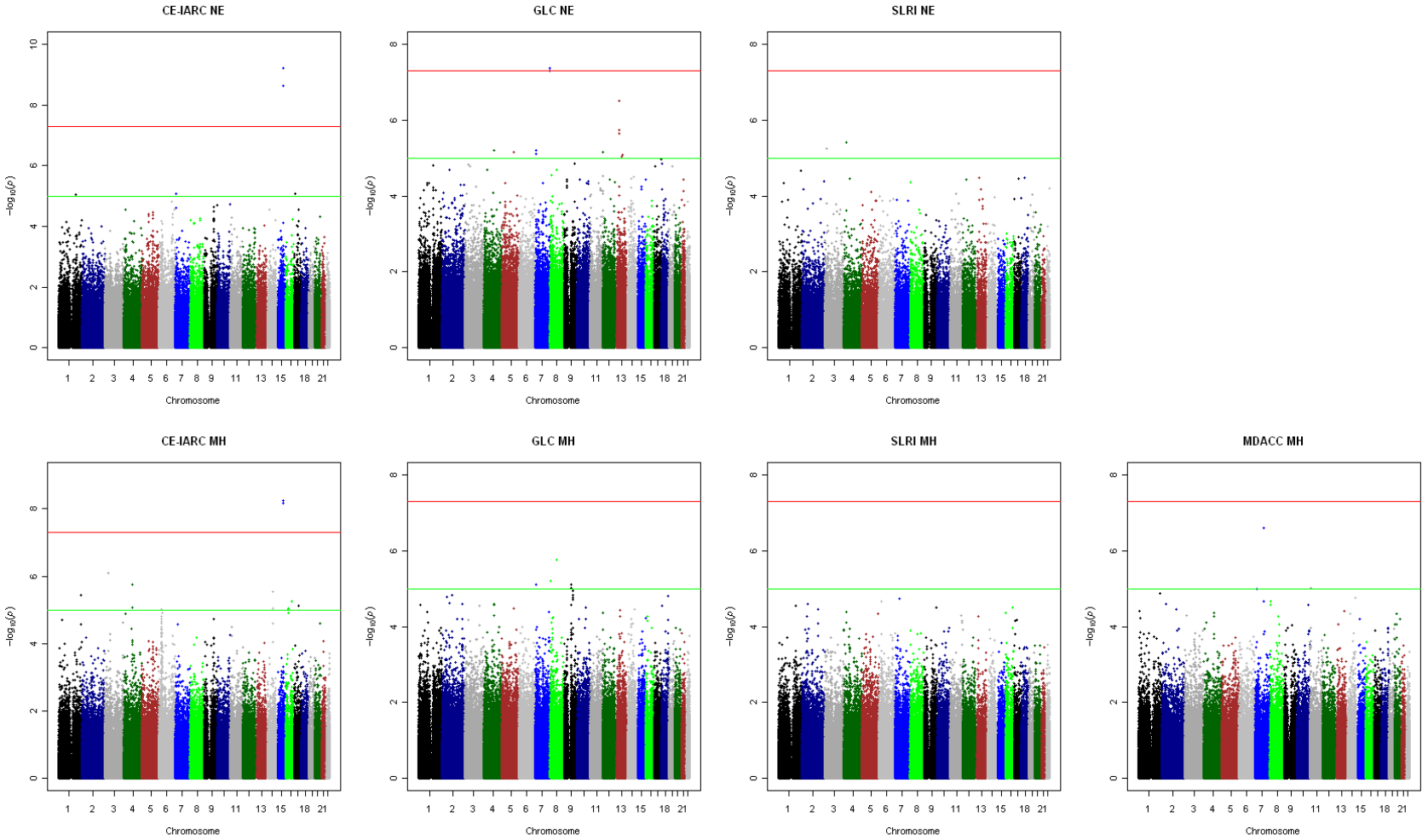
GWAS	*E*	*SNP*	*CHR*	*MA*	*Gene*	$p\ T^{J}_{EHB\text{-}GE(NN)}$	$p\ T^{J}_{cc}$	$p\ T^{J}_{co}$	$p\ T^{J}_{MUK\text{-}EB}$
	NE	rs2779286**	1	G	*C1orf21*	$9{,}00\times10^{-06}$	$1{,}30\times10^{-03}$	$8{,}72\times10^{-06}$	$1{,}64\times10^{-04}$
		rs38012	7	G	*GLCCI1*	$8{,}09\times10^{-06}$	$1{,}72\times10^{-04}$	$7{,}94\times10^{-06}$	$5{,}38\times10^{-05}$
		rs3784179	14	C	*AKAP6*	$2{,}79\times10^{-05}$	$3{,}67\times10^{-06}$	$2{,}86\times10^{-05}$	$4{,}02\times10^{-05}$
		rs1051730*	**15**	**A**	***CHRNA3***	$\mathbf{5{,}97\times10^{-10}}$	$\mathbf{1{,}67\times10^{-09}}$	$\mathbf{5{,}93\times10^{-10}}$	$\mathbf{6{,}42\times10^{-10}}$
		rs8034191*	**15**	**C**	***AGPHD1***	$\mathbf{2{,}37\times10^{-09}}$	$\mathbf{5{,}34\times10^{-09}}$	$\mathbf{2{,}36\times10^{-09}}$	$\mathbf{2{,}63\times10^{-09}}$
		rs9302935	17	G	*LOC400618*	$1{,}27\times10^{-03}$	$1{,}93\times10^{-06}$	$1{,}29\times10^{-03}$	$3{,}09\times10^{-05}$
		rs1006957	17	T	*UBB*	$8{,}25\times10^{-06}$	$1{,}35\times10^{-05}$	$8{,}27\times10^{-06}$	$8{,}21\times10^{-06}$
	MH	rs6685121	1	G	*LOC100505872*	$3{,}71\times10^{-06}$	$5{,}29\times10^{-06}$	$3{,}86\times10^{-06}$	$4{,}45\times10^{-06}$
		rs7620618**	3	T	*±GOLGA4*	$8{,}10\times10^{-07}$	$1{,}99\times10^{-03}$	$8{,}26\times10^{-07}$	$2{,}47\times10^{-05}$
		rs13149938	4	G	*SLC10A6*	$1{,}80\times10^{-06}$	$4{,}09\times10^{-06}$	$1{,}79\times10^{-06}$	$1{,}81\times10^{-06}$
		rs13106574	4	C	*SLC10A6*	$8{,}54\times10^{-06}$	$1{,}44\times10^{-05}$	$8{,}46\times10^{-06}$	$8{,}77\times10^{-06}$
CE-IARC		rs2736100	5	C	*TERT*	$4{,}70\times10^{-04}$	$8{,}50\times10^{-06}$	$4{,}95\times10^{-04}$	$1{,}34\times10^{-04}$
		rs4563628**	5	C	*±TAG*	$5{,}76\times10^{-04}$	$7{,}52\times10^{-06}$	$5{,}87\times10^{-04}$	$6{,}79\times10^{-04}$
		rs4324798	6	A	*LOC401242*	$9{,}35\times10^{-06}$	$3{,}00\times10^{-05}$	$9{,}51\times10^{-06}$	$1{,}30\times10^{-05}$
		rs1076204	11	C	*ABCC8*	$3{,}30\times10^{-05}$	$9{,}33\times10^{-06}$	$3{,}48\times10^{-05}$	$1{,}46\times10^{-05}$
		rs2302591**	14	T	*TTLL5*	$2{,}90\times10^{-06}$	$5{,}37\times10^{-03}$	$2{,}88\times10^{-06}$	$4{,}04\times10^{-06}$
		rs175891**	14	G	*TTLL5*	$9{,}29\times10^{-06}$	$9{,}45\times10^{-03}$	$8{,}93\times10^{-06}$	$1{,}47\times10^{-03}$
		rs1051730*	**15**	**A**	***CHRNA3***	$\mathbf{5{,}76\times10^{-09}}$	$\mathbf{3{,}42\times10^{-09}}$	$\mathbf{5{,}77\times10^{-09}}$	$\mathbf{5{,}76\times10^{-09}}$
		rs8034191*	**15**	**C**	***AGPHD1***	$\mathbf{6{,}72\times10^{-09}}$	$\mathbf{6{,}33\times10^{-09}}$	$\mathbf{6{,}72\times10^{-09}}$	$\mathbf{6{,}72\times10^{-09}}$
		rs200528	16	A	*TNRC6A*	$9{,}29\times10^{-06}$	$3{,}37\times10^{-03}$	$8{,}74\times10^{-06}$	$5{,}29\times10^{-04}$
		rs2112783	16	A	*TNRC6A*	$9{,}70\times10^{-06}$	$3{,}87\times10^{-03}$	$9{,}13\times10^{-06}$	$5{,}86\times10^{-04}$
		rs9937754	16	T	*LOC1009*	$5{,}47\times10^{-06}$	$8{,}66\times10^{-05}$	$5{,}96\times10^{-06}$	$2{,}25\times10^{-05}$
		rs12944442	17	A	*ANKFN1*	$7{,}57\times10^{-06}$	$9{,}88\times10^{-06}$	$7{,}47\times10^{-06}$	$9{,}51\times10^{-06}$
		rs2866908*	4	T	*DKK2*	$6{,}26\times10^{-06}$	$1{,}42\times10^{-05}$	$6{,}26\times10^{-06}$	$6{,}86\times10^{-06}$
	NE	rs6891265	5	C	*±SLC27A6*	$6{,}81\times10^{-06}$	$2{,}89\times10^{-05}$	$6{,}82\times10^{-06}$	$9{,}77\times10^{-06}$
		rs13244987**	**7**	**A**	***LOC645249***	$\mathbf{5{,}09\times10^{-08}}$	$\mathbf{2{,}37\times10^{-05}}$	$\mathbf{5{,}04\times10^{-08}}$	$\mathbf{5{,}09\times10^{-08}}$
		rs13438768**	**7**	**C**	***±LOC645249***	$\mathbf{4{,}19\times10^{-08}}$	$\mathbf{1{,}32\times10^{-04}}$	$\mathbf{4{,}23\times10^{-08}}$	$\mathbf{4{,}19\times10^{-08}}$
		rs847916*	7	G	*±SCIN*	$7{,}91\times10^{-06}$	$1{,}05\times10^{-05}$	$7{,}90\times10^{-06}$	$9{,}87\times10^{-06}$
GLC		rs847918	7	T	*±SCIN*	$6{,}34\times10^{-06}$	$1{,}93\times10^{-05}$	$6{,}33\times10^{-06}$	$6{,}89\times10^{-06}$
		rs10849065	12	T	*B4GALNT3*	$6{,}78\times10^{-06}$	$7{,}79\times10^{-04}$	$6{,}79\times10^{-06}$	$3{,}42\times10^{-04}$
		rs7982922**	13	A	*ENOX1*	$3{,}01\times10^{-07}$	$1{,}20\times10^{-06}$	$3{,}00\times10^{-07}$	$5{,}46\times10^{-07}$
		rs10492572**	13	T	*ENOX1*	$1{,}85\times10^{-06}$	$1{,}15\times10^{-05}$	$1{,}84\times10^{-06}$	$1{,}86\times10^{-06}$

		rs10492573**	13	G	*ENOX1*	$2{,}27\times10^{-06}$	$1{,}32\times10^{-05}$	$2{,}29\times10^{-06}$	$2{,}31\times10^{-06}$
		rs10507886	13	T	*±POU4F1*	$8{,}34\times10^{-06}$	$1{,}45\times10^{-04}$	$8{,}28\times10^{-06}$	$1{,}79\times10^{-05}$
		rs1014744	13	T	*±ATXN8OS*	$8{,}85\times10^{-06}$	$9{,}09\times10^{-05}$	$8{,}81\times10^{-06}$	$4{,}75\times10^{-05}$
		rs9911873	17	G	*LUC7L3*	$3{,}46\times10^{-04}$	$8{,}89\times10^{-06}$	$3{,}47\times10^{-04}$	$5{,}14\times10^{-05}$
	MH	rs2866908*	4	T	*DKK2*	$4{,}98\times10^{-05}$	$3{,}84\times10^{-06}$	$4{,}98\times10^{-05}$	$1{,}10\times10^{-05}$
		rs9643575*	8	C	*TRIM55*	$1{,}69\times10^{-06}$	$6{,}69\times10^{-06}$	$1{,}68\times10^{-06}$	$1{,}69\times10^{-06}$
		rs4876151	8	C	*±MYOM2*	$6{,}12\times10^{-06}$	$5{,}66\times10^{-05}$	$6{,}13\times10^{-06}$	$6{,}13\times10^{-06}$
		rs656875**	9	C	*TRPM3*	$7{,}56\times10^{-06}$	$7{,}08\times10^{-04}$	$7{,}54\times10^{-06}$	$8{,}55\times10^{-06}$
		rs1421156**	9	G	*TRPM3*	$9{,}52\times10^{-06}$	$6{,}09\times10^{-04}$	$9{,}57\times10^{-06}$	$1{,}33\times10^{-05}$
SLRI	**NE**	rs10517026*	4	G	*LOC100507930*	$1{,}98\times10^{-06}$	$4{,}23\times10^{-07}$	$1{,}98\times10^{-06}$	$2{,}40\times10^{-06}$
		rs10517031*	4	G	*LOC100507930*	$1{,}61\times10^{-06}$	$3{,}39\times10^{-07}$	$1{,}61\times10^{-06}$	$1{,}86\times10^{-06}$
		rs12956176**	18	A	*KLHL14*	$5{,}40\times10^{-05}$	$2{,}25\times10^{-06}$	$5{,}43\times10^{-05}$	$4{,}43\times10^{-05}$
	MH	rs10517026	4	G	*LOC100507930*	$7{,}01\times10^{-05}$	$6{,}66\times10^{-06}$	$7{,}01\times10^{-05}$	$7{,}13\times10^{-05}$
		rs10517031*	4	G	*LOC100507930*	$3{,}48\times10^{-05}$	$3{,}28\times10^{-06}$	$3{,}48\times10^{-05}$	$3{,}65\times10^{-05}$
		rs9347645	6	C	*PARK2*	$2{,}81\times10^{-03}$	$3{,}81\times10^{-06}$	$2{,}82\times10^{-03}$	$1{,}47\times10^{-03}$
		rs482449	11	T	*±MIR100HG*	$1{,}26\times10^{-01}$	$9{,}45\times10^{-07}$	$1{,}26\times10^{-01}$	$7{,}43\times10^{-04}$
		rs11631489	15	G	*AGBL1*	$1{,}10\times10^{-01}$	$3{,}29\times10^{-06}$	$1{,}10\times10^{-01}$	$5{,}15\times10^{-03}$
MDACC	**MH**	rs2538909	7	A	*±ZNF804B*	$2{,}52\times10^{-07}$	$6{,}92\times10^{-07}$	$2{,}53\times10^{-07}$	$7{,}82\times10^{-07}$
		rs552247	7	G	*±MEOX2*	$9{,}90\times10^{-06}$	$7{,}79\times10^{-06}$	$9{,}93\times10^{-06}$	$1{,}19\times10^{-05}$
		rs12276659	11	G	*PARVA*	$9{,}53\times10^{-06}$	$3{,}07\times10^{-03}$	$9{,}55\times10^{-06}$	$9{,}54\times10^{-06}$
		rs4239730	20	A	*SLC24A3*	$4{,}59\times10^{-05}$	$2{,}90\times10^{-06}$	4.63×10^{-05}	2.49×10^{-05}
		rs2876537	20	C	*SLC24A3*	1.91×10^{-04}	3.30×10^{-06}	1.93×10^{-04}	2.63×10^{-05}
		rs1555852	20	A	*SLC24A3*	8.61×10^{-05}	4.57×10^{-06}	8.69×10^{-05}	3.66×10^{-05}

Abbreviations: GWAS, genome-wide association study; SNP, single nucleotide polymorphism; CHR, chromosome number; MA, minor allele; EHB-GENN, a parametric empirical hierarchical Bayes approach for G×E interaction; ILCCO, International Lung and Cancer Consortium/ TRICL, Transdisciplinary Research in Cancer of the Lung; ±, denote that SNP locates ±500Kb of the gene; CC, classical case-control interaction estimator; CO, case-only interaction estimator; MUK-EB, empirical Bayes shrinkage estimator; TJ, joint test statistics; p, p-value associated with the joint test

E = environmental coding *(*NE = *never* vs. *ever*, MH = *moderate* vs. *heavy;* Gene = SNP to gene or nearest gene annotation; * denotes SNPs with *p-value* $\leq 10^{-5}$ testing for classical genetic main effect; ** denotes SNPs with *p-value* $\leq 10^{-5}$ testing for G×E interaction only by EHB-GE_{NN} in bold are markers with *p-value*$\leq 10^{-7}$;

Figure 6.5 Manhattan plots of p-values for SNPs joint effect based on the EHB-GE_{NN} test for G×E interaction component



Red line specifies 10^{-8} level of significance; Green line specifies 10^{-5} level of significance;

Chapter 7

7. Discussion

One objective of research in human genetics is to understand how genetic and environmental factors interact to cause different diseases. In statistical terms, G×E interaction is present when the effect of the genotype on disease risk depends on the level of exposure to an environmental factor, or vice versa (Clayton and McKeigue 2001). In this dissertation, three major concerns to studies of G×E interaction were addressed: the extent of bias due to the uncovered population stratification; the presence of G-E correlation; and the lower power of common tests to identify an interaction.

In Chapter 3, we focused on the evaluation of bias due to population stratification in studies of G×E interaction. We derived an equation to evaluate the population stratification bias for the case-control estimator of the interaction odds ratio. We demonstrated analytically that population stratification bias can reach an intolerable level for case-control studies of G×E interaction. We compared bias in estimates of G×E interaction effects in case-control and case-only studies with bias in genetic main effect estimates. We concluded that the case-control design is significantly more robust to population stratification than the case-only design. On average, the degree of bias for the G×E interaction effect estimate in case-control studies is similar to that in genetic main effect studies and constitutes about 2%-3%. Exceptions are some extreme situations that cannot easily be avoided, an example of which is the admixture of two subpopulations in a study sample. In this situation, the bias can reach on average 10%-15%. Researchers should be aware that bias can theoretically rise up to 30%-40% in case-control studies of interactions and to over 50% in case-only studies.

We compared common methods to detect G×E interaction in terms of their robustness to population stratification. These methods are the classic case-control approach, the case-only

approach, Mukherjee's empirical Bayes method, Murcray's two step approach, and the empirical hierarchical Bayes method to G×E interaction, based on the chi distribution. We demonstrated that principal component analysis adjustment reduced population stratification bias to almost zero and is thus also appropriate to correct for this bias in G×E interaction studies.

EHB-GE_{CHI} was previously demonstrated as an attractive method for GWAS of G×E interactions (Sohns, Viktorova et al. 2013). Therefore, further extensions of the EHB-GE_{CHI} approach were presented in Chapter 4. Extending the original work of Melanie Sohns (Sohns 2012), we demonstrated that the method can handle multilevel and continuous genotype and exposure variables. We also showed that it is applicable under the assumption of the log-additive genetic model on the multiplicative scale and can deal with covariate adjustment. These extensions are essential, as they allow more flexibility in the use of the original EHB-GE_{CHI} approach. In many situations, information on the exposure is collected as a categorical or continuous variable. Therefore, the ability of the approach to work with the original data without their reduction to a binary variable is important and may help in obtaining a more precise estimation.

Generally, the ability of the particular statistical approach to handle various genetic models such as additive, dominant, and recessive, makes it more attractive, since a properly chosen model adds power to the interaction test. By means of the simulation study and reliance on asymptotic theory, we revealed that EHB-GE_{CHI} is valid under the assumption of the log-additive risk inheritance model. The validity of the approach for dominant and recessive models is illustrated in the dissertation of Sohns (Sohns 2012). The additive risk model is commonly used to model the risk inheritance mode in epidemiologic study, since it takes into account natural genotype coding, according to the minor allele count. Furthermore, it performs optimally under the unknown true inheritance mode scenario.

The adjustment for major covariates such as sex, age, and principal components for ethnicity (as discussed in Chapter 3) is often needed in genetic association and interaction studies. Therefore, proof of the validity of the EHB-GE_{CHI} approach under the adjusted analysis is clearly important. We showed that if independence of the interaction OR and the covariate is given, then separate adjustment within cases and controls leads to the same estimates as those resulting from adjustment in the whole case-control sample. This proof validates EHB-GE_{CHI} after separate adjustment within cases and controls, as required by the approach. We proposed using log-linear models when the independence assumption is not valid, in order to obtain adjusted estimates for cases and controls simultaneously. However, a limitation of the log-linear regression framework is its ability to model only categorical variables but not continuous ones.

In Chapter 5, we developed an alternative, computationally much faster approach (it requires three times less computer time (CPU)) compared to the EHB-GE_{CHI}, approach. Another prominent advantage of our novel EHB-GE_{NN} method is that the analytically closed form of the posterior distribution for the test statistics of this approach is available. EHB-GE_{NN} is based on a two-stage hierarchical model, necessary to estimate G-E correlation effects in controls effectively. It is proposed as a tool to account for population-based G-E correlation, one of the biggest concerns in studies of G×E interactions. It is well known that the Gaussian family is a self-conjugate with respect to a Gaussian likelihood function. Therefore, choosing a normal distribution as a prior probability distribution for the mean of G-E correlations ensures that the posterior distribution is also normal. We assumed normal distributions at both stages, benefitting from the resulting analytical normal form of the posterior distribution and also from the closed form of the posterior variance of G-E correlation estimates. This is in contrast to EHB-GE_{CHI}. Our novel approach controls type I error substantially better than EHB-GE_{CHI} and suffers only minor power loss. The EHB-GE_{NN} approach is more stable in terms of the

hyperparameter estimation, as it requires only one common parameter to be estimated, gathering information on the whole available data in contrast to only three hyperparameters for EHB-GE_{CHI}. It is easily extendable to handle multilevel or continuous genotype and exposure variables, as this works in the same manner as shown for EHB-GE_{CHI} in Chapter 4. We implemented both EHB-GE_{CHI} and EHB-GE_{NN} in an R package that has been named EHBg×e. Performing an extensive simulation study, we evaluated properties of the EHB-GE_{NN} approach. Based on the observed results, we recommend performing EHB-GE_{NN} to test for the interaction when a large number of G-E correlation signals with moderate to high effect size are expected to exist in the study sample. We also suggest applying EHB-GE_{NN} in studies with frequent exposure variable, so that the strata are large enough for the hyperparameter estimation. EHB-GE_{NN} can be applied for significance testing in GWAS to search for G×E interaction signals without assuming G-E independence. This is in contrast to the case-only or Mukherjee's empirical Bayes tests. It maintains adequate power and almost always performs better than the case-control or Murcray's two step tests. Case-control or Murcray's two step also do not require any assumption of G-E independence.

Joint tests are performed to detect variants that have moderate marginal effects on an outcome, differing according to an environmental factor that would be potentially missed by the main effect genome-wide analysis or pure interaction analysis. EHB-GE_{NN} can easily be used to construct a joint test for genetic marginal and G×E interaction effect, similar to the joint tests proposed by Dai and colleagues (Dai, Logsdon et al. 2012). In contrast to the CO test that was employed in Dai's 2 *df* test construction, our EHB-GE_{NN} approach, as well as its joint version, do not require any assumption of G-E independence, which can be critical in the context of a large-scale genome-wide association study. Therefore, we constructed the joint EHB-GE_{NN}^{J} test for simultaneous testing of genetic main and G×E interaction effects in a similar fashion to Dai.

Our work was motivated by lung cancer GWAS data from the ILCCO/TRICL consortium with smoking being the established environmental risk factor. With the aim of identifying promising association signals for lung cancer, we conducted a statistical analysis on four lung cancer GWAS datasets. We replicated previous findings, namely two known SNPs on chromosome *15q24-25* that belong to the nicotine acetylcholine acceptor subunit *CHRNA3* and *AGPHD1*genes with slightly lower *p-values* than previously reported and described signals in our data worth further investigation, e.g. SNPs located in *TERT* and *ENOX1* genes. Nowadays, *TERT* is one of the most interesting genes in the study of lung cancer risk. SNP *rs2736100* in the *TERT* gene was reported as being in association with adenocarcinoma risk on the basis of a large genome-wide association study involving 13,300 cases and 19,666 controls of European descent and 3,333 subjects with adenocarcinoma among them (Landi, Chatterjee et al. 2009). The same variant was found to influence the risk of lung cancer in two meta-analyses; the first with 16 pooled GWASs involving 14,900 cases and 29,485 controls (Timofeeva, Hung et al. 2012) and the second with 21 pooled GWASs involving 11,645 cases and 14,954 controls (Truong, Hung et al. 2010). To date, none of the single case-control GWAS were able to find these SNPs without requiring huge datasets and meta-analytical approaches. In our study with the joint test, we identified this SNP with *p-value* 8.5×10^{-6} based on only 1,989 cases and 2,625 controls in the CE-IARC data with the moderate-heavy smoking model. For the same data, the variant has a p-value of 2.5×10^{-4} when testing for G×E interaction with classic CC test and a p-value of 1.6×10^{-3} when testing for genetic main effect. As a result, it was previously missed by both interaction and main effect tests in our data. This demonstrates that joint tests are useful in the identification of missing genetic main effect signals and require considerably smaller sample size than compared to meta-analytic approaches. This can be crucial in many situations and for many complex diseases such as cancer.

Both EHB-GE_{NN} and EHB-GE_{NN}^{J} tests indicated a novel association signal of SNPs *rs7982922, rs10492572,* and *rs10492573,* located on the *ENOX1* gene on chromosome *13q14*. ENOX proteins *(ENOX1*, *ENOX2* and *ENOX3)* are a unique family of cell surface proteins, playing an essential role in the enlargement phase of cell growth (*ENOX1*) and unregulated cancer cell growth (*ENOX2)*. Both *ENOX1* and *ENOX2* are found in the sera of cancer patients. These proteins highly relate to each other and in fact share 64% of identity and 80% of similarity in humans (Morré and Morré 2012). Deletion of three distinct regions on chromosome *13* including the *13q14* region in which *ENOX1* is located was reported for NSCLC (Tamura, Zhang et al. 1997). This suggests that *ENOX1* variants, namely *rs7982922, rs10492572,* and *rs10492573,* might form an interesting signal for the risk of lung cancer development.

Future research is necessary to study these signals in more detail with regard to their functions and molecular biology, as well as to replicate these association results in other studies of Europeans or other populations.

A meta-analysis across more GWASs based on the joint testing techniques and allowing for lung cancer subtypes may lead to a consolidation of the results.

8. References

Albert, P. S., D. Ratnasinghe, J. Tangrea and S. Wacholder (2001). "Limitations of the case-only design for identifying gene-environment interactions." Am J Epidemiol 154(8): 687-693.

Amos, C. I. (2007). Transdisciplinary Research in Cancer of the Lung, Website \url{http://u19tricl.org/}.

Amos, C. I., I. P. Gorlov, Q. Dong, X. Wu, H. Zhang, E. Y. Lu, P. Scheet, A. J. Greisinger, G. B. Mills and M. R. Spitz (2010). "Nicotinic acetylcholine receptor region on chromosome 15q25 and lung cancer risk among African Americans: a case-control study." J Natl Cancer Inst 102(15): 1199-1205.

Amos, C. I., S. M. Pinney, Y. Li, E. Kupert, J. Lee, M. A. de Andrade, P. Yang, A. G. Schwartz, P. R. Fain, A. Gazdar, J. Minna, J. S. Wiest, D. Zeng, H. Rothschild, D. Mandal, M. You, T. Coons, C. Gaba, J. E. Bailey-Wilson and M. W. Anderson (2010). "A susceptibility locus on chromosome 6q greatly increases lung cancer risk among light and never smokers." Cancer Res 70(6): 2359-2367.

Amos, C. I., X. Wu, P. Broderick, I. P. Gorlov, J. Gu, T. Eisen, Q. Dong, Q. Zhang, X. Gu, J. Vijayakrishnan, K. Sullivan, A. Matakidou, Y. Wang, G. Mills, K. Doheny, Y.-Y. Tsai, W. V. Chen, S. Shete, M. R. Spitz and R. S. Houlston (2008). "Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1." Nat Genet 40(5): 616-622.

Andersen, E. B. (1970). "Asymptotic properties of conditional maximum-likelihood estimators." Journal of the Royal Statistical Society. Series B (Methodological) 32(2): 283-301.

Ardlie, K. G., L. Kruglyak and M. Seielstad (2002). "Patterns of linkage disequilibrium in the human genome." Nat Rev Genet 3(4): 299-309.

Armstrong, B. G. (2003). "Fixed factors that modify the effects of time-varying factors: applying the case-only approach." Epidemiology 14(4): 467-472.

Bailey-Wilson, J. E., C. I. Amos, S. M. Pinney, G. M. Petersen, M. de Andrade, J. S. Wiest, P. Fain, A. G. Schwartz, M. You, W. Franklin, C. Klein, A. Gazdar, H. Rothschild, D. Mandal,

T. Coons, J. Slusser, J. Lee, C. Gaba, E. Kupert, A. Perez, X. Zhou, D. Zeng, Q. Liu, Q. Zhang, D. Seminara, J. Minna and M. W. Anderson (2004). "A major lung cancer susceptibility locus maps to chromosome 6q23-25." Am J Hum Genet 75(3): 460-474.

Bayes, T. (1991). "An essay towards solving a problem in the doctrine of chances. 1763." MD Comput 8(3): 157-171.

Berger, J. O. (1985). Statistical Decision Theory and Bayesian Analysis. New York, Springer.

Bhattacharjee, S., Z. Wang, J. Ciampa, P. Kraft, S. Chanock, K. Yu and N. Chatterjee (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(3): 331-342.

Bierut, L. J., P. A. Madden, N. Breslau, E. O. Johnson, D. Hatsukami, O. F. Pomerleau, G. E. Swan, J. Rutter, S. Bertelsen, L. Fox, D. Fugman, A. M. Goate, A. L. Hinrichs, K. Konvicka, N. G. Martin, G. W. Montgomery, N. L. Saccone, S. F. Saccone, J. C. Wang, G. A. Chase, J. P. Rice and D. G. Ballinger (2007). "Novel genes identified in a high-density genome wide association study for nicotine dependence." Hum Mol Genet 16(1): 24-35.

Bishop, Y. M. M., S. E. Fienberg and P. W. Holland (2007). Discrete Multivariate Analysis Theory and Practice, Springer New York.

Brenner, D. R., P. Boffetta, E. J. Duell, H. Bickeböller, A. Rosenberger, V. McCormack, J. E. Muscat, P. Yang, H.-E. Wichmann, I. Brueske-Hohlfeld, A. G. Schwartz, M. L. Cote, A. Tjønneland, S. Friis, L. Le Marchand, Z.-F. Zhang, H. Morgenstern, N. Szeszenia-Dabrowska, J. Lissowska, D. Zaridze, P. Rudnai, E. Fabianova, L. Foretova, V. Janout, V. Bencko, M. Schejbalova, P. Brennan, I. N. Mates, P. Lazarus, J. K. Field, O. Raji, J. R. McLaughlin, G. Liu, J. Wiencke, M. Neri, D. Ugolini, A. S. Andrew, Q. Lan, W. Hu, I. Orlow, B. J. Park and R. J. Hung (2012). "Previous Lung Diseases and Lung Cancer Risk: A Pooled Analysis From the International Lung Cancer Consortium." American Journal of Epidemiology 176(7): 573-585.

Broderick, P., Y. Wang, J. Vijayakrishnan, A. Matakidou, M. R. Spitz, T. Eisen, C. I. Amos and R. S. Houlston (2009). "Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study." Cancer Res 69(16): 6633-6641.

Brüske-Hohlfeld, I. (2009). Environmental and Occupational Risk Factors for Lung Cancer. Cancer Epidemiology. M. Verma, Humana Press. 472: 3-23.

Bryant, A. and R. J. Cerfolio (2007). "Differences in epidemiology, histology, and survival between cigarette smokers and never-smokers who develop non-small cell lung cancer." Chest Journal 132(1): 185-192.

Buselmaier, W. and G. Tariverdian (1999). Humangenetik, Springer.

Caporaso, N., F. Gu, N. Chatterjee, J. Sheng-Chih, K. Yu, M. Yeager, C. Chen, K. Jacobs, W. Wheeler, M. T. Landi, R. G. Ziegler, D. J. Hunter, S. Chanock, S. Hankinson, P. Kraft and A. W. Bergen (2009). "Genome-wide and candidate gene association study of cigarette smoking behaviors." PLoS One 4(2): e4653.

Catelinois, O., A. Rogel, D. Laurier, S. Billon, D. Hemon, P. Verger and M. Tirmarche (2006). "Lung cancer attributable to indoor radon exposure in france: impact of the risk models and uncertainty analysis." Environmental health perspectives 114(9): 1361-1366.

Cavalli-Sforza, L. L., P. Menozzi and A. Piazza (1994). The history and geography of human genes, Princeton University Press.

Chatterjee, N., Z. Kalaylioglu, R. Moslehi, U. Peters and S. Wacholder (2006). "Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions." Am J Hum Genet 79(6): 1002-1016.

Chen, J., M. J. Stampfer, H. L. Hough, M. Garcia-Closas, W. C. Willett, C. H. Hennekens, K. T. Kelsey and D. J. Hunter (1998). "A prospective study of N-acetyltransferase genotype, red meat intake, and risk of colorectal cancer." Cancer Res 58(15): 3307-3311.

Cheng, K. F. (2006). "A maximum likelihood method for studying gene-environment interactions under conditional independence of genotype and exposure." Stat Med 25(18): 3093-3109.

Chiu, H. F., M. H. Cheng, S. S. Tsai, T. N. Wu, H. W. Kuo and C. Y. Yang (2006). "Outdoor air pollution and female lung cancer in Taiwan." Inhal Toxicol 18(13): 1025-1031.

Cho, N., P.-J. Chueh, C. Kim, S. Caldwell, D. Morré and J. Morré (2002). "Monoclonal antibody to a cancer-specific and drug-responsive hydroquinone (NADH) oxidase from the sera of cancer patients." Cancer Immunology, Immunotherapy 51(3): 121-129.

Clarke, G. M. and A. P. Morris (2010). "A comparison of sample size and power in case-only association studies of gene-environment interaction." Am J Epidemiol 171(4): 498-505.

Clayton, D. and P. M. McKeigue (2001). "Epidemiological methods for studying genes and environmental factors in complex diseases." Lancet 358(9290): 1356-1360.

Collins, F. S., E. D. Green, A. E. Guttmacher and M. S. Guyer (2003). "A vision for the future of genomics research." Nature 422(6934): 835-847.

Cornfield, J. (1956). "A statistical problem arising from retrospective studies." Berkeley, Calif.: University of California Press: 179.

Coté, M. L., M. Liu, S. Bonassi, M. Neri, A. G. Schwartz, D. C. Christiani, M. R. Spitz, J. E. Muscat, G. Rennert, K. K. Aben, A. S. Andrew, V. Bencko, H. Bickeböller, P. Boffetta, P. Brennan, H. Brenner, E. J. Duell, E. Fabianova, J. K. Field, L. Foretova, S. Friis, C. C. Harris, I. Holcatova, Y.-C. Hong, D. Isla, V. Janout, L. A. Kiemeney, C. Kiyohara, Q. Lan, P. Lazarus, J. Lissowska, L. Le Marchand, D. Mates, K. Matsuo, J. I. Mayordomo, J. R. McLaughlin, H. Morgenstern, H. Müeller, I. Orlow, B. J. Park, M. Pinchev, O. Y. Raji, H. S. Rennert, P. Rudnai, A. Seow, I. Stucker, N. Szeszenia-Dabrowska, M. Dawn Teare, A. Tjønnelan, D. Ugolini, H. F. M. van der Heijden, E. Wichmann, J. K. Wiencke, P. J. Woll, P. Yang, D. Zaridze, Z.-F. Zhang, C. J. Etzel and R. J. Hung (2012). "Increased risk of lung cancer in individuals with a family history of the disease: A pooled analysis from the International Lung Cancer Consortium." European Journal of Cancer 48(13): 1957-1968.

Couraud, S., G. Zalcman, B. Milleron, F. Morin and P.-J. Souquet (2012). "Lung cancer in never smokers – A review." European Journal of Cancer 48(9): 1299-1311.

Coyle, Y. M., A. T. Minahjuddin, L. S. Hynan and J. D. Minna (2006). "An ecological study of the association of metal air pollutants with lung cancer incidence in Texas." J Thorac Oncol 1(7): 654-661.

Dai, J. Y., C. Kooperberg, M. Leblanc and R. L. Prentice (2012). "Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction." Biometrika 9: 929-944.

Dai, J. Y., B. A. Logsdon, Y. Huang, L. Hsu, A. P. Reiner, R. L. Prentice and C. Kooperberg (2012). "Simultaneously testing for marginal genetic association and gene-environment interaction." Am J Epidemiol 176(2): 164-173.

Darwin, C. R. (1869). On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. London, John Murray.

Dehling, H. and B. Haupt (2004). Einführung in die Wahrscheinlichkeitstheorie und Statistik, Springer Fachmedien Wiesbaden.

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

Eiichiro, F. (2004). "Worcester's log-linear model for three-dimensional contingency table."

Fehringer, G., G. Liu, M. Pintilie, J. Sykes, D. Cheng, N. Liu, Z. Chen, L. Seymour, S. D. Der, F. A. Shepherd, M. S. Tsao and R. J. Hung (2012). "Association of the 15q25 and 5p15 lung cancer susceptibility regions with gene expression in lung tumor tissue." Cancer Epidemiol Biomarkers Prev 21(7): 1097-1104.

Ferlay, J., P. Autier, M. Boniol, M. Heanue, M. Colombet and P. Boyle (2007). "Estimates of the cancer incidence and mortality in Europe in 2006." Ann Oncol 18(3): 581-592.

Finlin, B. S., C. L. Gau, G. A. Murphy, H. Shao, T. Kimel, R. S. Seitz, Y. F. Chiu, D. Botstein, P. O. Brown, C. J. Der, F. Tamanoi, D. A. Andres and C. M. Perou (2001). "RERG is a novel ras-related, estrogen-regulated and growth-inhibitory gene in breast cancer." J Biol Chem 276(45): 42259-42267.

Frazer, K. A., S. S. Murray, N. J. Schork and E. J. Topol (2009). "Human genetic variation and its contribution to complex traits." Nature Reviews Genetics 10(4): 241-251.

Gelman, A., J. B. Carlin, H. S. Stern and D. B. Rubin (1995). Bayesian Data Analysis. London, Chapman & Hall.

Gorlova, O. Y., S.-F. Weng, Y. Zhang, C. I. Amos and M. R. Spitz (2007). "Aggregation of cancer among relatives of never-smoking lung cancer patients." International Journal of Cancer 121(1): 111-118.

Gu, M., X. Dong, X. Zhang, X. Wang, Y. Qi, J. Yu and W. Niu (2012). "Strong association between two polymorphisms on 15q25.1 and lung cancer risk: a meta-analysis." PLoS One 7(6): e37970.

Hardy, J. and A. Singleton (2009). "Genomewide association studies and human disease." N Engl J Med 360(17): 1759-1768.

Haugen, A., D. Ryberg, S. Mollerup, S. Zienolddiny, V. Skaug and D. H. Svendsrud (2000). "Gene-environment interactions in human lung cancer." Toxicol Lett 112-113: 233-237.

Heller, G., C. C. Zielinski and S. Zochbauer-Muller (2010). "Lung cancer: from single-gene methylation to methylome profiling." Cancer Metastasis Rev 29(1): 95-107.

Herbst, R. S., J. V. Heymach and S. M. Lippman (2008). "Lung cancer." N Engl J Med 359(13): 1367-1380.

Heron, E. A., C. O'Dushlaine, R. Segurado, L. Gallagher and M. Gill (2011). "Exploration of empirical Bayes hierarchical modeling for the analysis of genome-wide association study data." Biostatistics 12(3): 445-461.

Hishimoto, A., Q. R. Liu, T. Drgon, O. Pletnikova, D. Walther, X. G. Zhu, J. C. Troncoso and G. R. Uhl (2007). "Neurexin 3 polymorphisms are associated with alcohol dependence and altered expression of specific isoforms." Hum Mol Genet 16(23): 2880-2891.

Holle, R., M. Happich, H. Löwel and H. E. Wichmann (2005). "KORA - a research platform for population based health research." Gesundheitswesen 67(S 01): 19-25.

Holsinger, K. E. and B. S. Weir (2009). "Genetics in geographically structured populations: defining, estimating and interpreting FST." Nat Rev Genet 10(9): 639-650.

Hostetler, B., N. Weston, C. Kim, D. Morré and D. J. Morré (2009). "Cancer site-specific isoforms of ENOX2 (tNOX), a cancer-specific cell surface oxidase." Clinical Proteomics 5(1): 46-51.

Hsiung, C. A., Q. Lan, Y.-C. Hong, C.-J. Chen, H. D. Hosgood, III, I. S. Chang, N. Chatterjee, P. Brennan, C. Wu, W. Zheng, G.-C. Chang, T. Wu, J. Y. Park, C.-F. Hsiao, Y. H. Kim, H. Shen, A. Seow, M. Yeager, Y.-H. Tsai, Y. T. Kim, W.-H. Chow, H. Guo, W.-C. Wang, S. W. Sung, Z. Hu, K.-Y. Chen, J. H. Kim, Y. Chen, L. Huang, K.-M. Lee, Y.-L. Lo, Y.-T. Gao, J. H. Kim, L. Liu, M.-S. Huang, T. H. Jung, G. Jin, N. Caporaso, D. Yu, C. H. Kim, W.-C. Su, X.-O. Shu, P. Xu, I.-S. Kim, Y.-M. Chen, H. Ma, M. Shen, S. I. Cha, W. Tan, C.-H. Chang, J. S. Sung, M. Zhang, T.-Y. Yang, K. H. Park, J. Yuenger, C.-L. Wang, J.-S. Ryu, Y. Xiang, Q. Deng, A. Hutchinson, J. S. Kim, Q. Cai, M. T. Landi, C.-J. Yu, J.-Y. Park, M. Tucker, J.-Y. Hung, C.-C. Lin, R.-P. Perng, P. Boffetta, C.-Y. Chen, K.-C. Chen, S.-Y. Yang, C.-Y. Hu, C.-K. Chang, J. F. Fraumeni, Jr., S. Chanock, P.-C. Yang, N. Rothman and D. Lin (2010). "The 5p15.33 locus is associated with risk of lung adenocarcinoma in never-smoking females in asia." PLoS Genet 6(8): e1001051.

Hung, R. J., D. C. Christiani, A. Risch, O. Popanda, A. Haugen, S. Zienolddiny, S. Benhamou, C. Bouchardy, Q. Lan, M. R. Spitz, H.-E. Wichmann, L. LeMarchand, P. Vineis, G. Matullo, C. Kiyohara, Z.-F. Zhang, B. Pezeshki, C. Harris, L. Mechanic, A. Seow, D. P. K. Ng, N. Szeszenia-Dabrowska, D. Zaridze, J. Lissowska, P. Rudnai, E. Fabianova, D. Mates, L. Foretova, V. Janout, V. Bencko, N. Caporaso, C. Chen, E. J. Duell, G. Goodman, J. K. Field, R. S. Houlston, Y.-C. Hong, M. T. Landi, P. Lazarus, J. Muscat, J. McLaughlin, A. G. Schwartz, H. Shen, I. Stucker, K. Tajima, K. Matsuo, M. Thun, P. Yang, J. Wiencke, A. S. Andrew, S. Monnier, P. Boffetta and P. Brennan (2008). "International lung cancer consortium: pooled analysis of sequence variants in DNA repair and cell cycle pathways." Cancer Epidemiology Biomarkers & Prevention 17(11): 3081-3089.

Hung, R. J., J. D. McKay, V. Gaborieau, P. Boffetta, M. Hashibe, D. Zaridze, A. Mukeria, N. Szeszenia-Dabrowska, J. Lissowska, P. Rudnai, E. Fabianova, D. Mates, V. Bencko, L. Foretova, V. Janout, C. Chen, G. Goodman, J. K. Field, T. Liloglou, G. Xinarianos, A. Cassidy, J. McLaughlin, G. Liu, S. Narod, H. E. Krokan, F. Skorpen, M. B. Elvestad, K. Hveem, L. Vatten, J. Linseisen, F. Clavel-Chapelon, P. Vineis, H. B. Bueno-de-Mesquita, E. Lund, C. Martinez, S. Bingham, T. Rasmuson, P. Hainaut, E. Riboli, W. Ahrens, S. Benhamou, P. Lagiou, D. Trichopoulos, I. Holcatova, F. Merletti, K. Kjaerheim, A. Agudo, G. Macfarlane, R. Talamini, L. Simonato, R. Lowry, D. I. Conway, A. Znaor, C. Healy, D. Zelenika, A. Boland, M. Delepine, M. Foglio, D. Lechner, F. Matsuda, H. Blanche, I. Gut, S. Heath, M.

Lathrop and P. Brennan (2008). "A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25." Nature 452(7187): 633-637.

Hunter, D. J. (2005). "Gene-environment interactions in human diseases." Nat Rev Genet 6(4): 287-298.

International HapMap Consortium (2003). "The International HapMap Project." Nature 426(6968): 789-796.

International HapMap Consortium (2005). "A haplotype map of the human genome." Nature 437(7063): 1299-1320.

International HapMap Consortium, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Y. Waye, S. K. W. Tsui, H. Xue, J. T.-F. Wong, L. M. Galver, J.-B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J.-F. o. Olivier, M. S. Phillips, S. p. Roumy, C. m. Sallee, A. Verner, T. J. Hudson, P.-Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L.-C. Tsui, W. Mak, Y. Q. Song, P. K. H. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. W. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y.

Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. M. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archeveque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. ennedy, R. Jamieson and J. Stewart (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature 449(7164): 851-861.

International Human Genome Sequencing Consortium (2004). "Finishing the euchromatic sequence of the human genome." Nature 431(7011): 931-945.

Itsara, A., G. M. Cooper, C. Baker, S. Girirajan, J. Li, D. Absher, R. M. Krauss, R. M. Myers, P. M. Ridker, D. I. Chasman, H. Mefford, P. Ying, D. A. Nickerson and E. E. Eichler (2009). "Population analysis of large copy number variants and hotspots of human genetic disease." Am J Hum Genet 84(2): 148-161.

Jemal, A., F. Bray, M. M. Center, J. Ferlay, E. Ward and D. Forman (2011). "Global cancer statistics." CA: A Cancer Journal for Clinicians 61(2): 69-90.

Jemal, A., R. Siegel, E. Ward, Y. Hao, J. Xu, T. Murray and M. J. Thun (2008). "Cancer statistics, 2008." CA: A Cancer Journal for Clinicians 58(2): 71-96.

Jemal, A., E. P. Simard, C. Dorell, A. M. Noone, L. E. Markowitz, B. Kohler, C. Eheman, M. Saraiya, P. Bandi, D. Saslow, K. A. Cronin, M. Watson, M. Schiffman, S. J. Henley, M. J. Schymura, R. N. Anderson, D. Yankey and B. K. Edwards (2013). "Annual report to the nation on the status of cancer, 1975-2009, featuring the burden and trends in human papillomavirus (HPV)-associated cancers and HPV vaccination coverage levels." J Natl Cancer Inst 105(3): 175-201.

Jiang, Z., N. M. Gorenstein, D. M. Morré and D. J. Morré (2008). "Molecular cloning and characterization of a candidate human growth-related and time-keeping constitutive cell surface hydroquinone (NADH) oxidase." Biochemistry 47(52): 14028-14038.

Johnson, A. D. (2009). "Single-nucleotide polymorphism bioinformatics: a comprehensive review of resources." Circulation: Cardiovascular Genetics 2(5): 530-536.

Kabir, Z., K. Bennett and L. Clancy (2007). "Lung cancer and urban air-pollution in Dublin: a temporal association?" Ir Med J 100(2): 367-369.

Kass, R. E. and D. Steffey (1989). "Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models)." Journal of the American Statistical Association 84(407): 717-726.

Khoury, M. J., T. H. Beaty and B. H. Cohen (1993). Fundamentals of Genetic Epidemiology. New York, Oxford University Press, Inc.

Khoury, M. J. and W. D. Flanders (1996). "Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls!" Am J Epidemiol 144(3): 207-213.

Köhler, K. and H. Bickeböller (2006). "Case-Control Association Tests Correcting for Population Stratification." Annals of Human Genetics 70(1): 98-115.

Kostenko, S., G. Dumitriu and U. Moens (2012). "Tumour promoting and suppressing roles of the atypical MAP kinase signalling pathway ERK3/4-MK5." J Mol Signal 7(1): 9.

Kraft, P., Y. C. Yen, D. O. Stram, J. Morrison and W. J. Gauderman (2007). "Exploiting gene-environment interaction to detect genetic associations." Hum Hered 63(2): 111-119.

Krebsregister, R. K.-I. R. u. d. G. d. e. and i. D. e. V. (GEKID). (2013). "Krebs in Deutschland 2009/2010. 8. Ausgabe. Gesundheitsberichterstattung des Bundes. Berlin." Retrieved Accessed 31 January 2014, from http://www.krebsdaten.de/Krebs/DE/Content/Publikationen/Krebs_in_Deutschland/krebs_in_deutschland_node.html.

Kuo, C. L. and D. V. Zaykin (2011). "Novel rank-based approaches for discovery and replication in genome-wide association studies." Genetics 189(1): 329-340.

Kusinska, R., P. Górniak, A. Pastorczak, W. Fendler, P. Potemski, W. Mlynarski and R. Kordek (2012). "Influence of genomic variation in FTO at 16q12.2, MC4R at 18q22 and NRXN3 at 14q31 genes on breast cancer risk." Molecular Biology Reports 39(3): 2915-2919.

Landi, M. T., N. Chatterjee, K. Yu, L. R. Goldin, A. M. Goldstein, M. Rotunno, L. Mirabello, K. Jacobs, W. Wheeler, M. Yeager, A. W. Bergen, Q. Li, D. Consonni, A. C. Pesatori, S. Wacholder, M. Thun, R. Diver, M. Oken, J. Virtamo, D. Albanes, Z. Wang, L. Burdette, K. F. Doheny, E. W. Pugh, C. Laurie, P. Brennan, R. Hung, V. Gaborieau, J. D. McKay, M. Lathrop, J. McLaughlin, Y. Wang, M. S. Tsao, M. R. Spitz, Y. Wang, H. Krokan, L. Vatten, F. Skorpen, E. Arnesen, S. Benhamou, C. Bouchard, A. Metspalu, T. Vooder, M. Nelis, K. Valk, J. K. Field, C. Chen, G. Goodman, P. Sulem, G. Thorleifsson, T. Rafnar, T. Eisen, W. Sauter, A. Rosenberger, H. Bickeboller, A. Risch, J. Chang-Claude, H. E. Wichmann, K. Stefansson, R. Houlston, C. I. Amos, J. F. Fraumeni, Jr., S. A. Savage, P. A. Bertazzi, M. A. Tucker, S. Chanock and N. E. Caporaso (2009). "A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma." Am J Hum Genet 85(5): 679-691.

Lee, P. M. (1997). Bayesian Statistics - An Introduction. London, Arnold.

Lee, S. J., H.-S. Jeon, J.-S. Jang, S. H. Park, G. Y. Lee, B.-H. Lee, C. H. Kim, Y. M. Kang, W. K. Lee, S. Kam, R. W. Park, I.-S. Kim, Y. L. Cho, T. H. Jung and J. Y. Park (2005). "DNMT3B polymorphisms and risk of primary lung cancer." Carcinogenesis 26(2): 403-409.

Lee, W. C. and L. Y. Wang (2008). "Simple formulas for gauging the potential impacts of population stratification bias." American Journal of Epidemiology 167(1): 86-89.

Lettre, G., C. Lange and J. N. Hirschhorn (2007). "Genetic model testing and statistical power in population-based association studies of quantitative traits." Genet Epidemiol 31(4): 358-362.

Lewinger, J. P., D. V. Conti, J. W. Baurley, T. J. Triche and D. C. Thomas (2007). "Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation." Genet Epidemiol 31(8): 871-882.

Li, C., Z. Yin, W. Wu, X. Li, Y. Ren and B. Zhou (2013). "Genetic variations in TERT-CLPTM1L genes and risk of lung cancer in chinese women nonsmokers." PLoS ONE 8(5): e64988.

Liang, B., S. Wang, X. Zhu, Y. Yu, Z. Ciu and Y. Yu (2005). "Increased expression of mitogen-activated protein kinase and its upstream regulating signal in human gastric cancer." World J Gastroenterol 11: 623 - 628.

Lips, E. H., V. Gaborieau, J. D. McKay, A. Chabrier, R. J. Hung, P. Boffetta, M. Hashibe, D. Zaridze, N. Szeszenia-Dabrowska, J. Lissowska, P. Rudnai, E. Fabianova, D. Mates, V. Bencko, L. Foretova, V. Janout, J. K. Field, T. Liloglou, G. Xinarianos, J. McLaughlin, G. Liu, F. Skorpen, M. B. Elvestad, K. Hveem, L. Vatten, E. Study, S. Benhamou, P. Lagiou, I. Holcátová, F. Merletti, K. Kjaerheim, A. Agudo, X. Castellsagué, T. V. Macfarlane, L. Barzan, C. Canova, R. Lowry, D. I. Conway, A. Znaor, C. Healy, M. P. Curado, S. Koifman, J. Eluf-Neto, E. Matos, A. Menezes, L. Fernandez, A. Metspalu, S. Heath, M. Lathrop and P. Brennan (2010). "Association between a 15q25 gene variant, smoking quantity and tobacco-related cancers among 17 000 individuals." International Journal of Epidemiology 39(2): 563-577.

Liu, J. Z., F. Tozzi, D. M. Waterworth, S. G. Pillai, P. Muglia, L. Middleton, W. Berrettini, C. W. Knouff, X. Yuan, G. Waeber, P. Vollenweider, M. Preisig, N. J. Wareham, J. H. Zhao, R. J. Loos, I. Barroso, K. T. Khaw, S. Grundy, P. Barter, R. Mahley, A. Kesaniemi, R. McPherson, J. B. Vincent, J. Strauss, J. L. Kennedy, A. Farmer, P. McGuffin, R. Day, K. Matthews, P. Bakke, A. Gulsvik, S. Lucae, M. Ising, T. Brueckl, S. Horstmann, H. E. Wichmann, R. Rawal, N. Dahmen, C. Lamina, O. Polasek, L. Zgaga, J. Huffman, S. Campbell, J. Kooner, J. C. Chambers, M. S. Burnett, J. M. Devaney, A. D. Pichard, K. M. Kent, L. Satler, J. M. Lindsay, R. Waksman, S. Epstein, J. F. Wilson, S. H. Wild, H. Campbell, V. Vitart, M. P. Reilly, M. Li, L. Qu, R. Wilensky, W. Matthai, H. H. Hakonarson, D. J. Rader, A. Franke, M. Wittig, A. Schafer, M. Uda, A. Terracciano, X. Xiao, F. Busonero, P. Scheet, D. Schlessinger, D. St Clair, D. Rujescu, G. R. Abecasis, H. J. Grabe, A. Teumer, H. Volzke, A. Petersmann, U. John, I. Rudan, C. Hayward, A. F. Wright, I. Kolcic, B. J. Wright, J. R. Thompson, A. J. Balmforth, A. S. Hall, N. J. Samani, C. A. Anderson, T. Ahmad, C. G. Mathew, M. Parkes, J. Satsangi, M. Caulfield, P. B. Munroe, M. Farrall, A. Dominiczak, J. Worthington, W. Thomson, S. Eyre, A. Barton, V. Mooser, C. Francks and J. Marchini (2010). "Meta-analysis and imputation refines the association of 15q25 with smoking quantity." Nat Genet 42(5): 436-440.

Malvezzi, M., P. Bertuccio, F. Levi, C. La Vecchia and E. Negri (2013). "European cancer mortality predictions for the year 2013." Annals of Oncology.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll and P. M. Visscher (2009). "Finding the missing heritability of complex diseases." Nature 461(7265): 747-753.

Mardis, E. R. (2008). "Next-generation DNA sequencing methods." Annu Rev Genomics Hum Genet 9: 387-402.

Matakidou, A., T. Eisen and R. S. Houlston (2005). "Systematic review of the relationship between family history and lung cancer risk." Br J Cancer 93(7): 825-833.

McCarroll, S. A. (2008). "Extending genome-wide association studies to copy-number variation." Hum Mol Genet 17(R2): R135-142.

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis and J. N. Hirschhorn (2008). "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." Nat Rev Genet 9(5): 356-369.

McKay, J. D., R. J. Hung, V. Gaborieau, P. Boffetta, A. Chabrier, G. Byrnes, D. Zaridze, A. Mukeria, N. Szeszenia-Dabrowska, J. Lissowska, P. Rudnai, E. Fabianova, D. Mates, V. Bencko, L. Foretova, V. Janout, J. McLaughlin, F. Shepherd, A. Montpetit, S. Narod, H. E. Krokan, F. Skorpen, M. B. Elvestad, L. Vatten, I. Njølstad, T. Axelsson, C. Chen, G. Goodman, M. Barnett, M. M. Loomis, J. Lubiński, J. Matyjasik, M. Lener, D. Oszutowska, J. Field, T. Liloglou, G. Xinarianos, A. Cassidy, E. P. I. C. Study, P. Vineis, F. Clavel-Chapelon, D. Palli, R. Tumino, V. Krogh, S. Panico, C. A. González, J. R. Quirós, C. Martínez, C. Navarro, E. Ardanaz, N. Larrañaga, K. T. Kham, T. Key, H. B. B. de Mesquita, P. H. Peeters, A. Trichopoulou, J. Linseisen, H. Boeing, G. Hallmans, K. Overvad, A. Tjønneland, M. Kumle, E. Riboli, D. Zelenika, A. Boland, M. Delepine, M. Foglio, D. Lechner, F. Matsuda, H. Blanche, I. Gut, S. Heath, M. Lathrop and P. Brennan (2008). "Lung cancer susceptibility locus at 5p15.33." Nat Genet 40(12): 1404-1406.

Meyer, C., A. Brieger, G. Plotz, N. Weber, S. Passmann, T. Dingermann, S. Zeuzem, J. Trojan and R. Marschalek (2009). "An interstitial deletion at 3p21.3 results in the genetic fusion of MLH1 and ITGA9 in a Lynch syndrome family." Clin Cancer Res 15(3): 762-769.

Moncho-Amor, V., I. Ibanez de Caceres, E. Bandres, B. Martinez-Poveda, J. L. Orgaz, I. Sanchez-Perez, S. Zazo, A. Rovira, J. Albanell, B. Jimenez, F. Rojo, C. Belda-Iniesta, J. Garcia-Foncillas and R. Perona (2011). "DUSP1/MKP1 promotes angiogenesis, invasion and metastasis in non-small-cell lung cancer." Oncogene 30(6): 668-678.

Morré, D. J. and D. M. Morré (2012). ECTO-NOX proteins: growth, cancer, and aging, Springer New York Heidelberg Dordrecht London.

Morris, C. N. (1983). "Parametric Empirical Bayes Inference: Theory and Applications." Journal of the American Statistical Association 78(381): 47-55.

Mukherjee, B., J. Ahn, S. B. Gruber and N. Chatterjee (2012). "Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons." Am J Epidemiol 175(3): 177-190.

Mukherjee, B., J. Ahn, S. B. Gruber, G. Rennert, V. Moreno and N. Chatterjee (2008). "Tests for gene-environment interaction from case-control data: a novel study of type I error, power and designs." Genet Epidemiol 32(7): 615-626.

Mukherjee, B. and N. Chatterjee (2008). "Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency." Biometrics 64(3): 685-694.

Murcray, C. E., J. P. Lewinger, D. V. Conti, D. C. Thomas and W. J. Gauderman (2011). "Sample size requirements to detect gene-environment interactions in genome-wide association studies." Genet Epidemiol 35(3): 201-210.

Murcray, C. E., J. P. Lewinger and W. J. Gauderman (2009). "Gene-environment interaction in genome-wide association studies." Am J Epidemiol 169(2): 219-226.

Murcray, C. E., J. P. Lewinger and W. J. Gauderman (2009). "Gene-environment interaction in genome-wide association studies." American Journal of Epidemiology 169(2): 219-226.

Myneni, A. A., S. C. Chang, R. Niu, L. Liu, H. M. Ochs-Balcom, Y. Li, C. Zhang, B. Zhao, J. Shi, X. Han, J. Li, J. Su, L. Cai, S. Yu, Z. F. Zhang and L. Mu (2013). "Genetic polymorphisms of TERT and CLPTM1L and risk of lung cancer--a case-control study in a Chinese population." Lung Cancer 80(2): 131-137.

O'Reilly, K. M., A. M. McLaughlin, W. S. Beckett and P. J. Sime (2007). "Asbestos-related lung disease." Am Fam Physician 75(5): 683-688.

Ober, C. and D. Vercelli (2011). "Gene-environment interactions in human disease: nuisance or opportunity?" Trends Genet 27(3): 107-115.

Online Mendelian Inheritance in Man (2012). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), Available at \url{http://www.omim.org/}[Accessed 24 February 2012].

Osann, K. E. (1991). "Lung cancer in women: the importance of smoking, family history of cancer, and medical history of respiratory disease." Cancer Research 51(18): 4893-4897.

Patterson, N., A. L. Price and D. Reich (2006). "Population Structure and Eigenanalysis." PLoS Genet 2(12): e190.

Piegorsch, W. W. and G. Casella (1996). "Empirical Bayes estimation for logistic regression and extended parametric regression models." Journal of Agricultural, Biological, and Environmental Statistics 1(2): 231-249.

Piegorsch, W. W., C. R. Weinberg and J. A. Taylor (1994). "Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies." Stat Med 13(2): 153-162.

Prentice, R. L. and R. Pyke (1979). "Logistic disease incidence models and case-control studies." Biometrika 66(3): 403-411.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick and D. Reich (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nat Genet 38(8): 904-909.

Price, A. L., N. A. Zaitlen, D. Reich and N. Patterson (2010). "New approaches to population stratification in genome-wide association studies." Nat Rev Genet 11(7): 459-463.

Pritchard, J. K. and N. A. Rosenberg (1999). "Use of unlinked genetic markers to detect population stratification in association studies." Am J Hum Genet 65(1): 220-228.

Pritchard, J. K., M. Stephens and P. Donnelly (2000). "Inference of population structure using multilocus genotype data." Genetics 155(2): 945-959.

Pritchard, J. K., M. Stephens, N. A. Rosenberg and P. Donnelly (2000). "Association mapping in structured populations." Am J Hum Genet 67(1): 170-181.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly and P. C. Sham (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." American journal of human genetics 81(3): 559-575.

Raimondi, S., V. Paracchini, H. Autrup, J. M. Barros-Dios, S. Benhamou, P. Boffetta, M. L. Cote, I. A. Dialyna, V. Dolzan, R. Filiberti, S. Garte, A. Hirvonen, K. Husgafvel-Pursiainen, E. N. Imyanitov, I. Kalina, D. Kang, C. Kiyohara, T. Kohno, P. Kremers, Q. Lan, S. London, A. C. Povey, A. Rannug, E. Reszka, A. Risch, M. Romkes, J. Schneider, A. Seow, P. G. Shields, R. C. Sobti, M. Sorensen, M. Spinola, M. R. Spitz, R. C. Strange, I. Stucker, H. Sugimura, J. To-Figueras, S. Tokudome, P. Yang, J. M. Yuan, M. Warholm and E. Taioli (2006). "Meta- and pooled analysis of GSTT1 and lung cancer: a HuGE-GSEC review." Am J Epidemiol 164(11): 1027-1042.

Redon, R., S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer and M. E. Hurles (2006). "Global variation in copy number in the human genome." Nature 444(7118): 444-454.

Rees, J. L. (2004). "The genetics of sun sensitivity in humans." Am J Hum Genet 75(5): 739-751.

Robert, C. P. (1994). The Bayesian Choice: a decision-theoretic motivation. New York, Springer.

Robert Koch-Institut und die Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. (2012). Krebs in Deutschland 2007/2008, Available at \url{http://www.krebsdaten.de/Krebs/DE/Content/Publikationen/Krebs_in_Deutschland/krebs_in_deutschland_node.html}[Accessed 15 March 2012].

Rothman, K. J., S. Greenland and A. M. Walker (1980). "Concepts of interaction." Am J Epidemiol 112(4): 467-470.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander and D. Altshuler (2001). "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms." Nature 409(6822): 928-933.

Sasieni, P. D. (1997). "From genotypes to genes: doubling the sample size." Biometrics 53(4): 1253-1261.

Sauter, W., A. Rosenberger, L. Beckmann, S. Kropp, K. Mittelstrass, M. Timofeeva, G. Wölke, A. Steinwachs, D. Scheiner, E. Meese, G. Sybrecht, F. Kronenberg, H. Dienemann, J. Chang-Claude, T. Illig, H.-E. Wichmann, H. Bickeböller and A. Risch (2008). "Matrix metalloproteinase 1 (MMP1) is associated with early-onset lung cancer." Cancer Epidemiology Biomarkers & Prevention 17(5): 1127-1135.

Scelo, G., V. Constantinescu, I. Csiki, D. Zaridze, N. Szeszenia-Dabrowska, P. Rudnai, J. Lissowska, E. Fabianova, A. Cassidy, A. Slamova, L. Foretova, V. Janout, J. Fevotte, T. Fletcher, A. Mannetje, P. Brennan and P. Boffetta (2004). "Occupational exposure to vinyl chloride, acrylonitrile and styrene and lung cancer risk (europe)." Cancer Causes Control 15(5): 445-452.

Schmidt, S. and D. J. Schaid (1999). "Potential misinterpretation of the case-only study to assess gene-environment interaction." Am J Epidemiol 150(8): 878-885.

Schuster, S. C. (2008). "Next-generation sequencing transforms today's biology." Nat Methods 5(1): 16-18.

Shi, J., N. Chatterjee, M. Rotunno, Y. Wang, A. C. Pesatori, D. Consonni, P. Li, W. Wheeler, P. Broderick, M. Henrion, T. Eisen, Z. Wang, W. Chen, Q. Dong, D. Albanes, M. Thun, M. R. Spitz, P. A. Bertazzi, N. E. Caporaso, S. J. Chanock, C. I. Amos, R. S. Houlston and M. T. Landi (2012). "Inherited variation at chromosome 12p13.33, including RAD52, influences the risk of squamous cell lung carcinoma." Cancer Discov 2(2): 131-139.

Smith, P. G. and N. E. Day (1984). "The design of case-control studies: the influence of confounding and interaction effects." Int J Epidemiol 13(3): 356-365.

Smoking, O. o. and and Health (2006). Publications and Reports of the Surgeon General. The health consequences of involuntary exposure to tobacco smoke: a report of the surgeon general. Atlanta (GA), Centers for Disease Control and Prevention (US).

Sohns, M. (2012). The empirical hierarchical Bayes approach for pathway integration and gene-environment interactions in genome-wide association studies Doctoral Thesis, Georg-August University of Göttingen.

Sohns, M., E. Viktorova, C. I. Amos, P. Brennan, G. Fehringer, V. Gaborieau, Y. Han, J. Heinrich, J. Chang-Claude, R. J. Hung, M. Muller-Nurasyid, A. Risch, J. P. Lewinger, D. C. Thomas and H. Bickeböller (2013). "Empirical hierarchical bayes approach to gene-environment interactions: development and application to genome-wide association studies of lung cancer in TRICL." Genet Epidemiol 37(6): 551-559.

Spencer, C. C., Z. Su, P. Donnelly and J. Marchini (2009). "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip." PLoS Genet 5(5): e1000477.

Steffens, M., C. Lamina, T. Illig, T. Bettecken, R. Vogler, P. Entz, E. K. Suk, M. R. Toliat, N. Klopp, A. Caliebe, I. R. Konig, K. Kohler, J. Ludemann, A. Diaz Lacava, R. Fimmers, P. Lichtner, A. Ziegler, A. Wolf, M. Krawczak, P. Nurnberg, J. Hampe, S. Schreiber, T. Meitinger, H. E. Wichmann, K. Roeder, T. F. Wienker and M. P. Baur (2006). "SNP-based analysis of genetic substructure in the German population." Hum Hered 62(1): 20-29.

Subramanian, J. and R. Govindan (2007). "Lung cancer in never smokers: a review." J Clin Oncol 25(5): 561-570.

Sun, S., J. H. Schiller and A. F. Gazdar (2007). "Lung cancer in never smokers-a different disease." Nat Rev Cancer 7(10): 778-790.

Sun, S., J. H. Schiller and A. F. Gazdar (2007). "Lung cancer in never smokers - a different disease." Nat Rev Cancer 7(10): 778-790.

Syvanen, A.-C. (2001). "Accessing genetic variation: genotyping single nucleotide polymorphisms." Nat Rev Genet 2(12): 930-942.

Takahashi, Y., A. R. R. Forrest, E. Maeno, T. Hashimoto, C. O. Daub and J. Yasuda (2009). "MiR-107 and MiR-185 Can Induce Cell Cycle Arrest in Human Non Small Cell Lung Cancer Cell Lines." PLoS ONE 4(8): e6677.

Tamura, K., X. Zhang, Y. Murakami, S. Hirohashi, H.-J. Xu, S.-X. Hu, W. F. Benedict and T. Sekiya (1997). "Deletion of three distinct regions on chromosome 13q in human non-small-cell lung cancer." International Journal of Cancer 74(1): 45-49.

Thomas, D. (2010a). "Gene-environment-wide association studies: emerging approaches." Nat Rev Genet 11(4): 259-272.

Thomas, D. (2010b). "Methods for Investigating Gene-Environment Interactions in Candidate Pathway and Genome-Wide Association Studies." Annual Review of Public Health 31(1): 21-36.

Thomas, D. C., J. P. Lewinger, C. E. Murcray and W. J. Gauderman (2012). "Invited commentary: GE-Whiz! Ratcheting gene-environment studies up to the whole genome and the whole exposome." Am J Epidemiol 175(3): 203-207; discussion 208-209.

Thomas, P. D. and A. Kejariwal (2004). "Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences in molecular effects." Proceedings of the National Academy of Sciences of the United States of America 101(43): 15398-15403.

Thorgeirsson, T. E., F. Geller, P. Sulem, T. Rafnar, A. Wiste, K. P. Magnusson, A. Manolescu, G. Thorleifsson, H. Stefansson, A. Ingason, S. N. Stacey, J. T. Bergthorsson, S. Thorlacius, J.

Gudmundsson, T. Jonsson, M. Jakobsdottir, J. Saemundsdottir, O. Olafsdottir, L. J. Gudmundsson, G. Bjornsdottir, K. Kristjansson, H. Skuladottir, H. J. Isaksson, T. Gudbjartsson, G. T. Jones, T. Mueller, A. Gottsater, A. Flex, K. K. Aben, F. de Vegt, P. F. Mulders, D. Isla, M. J. Vidal, L. Asin, B. Saez, L. Murillo, T. Blondal, H. Kolbeinsson, J. G. Stefansson, I. Hansdottir, V. Runarsdottir, R. Pola, B. Lindblad, A. M. van Rij, B. Dieplinger, M. Haltmayer, J. I. Mayordomo, L. A. Kiemeney, S. E. Matthiasson, H. Oskarsson, T. Tyrfingsson, D. F. Gudbjartsson, J. R. Gulcher, S. Jonsson, U. Thorsteinsdottir, A. Kong and K. Stefansson (2008). "A variant associated with nicotine dependence, lung cancer and peripheral arterial disease." Nature 452(7187): 638-642.

Thorgeirsson, T. E., D. F. Gudbjartsson, I. Surakka, J. M. Vink, N. Amin, F. Geller, P. Sulem, T. Rafnar, T. Esko, S. Walter, C. Gieger, R. Rawal, M. Mangino, I. Prokopenko, R. Magi, K. Keskitalo, I. H. Gudjonsdottir, S. Gretarsdottir, H. Stefansson, J. R. Thompson, Y. S. Aulchenko, M. Nelis, K. K. Aben, M. den Heijer, A. Dirksen, H. Ashraf, N. Soranzo, A. M. Valdes, C. Steves, A. G. Uitterlinden, A. Hofman, A. Tonjes, P. Kovacs, J. J. Hottenga, G. Willemsen, N. Vogelzangs, A. Doring, N. Dahmen, B. Nitz, M. L. Pergadia, B. Saez, V. De Diego, V. Lezcano, M. D. Garcia-Prats, S. Ripatti, M. Perola, J. Kettunen, A. L. Hartikainen, A. Pouta, J. Laitinen, M. Isohanni, S. Huei-Yi, M. Allen, M. Krestyaninova, A. S. Hall, G. T. Jones, A. M. van Rij, T. Mueller, B. Dieplinger, M. Haltmayer, S. Jonsson, S. E. Matthiasson, H. Oskarsson, T. Tyrfingsson, L. A. Kiemeney, J. I. Mayordomo, J. S. Lindholt, J. H. Pedersen, W. A. Franklin, H. Wolf, G. W. Montgomery, A. C. Heath, N. G. Martin, P. A. Madden, I. Giegling, D. Rujescu, M. R. Jarvelin, V. Salomaa, M. Stumvoll, T. D. Spector, H. E. Wichmann, A. Metspalu, N. J. Samani, B. W. Penninx, B. A. Oostra, D. I. Boomsma, H. Tiemeier, C. M. van Duijn, J. Kaprio, J. R. Gulcher, M. I. McCarthy, L. Peltonen, U. Thorsteinsdottir and K. Stefansson (2010). "Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior." Nat Genet 42(5): 448-453.

Timofeeva, M. N., R. J. Hung, T. Rafnar, D. C. Christiani, J. K. Field, H. Bickeboller, A. Risch, J. D. McKay, Y. Wang, J. Dai, V. Gaborieau, J. McLaughlin, D. Brenner, S. A. Narod, N. E. Caporaso, D. Albanes, M. Thun, T. Eisen, H. E. Wichmann, A. Rosenberger, Y. Han, W. Chen, D. Zhu, M. Spitz, X. Wu, M. Pande, Y. Zhao, D. Zaridze, N. Szeszenia-Dabrowska, J. Lissowska, P. Rudnai, E. Fabianova, D. Mates, V. Bencko, L. Foretova, V. Janout, H. E. Krokan, M. E. Gabrielsen, F. Skorpen, L. Vatten, I. Njolstad, C. Chen, G. Goodman, M. Lathrop, S. Benhamou, T. Vooder, K. Valk, M. Nelis, A. Metspalu, O. Raji, Y. Chen, J.

Gosney, T. Liloglou, T. Muley, H. Dienemann, G. Thorleifsson, H. Shen, K. Stefansson, P. Brennan, C. I. Amos, R. Houlston and M. T. Landi (2012). "Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls." Hum Mol Genet 21(22): 4980-4995.

Truong, T., R. J. Hung, C. I. Amos, X. Wu, H. Bickeböller, A. Rosenberger, W. Sauter, T. Illig, H.-E. Wichmann, A. Risch, H. Dienemann, R. Kaaks, P. Yang, R. Jiang, J. K. Wiencke, M. Wrensch, H. Hansen, K. T. Kelsey, K. Matsuo, K. Tajima, A. G. Schwartz, A. Wenzlaff, A. Seow, C. Ying, A. Staratschek-Jox, P. Nürnberg, E. Stoelben, J. Wolf, P. Lazarus, J. E. Muscat, C. J. Gallagher, S. Zienolddiny, A. Haugen, H. F. M. van der Heijden, L. A. Kiemeney, D. Isla, J. I. Mayordomo, T. Rafnar, K. Stefansson, Z.-F. Zhang, S.-C. Chang, J. H. Kim, Y.-C. Hong, E. J. Duell, A. S. Andrew, F. Lejbkowicz, G. Rennert, H. Müller, H. Brenner, L. Le Marchand, S. Benhamou, C. Bouchardy, M. D. Teare, X. Xue, J. McLaughlin, G. Liu, J. D. McKay, P. Brennan and M. R. Spitz (2010). "Replication of Lung Cancer Susceptibility Loci at Chromosomes 15q25, 5p15, and 6p21: A Pooled Analysis From the International Lung Cancer Consortium." Journal of the National Cancer Institute 102(13): 959-971.

Uhlen, M., P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, H. Wernerus, L. Bjorling and F. Ponten (2010). "Towards a knowledge-based Human Protein Atlas." Nat Biotech 28(12): 1248-1250.

Umbach, D. M. and C. R. Weinberg (1997). "Designing and analysing case-control studies to exploit independence of genotype and exposure." Stat Med 16(15): 1731-1743.

Vanderweele, T. J., Y. A. Ko and B. Mukherjee (2013). "Environmental confounding in gene-environment interaction studies." Am J Epidemiol 178(1): 144-152.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian,

W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh and X. Zhu (2001). "The sequence of the human genome." Science 291(5507): 1304-1351.

von Bubnoff, A. (2008). "Next-generation sequencing: the race is on." Cell 132(5): 721-723.

Wacholder, S., N. Rothman and N. Caporaso (2000). "Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias." J Natl Cancer Inst 92(14): 1151-1158.

Walsh, K. M., C. I. Amos, A. S. Wenzlaff, I. P. Gorlov, J. D. Sison, X. Wu, M. R. Spitz, H. M. Hansen, E. Y. Lu, C. Wei, H. Zhang, W. Chen, S. M. Lloyd, M. L. Frazier, P. M. Bracci, M. F. Seldin, M. R. Wrensch, A. G. Schwartz and J. K. Wiencke (2012). "Association study of nicotinic acetylcholine receptor genes identifies a novel lung cancer susceptibility locus near CHRNA1 in African-Americans." Oncotarget 3(11): 1428-1438.

Wang, H. M., X. Y. Zhang and B. Jin (2013). "TERT genetic polymorphism rs2736100 was associated with lung cancer: a meta-analysis based on 14,492 subjects." Genet Test Mol Biomarkers 17(12): 937-941.

Wang, L. Y. and W. C. Lee (2008). "Population stratification bias in the case-only study for gene-environment interactions." Am J Epidemiol 168(2): 197-201.

Wang, Y., P. Broderick, E. Webb, X. Wu, J. Vijayakrishnan, A. Matakidou, M. Qureshi, Q. Dong, X. Gu, W. V. Chen, M. R. Spitz, T. Eisen, C. I. Amos and R. S. Houlston (2008). "Common 5p15.33 and 6p21.33 variants influence lung cancer risk." Nat Genet 40(12): 1407-1409.

Weale, M. E. (2010). "Quality control for genome-wide association studies." Methods In Molecular Biology 628(Genetic Variation): 341-372.

WHO, W. H. O. from http://www.who.int/gho/ncd/mortality_morbidity/cancer_text/en/.

Wichmann, H. E., C. Gieger, T. Illig and M. K. S. G. for the (2005). "KORA-gen - Resource for Population Genetics, Controls and a Broad Spectrum of Disease Phenotypes." Gesundheitswesen 67(S 01): 26-30.

Witte, J. S. (2010). "Genome-wide association studies and beyond." Annu Rev Public Health 31: 9-20 24 p following 20.

Wu, C., Z. Hu, D. Yu, L. Huang, G. Jin, J. Liang, H. Guo, W. Tan, M. Zhang, J. Qian, D. Lu, T. Wu, D. Lin and H. Shen (2009). "Genetic variants on chromosome 15q25 associated with lung cancer risk in chinese populations." Cancer Research 69(12): 5065-5072.

Zeggini, E. and A. Morris (2010). Analysis of Complex Disease Association Studies: A Practical Guide, Elsevier Science Oxford.

Zhang, Y. (2008). "ERRFI1 (ERBB receptor feedback inhibitor 1)." Atlas Genet Cytogenet Oncol Haematol.

Ziegler, A. and I. R. König (2006). A statistical approach to genetic epidemiology: concepts and applications, Wiley-VCH Weinheim.

Ziegler, A., I. R. König and J. R. Thompson (2008). "Biostatistical aspects of genome-wide association studies." Biom J 50(1): 8-28.

9. Curriculum Vitae

Elena Viktorova

Personal Data

Date of birth	27th June 1986
Place of birth	Ufa, Russian Federation

Education

since 01/2011	**PhD position** at the Institute of Genetic Epidemiology, Medical School, Georg-August-University Göttingen Supervised by Prof. H. Bickeböller
08/2008 – 08/2010	**Master of Science (M.Sci)** in Statistics at Stephen F. Austin State University, USA
09/2007 – 06/2009	**Master of Science (M.Sci)** in Applied Mathematics at Bashkir State University, Russian Federation
09/2003 – 06/2007	**Bachelor of Science (B.Sci)** in Applied Mathematics at Bashkir State University, Russian Federation

Publications

under review	E Viktorova, CI Amos, P Brennan, J Chang-Claude, G Fehringer, V Gaborieau, Y Han, RJ Hung, J Heinrich, M Müller-Nurasyid, A Risch, JP Lewinger, and H Bickeböller, An Empirical Bayes Gene-Environment Interaction Method for Genome-Wide Association Studies with an Application to Lung Cancer in TRICL // Human Heredity
2013	M Sohns*, E Viktorova*, CI Amos, P Brennan, G Fehringer, V Gaborieau, Y Han, J Heinrich, J Chang-Claude, RJ Hung, M Müller-Nurasyid, A Risch, DC Thomas, H Bickeböller, ***equally contributed authors**, Empirical Hierarchical Bayes Approach to Gene-Environment Interactions: Development and Application to Genome-Wide Association Studies of Lung Cancer in TRICL // Genetic Epidemiology Journal, **2013**